Lessons In Industrial Instrumentation

By Tony R. Kuphaldt

Version 1.0 – Released September 28, 2009

Revision history[1]

- Version 0.1 – July 2008 to September 2008 (initial development)
- Version 0.2 – released September 29, 2008 for Fall quarter student use (SafeCreative registration code 0810111072182)
- Version 0.3 – September 2008 to January 2009 (continued development)
- Version 0.4 – released January 12, 2009 for Winter quarter student use (SafeCreative registration code 0901122394919)
- Version 0.5 – January 2009 to April 2009 (continued development)
- Version 0.6 – released April 21, 2009 for public use (SafeCreative registration code 0904213106101)
- Version 0.7 – April 2009 to September 2009 (continued development)
- Version 0.8 – released September 8, 2009 for public use (SafeCreative registration code 0909094400980)
- Version 0.9 – September 8, 2009 to September 28, 2009 (finishing touches for version 1.0)

[1] Version numbers ending in odd digits are developmental (e.g. 0.7, 1.23, 4.5), with only the latest revision made accessible to the public. Version numbers ending in even digits (e.g. 0.6, 1.0, 2.14) are considered "public-release" and will be archived. Version numbers beginning with zero (e.g. 0.1, 0.2, etc.) represent incomplete editions lacking major chapters or topic coverage.

Contents

Preface

I did not want to write this book . . . honestly.

My first book project began in 1998, titled *Lessons In Electric Circuits*, and I didn't call "quit" until six volumes and five years later. Even then it was not complete, but being an open-source project it gained traction on the internet to the point where other people took over its development and it grew fine without me. The impetus for writing this first tome was a general dissatisfaction with available electronics textbooks. Plenty of textbooks exist to describe things, but few really *explain* things well for students, and the field of electronics is no exception. I wanted my book(s) to be different, and so they were. No one told me how time-consuming it was going to be to write them, though!

The next few years' worth of my spare time went to developing a set of question-and-answer worksheets designed to teach electronics theory in a Socratic, active-engagement style. This project proved quite successful in my professional life as an instructor of electronics. In the summer of 2006, my job changed from teaching electronics to teaching industrial instrumentation, and I decided to continue the Socratic mode of instruction with another set of question-and-answer worksheets.

However, the field of industrial instrumentation is not as well-represented as general electronics, and thus the array of available textbooks is not as vast. I began to re-discover the drudgery of trying to teach with inadequate texts as source material. The basis of my active teaching style was that students would spend time researching the material on their own, then engage in Socratic-style discussion with me on the subject matter when they arrived for class. This teaching technique functions in direct proportion to the quality and quantity of the research sources at the students' disposal. Despite much searching, I was unable to find a textbook adequately addressing my students' learning needs. Many textbooks I found were written in a shallow, "math-phobic" style well below the level I intended to teach to. Some reference books I found contained great information, but were often written for degreed engineers with lots of Laplace transforms and other mathematical techniques well above the level I intended to teach to. Few on either side of the spectrum actually made an effort to explain certain concepts students generally struggle to understand. I needed a text giving good, practical information and theoretical coverage at the same time.

In a futile effort to provide my students with enough information to study outside of class, I scoured the internet for free tutorials written by others. While some manufacturer's tutorials were nearly perfect for my needs, others were just as shallow as the textbooks I had found, and/or were little more than sales brochures. I found myself starting to write my own tutorials on specific topics to "plug the gaps," but then another problem arose: it became troublesome for students to navigate through dozens of tutorials in an effort to find the information they needed in their studies. What my students really needed was a *book*, not a smorgasbord of tutorials.

So here I am again, writing another textbook. This time around I have the advantage of wisdom gained from the first textbook project. For this project, I will *not*:

- . . . attempt to maintain a parallel book in HTML markup (for direct viewing on the internet). I had to go to the trouble of inventing my own quasi-XML markup language last time in an effort to generate multiple format versions of the book from the same source code. Instead, this time I will use stock LaTeX as the source code format and regular Adobe PDF format for the final output, which anyone may read thanks to its ubiquity. If anyone else desires the book in a different format, I will gladly let them deal with issues of source code translation. Not that this should be a terrible problem for anyone technically competent in markup languages, as LaTeX source is rather easy to work with.
- . . . use a GNU GPL-style copyleft license. Instead, I will use the Creative Commons Attribution-only license, which is far more permissive for anyone wishing to incorporate my work into derivative works. My interest is maximum flexibility for those who may adapt my material to their own needs, not the imposition of certain philosophical ideals.
- . . . start from a conceptual state of "ground zero." I will assume the reader has certain familiarity with electronics and mathematics, which I will build on. If a reader finds they need to learn more about electronics, they should go read *Lessons In Electric Circuits*.
- . . . avoid using calculus to help explain certain concepts. Not all my readers will understand these parts, and so I will be sure to explain what I can without using calculus. However, I want to give my more mathematically adept students an opportunity to see the power of calculus applied to instrumentation where appropriate. By occasionally applying calculus and explaining my steps, I also hope this text will serve as a practical guide for students who might wish to learn calculus, so they can see its utility and function in a context that interests them.

There do exist many fine references on the subject of industrial instrumentation. I only wish I could condense their best parts into a single volume for my students. Being able to do so would certainly save me from having to write my own! Listed here are some of the best books I can recommend for those wishing to explore instrumentation outside of my own presentation:

- *Instrument Engineers' Handbook* series (Volumes I, II, and III), edited by Béla Lipták – by far my favorite modern references on the subject. Unfortunately, there is a fair amount of material within that lies well beyond my students' grasp (Laplace transforms, etc.), and the volumes are incredibly bulky and expensive (nearly 2000 pages, and at a cost of nearly $200.00, *apiece!*). These texts also lack some of the basic content my students do need, and I don't have the heart to tell them to buy yet *another* textbook to fill the gaps.
- *Handbook of Instrumentation and Controls*, by Howard P. Kallen. Perhaps the best-written textbook on general instrumentation I have ever encountered. Too bad it is both long out of print – my copy dates 1961 – and technologically dated. Like most American textbooks written during the years immediately following Sputnik, it is a masterpiece of practical content and conceptual clarity. I consider books like this useful for their presentations of "first principles," which of course are timeless.
- *Industrial Instrumentation Fundamentals*, by Austin E. Fribance. Another great post-Sputnik textbook – my copy dates 1962.

- *Instrumentation for Process Measurement and Control*, by Norman A. Anderson. An inspiring effort by someone who knows the art of teaching as well as the craft of instrumentation. Too bad the content doesn't seem to have been updated since 1980.
- Practically anything written by Francis Greg Shinskey.

Whether or not I achieve my goal of writing a better textbook is a judgment left for others to make. One decided advantage my book will have over all the others is its *openness*. If you don't like anything you see in these pages, you have the right to modify it to your liking! Delete content, add content, modify content – it's all fair game thanks to the Creative Commons licensing. My only condition is declared in the license: you must give me credit for my original authorship. What you do with it beyond that is wholly up to you[2]. This way, perhaps I can spare someone else from having to write their own textbook from scratch!

[2]This includes selling copies of it, either electronic or print. Of course, you must include the Creative Commons license as part of the text you sell (see Section 4, subsection 1 of the license for details), which means anyone will be able to tell it is an open text and can probably figure out how to download an electronic copy off the internet for free. The only way you're going to make significant money selling this text is to add your own value to it, either in the form of expansions or bundled product (e.g. simulation software, learning exercises, etc.), which of course is perfectly fair – you must profit from your *own* labors. All my work does for you is give you a starting point.

Chapter 1

Mathematics

Mathematics is the investigation of an artificial world: a universe populated by abstract entities and rigid rules governing those entities. Mathematicians devoted to the study and advancement of pure mathematics have an extremely well-developed respect for these rules, for the integrity of this artificial world depends on them. In order to preserve the integrity of their artificial world, their collective work must be *rigorous*, never allowing for sloppy handling of the rules or allowing intuitive leaps to be left unproven.

However, many of the tools and techniques developed by mathematicians for their artificial world happen to be extremely useful for understanding the real world in which we live and work, and therein lies a problem. In applying mathematical rules to the study of real-world phenomena, we often take a far more pragmatic approach than any mathematician would feel comfortable with.

The tension between pure mathematicians and those who apply math to real-world problems is not unlike the tension between linguists and those who use language in everyday life. All human languages have rules (though none as rigid as in mathematics!), and linguists are the guardians of those rules, but the vast majority of human beings play fast and loose with the rules as they use language to describe and understand the world around them. Whether or not this "sloppy" adherence to rules is good depends on which camp you are in. To the purist, it is offensive; to the pragmatist, it is convenient.

I like to tell my students that mathematics is very much like a language. The more you understand mathematics, the larger "vocabulary" you will possess to describe principles and phenomena you encounter in the world around you. Proficiency in mathematics also empowers you to grasp relationships between different things, which is a powerful tool in learning new concepts.

This book is not written for (or by!) mathematicians. Rather, it is written for people wishing to make sense of industrial process measurement and control. This chapter of the book is devoted to a very pragmatic coverage of certain mathematical concepts, for the express purpose of applying these concepts to real-world systems.

Mathematicians, cover your eyes for the rest of this chapter!

1.1 Introduction to calculus

Few areas of mathematics are as powerfully useful in describing and analyzing as calculus: the mathematical study of *changes*. Calculus also happens to be tremendously confusing to most students first encountering it. A great deal of this confusion stems from mathematicians' insistence on rigor and denial of intuition.

Look around you right now. Do you see any mathematicians? If not, good – you can proceed in safety. If so, find another location to begin reading the rest of this chapter. I will frequently appeal to practical example and intuition in describing the basic principles of single-variable calculus, for the purpose of expanding your mathematical "vocabulary" to be able to describe and better understand phenomena of change related to instrumentation.

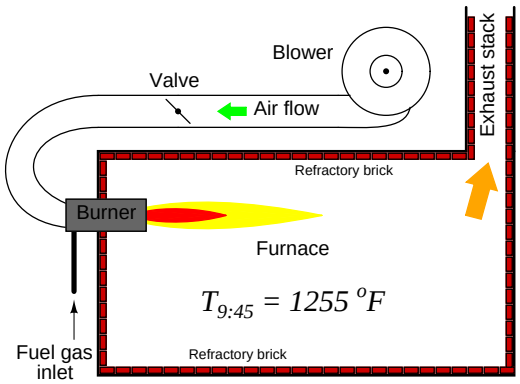
Silvanus P. Thompson, in his wonderful book *Calculus Made Simple* originally published in 1910, began his text with a short chapter entitled, "To Deliver You From The Preliminary Terrors[1]." I will follow his lead by similarly introducing you to some of the notations frequently used in calculus, along with very simple (though not mathematically rigorous) definitions.

[1] The book's subtitle happens to be, *Being a very-simplest introduction to those beautiful methods of reckoning which are generally called by the terrifying names of the differential calculus and the integral calculus.* Not only did Thompson recognize the anti-pragmatic tone with which calculus is too often taught, but he also infused no small amount of humor in his work.

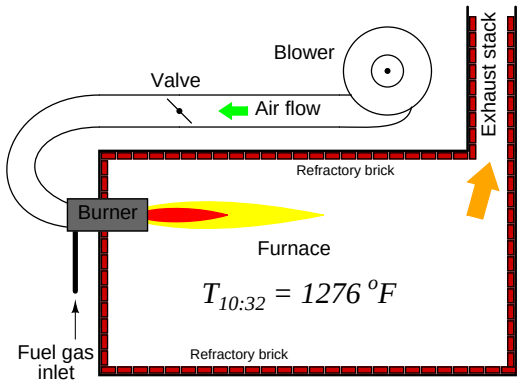
When we wish to speak of a change in some variable's value (let's say x), it is common to precede the variable with the capital Greek letter "delta" as such:

$$\Delta x = \text{“Change in } x\text{”}$$

For example, if the temperature of a furnace (T) increases over time, we might wish to describe that change in temperature as ΔT:



Temperature of furnace at 9:45 AM = 1255 °F



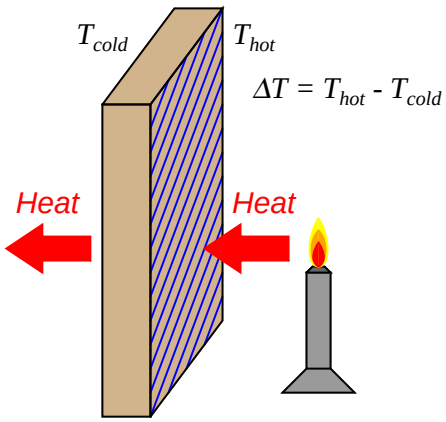
Temperature of furnace at 10:32 AM = 1276 °F

$$\Delta T = T_{10:32} - T_{9:45}$$

$$\Delta T = 1276\ ^{o}F - 1255\ ^{o}F = 21\ ^{o}F$$

The value of ΔT is nothing more than the difference (subtraction) between the recent temperature and the older temperature. A rising temperature over time thus yields a positive value for ΔT, while a falling temperature over time yields a negative value for ΔT.

We could also describe differences between the temperature of two *locations* (rather than a difference of temperature between two *times*) by the notation ΔT, such as this example of heat transfer through a heat-conducting wall where one side of the wall is hotter than the other:



Once again, ΔT is calculated by subtracting one temperature from another. Here, the sign (positive or negative) of ΔT denotes the *direction* of heat flow through the thickness of the wall.

One of the major concerns of calculus is changes or differences between variable values lying *very close to each other*. In the context of a heating furnace, this could mean increases in temperature over miniscule time intervals. In the context of heat flowing through a wall, this could mean differences in temperature sampled between points within the wall immediately next to each other. For such applications, we use a different notation instead of the capital Greek letter delta (Δ); instead, we use a lower-case Roman letter d (or in some cases, the lower-case Greek letter delta: δ).

Thus, a change in furnace temperature from one instant in time to the next instant could be expressed as dT (or δT), while a difference in temperature between two adjacent points within the heat-conducting wall could also be expressed as dT (or δT). We even have a unique name for this concept of extremely small differences: whereas ΔT is called a *difference* in temperature, dT is called a *differential* of temperature.

The concept of a differential may seem useless to you right now, but they are actually quite powerful for describing *continuous changes*, especially when one differential is related to another differential by ratio (something we call a *derivative*).

Another major concern in calculus is how quantities accumulate, especially how differential (extremely small differences in) quantities accumulate to form a larger whole. If we were concerned with how hot the furnace would become over time (T), we could express its eventual temperature as the accumulation, or sum, of temperature differences measured over time (ΔT). Supposing we measured the furnace's temperature once every minute from 9:45 to 10:32 AM:

$$T = \Delta T_{9:45} + \Delta T_{9:46} + \cdots \Delta T_{10:32} = \text{Accumulated temperature rise over time, from 9:45 to 10:32}$$

A more sophisticated way of expressing the summation of differences is to use the capital Greek letter sigma (meaning "sum of" in mathematics) with notations specifying which temperature differences to sum:

$$T = \sum_{n=9:45}^{10:32} \Delta T_n = \text{Accumulated temperature rise over time, from 9:45 to 10:32}$$

However, if our furnace temperature system sampled at an infinite pace, measuring temperature *differentials* (dT) in rapid succession, we could express the same accumulated temperature rise as a sum of infinitesimal (infinitely small) quantities. Just as we used a different mathematical symbol to represent differentials (d) instead of differences (Δ), we will use a different mathematical symbol to represent the summation of differentials ($\int$) instead of the summation of differences ($\sum$):

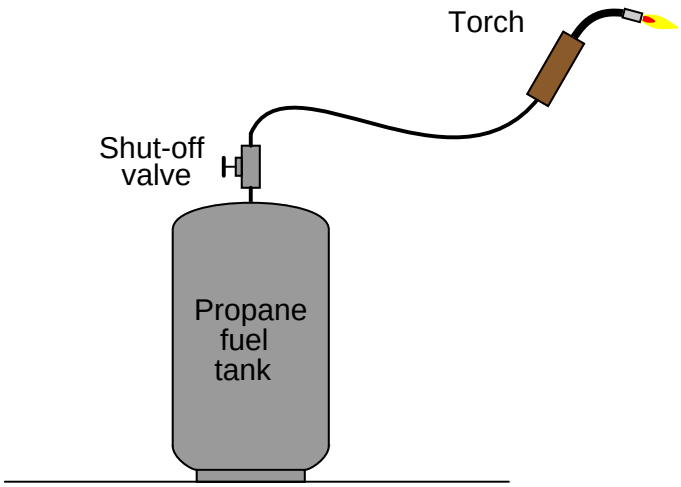
$$T = \int_{9:45}^{10:32} dT = \text{Accumulated temperature rise over time, from 9:45 to 10:32}$$

This summation of infinitesimal quantities is called *integration*, and the elongated "S" symbol ($\int$) is the *integral* symbol.

These are the two major ideas in calculus: *differentials* and *integrals*, and the notations used to represent each. Now that wasn't so frightening, was it?

1.2 The concept of differentiation

Suppose we wished to measure the rate of propane gas flow through a hose to a torch:



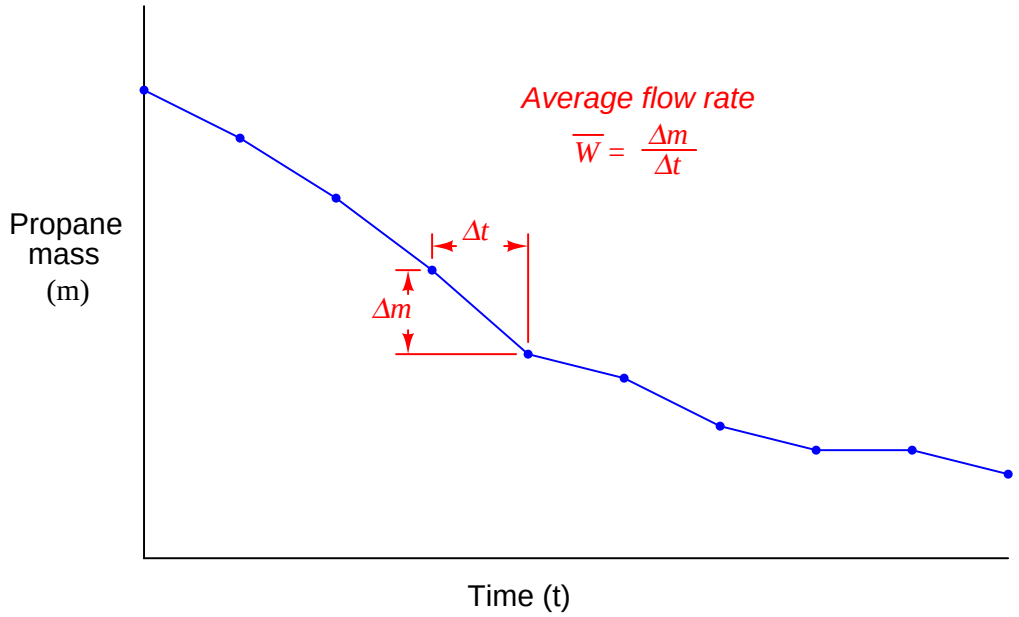
Flowmeters appropriate for measuring low flow rates of propane gas are quite expensive, and so it would be challenging to measure the flow rate of propane fuel gas consumed by the torch at any given moment. We could, however, *indirectly* measure the flow rate of propane by placing the tank on a scale where its mass (m) could be monitored over time. By taking measurements of mass between intervals of time (Δt), we could calculate the corresponding differences in mass (Δm), then calculate the ratio of mass lost over time to calculate average mass flow rate ($\overline{W}$) between those time intervals:

$$\overline{W} = \frac{\Delta m}{\Delta t} = \text{Average mass flow rate}$$

Where,
$\overline{W}$ = Average mass flow rate within each time interval (kilograms per minute)
Δm = Mass difference over time interval (kilograms)
Δt = Time interval (minutes)

Note that flow rate is a ratio (quotient) of mass change over time change. The units used to express flow even reflect this process of division: kilograms *per* minute.

Graphed as a function over time, the tank's mass will be seen to decrease as time elapses. Each dot represents a mass and time measurement coordinate pair (e.g. 20 kilograms at 7:38, 18.6 kilograms at 7:51, etc.):

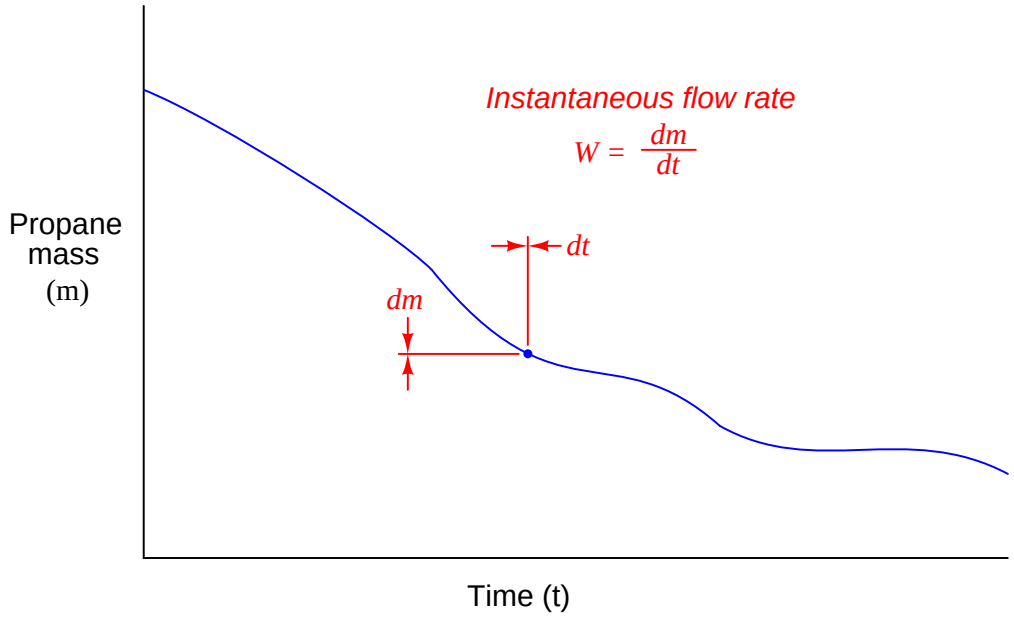


We should recall from basic geometry that the slope of a line is defined as its *rise* (vertical interval) divided by its *run* (horizontal interval). Thus, the average mass flow rate calculated within each time interval may be represented as the pitch (slope) of the line segments connecting dots, since mass flow rate is defined as a change in mass per (divided by) change in time.

Intervals of high propane flow (large flame from the torch) show up on the graph as steeply-pitched line segments. Intervals of no propane flow reveal themselves as flat portions on the graph (no rise or fall over time).

If the determination of average flow rates between significant gaps in time is good enough for our application, we need not do anything more. However, if we would like to be able to infer mass flow rate at any particular *instant* in time, we need to perform the same measurements of mass loss, time elapse, and division of the two at an infinitely fast rate.

Supposing such a thing were possible, what we would end up with is a smooth graph showing mass consumed over time. Instead of a few line segments roughly approximating a curve, we would have an *infinite* number of infinitely short line segments connected together to form a seamless curve. The flow rate at any particular point in time would be the ratio of the mass and time differentials (the slope of the infinitesimal line segment) at that point:



$$W = \frac{dm}{dt} = \text{Instantaneous mass flow rate}$$

Where,

W = Instantaneous mass flow rate at a given time (kilograms per minute)
Δm = Mass differential at a given time (kilograms)
Δt = Time differential at a given time (minutes)

Flow is calculated just the same as before: a quotient of mass and time intervals, except here the intervals are infinitesimal in magnitude. The unit of flow measurement reflects this process of division, just as before, with mass flow rate expressed in units of kilograms *per* minute.

Such a ratio of differential quantities is called a *derivative* in calculus[2]. Derivatives – especially time-based derivatives such as flow rate – find many applications in instrumentation as well as the general sciences. Some of the most common time-based derivative functions include the relationships between *position* (x), *velocity* (v), and *acceleration* (a).

[2]Isaac Newton referred to derivatives as *fluxions*, and in Silvanus Thompson's day they were known as *differential coefficients*.

Velocity is the rate at which an object changes position over time. Since position is typically denoted by the variable x and time by the variable t, the derivative of position with respect to time may be written as such:

$$v = \frac{dx}{dt}$$

The units of measurement for velocity (meters per second, miles per hour, etc.) betray this process of division: a differential of position (meters) divided by a differential of time (second).

Acceleration is the rate at which an object changes velocity over time. Thus, we may express acceleration as the time-derivative of velocity, just as velocity was expressed as the time-derivative of position:

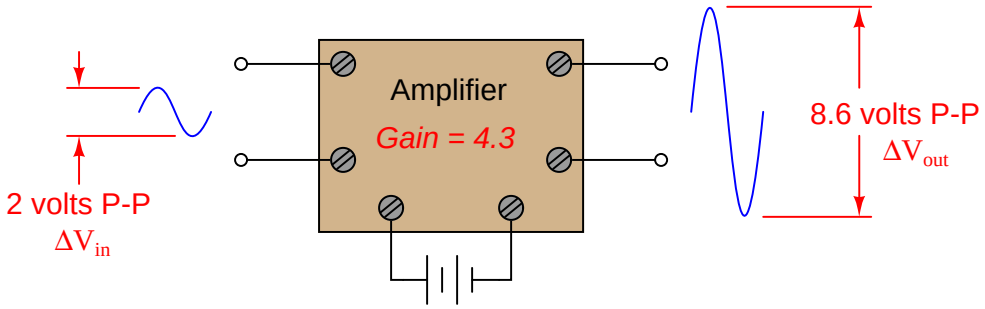
$$a = \frac{dv}{dt}$$

We may even express acceleration as a function of position (x), since it is the rate of change of the rate of change in position over time. This is known as a *second derivative*, since it is applying the process of "differentiation" twice:

$$a = \frac{d^2x}{dt^2}$$

As with velocity, the units of measurement for acceleration (meters per second squared, or alternatively meters per second per second) betray a compounded quotient.

It is also possible to express rates of change between different variables not involving time. A common example in the engineering realm is the concept of *gain*, generally defined as the ratio of output change to input change. An electronic amplifier, for example, with an input signal of 2 volts (peak-to-peak) and an output signal of 8.6 volts (peak-to-peak), would be said to have a gain of 4.3, since the change in output measured in peak-to-peak volts is 4.3 times larger than the corresponding change in input voltage:



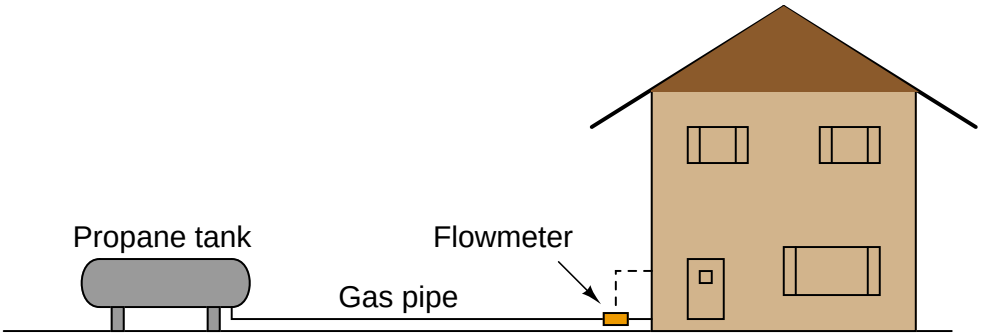
This gain could be expressed as a quotient of differences ($\frac{\Delta V_{out}}{\Delta V_{in}}$), or it could be expressed as a derivative instead:

$$\text{Gain} = \frac{dV_{out}}{dV_{in}}$$

If the amplifier's behavior is perfectly linear, there will be no difference between gain calculated using differences and gain calculated using differentials (the derivative), since the average slope of a straight line is the same as the instantaneous slope at any point along that line. If, however, the amplifier does not behave in a perfectly linear fashion, gain calculated from large changes in voltage ($\frac{\Delta V_{out}}{\Delta V_{in}}$) will not be the same as gain calculated from infinitesimal changes at different points along the amplifier's operating voltage range.

1.3 The concept of integration

Suppose we wished to measure the loss of mass over time in a large propane storage tank supplying a building with heating fuel, because the tank lacked a level indicator to show how much fuel was left at any given time. The flow rate is sufficiently large, and the task sufficiently important, to justify the installation of a mass flowmeter[3], which registers flow rate at an indicator inside the building:



By measuring true mass flow rate, it should be possible to indirectly measure how much propane has been used at any time following the most recent filling of the tank. For example, if the mass flow rate of propane into the building was measured to be an average of 5 kilograms per hour for 30 hours, it would be a simple matter of multiplication to arrive at the consumed mass:

$$\left(\frac{5 \text{ kg}}{\text{hr}}\right)\left(\frac{30 \text{ hrs}}{1}\right) = 150 \text{ kg}$$

Expressing this mathematically as a function of differences (in mass as well as time), we may write the following equation:

$$\Delta m = \overline{W}\,\Delta t$$

Where,
$\overline{W}$ = Average mass flow rate within the time interval (kilograms per hour)
Δm = Mass difference over time interval (kilograms)
Δt = Time interval (hours)

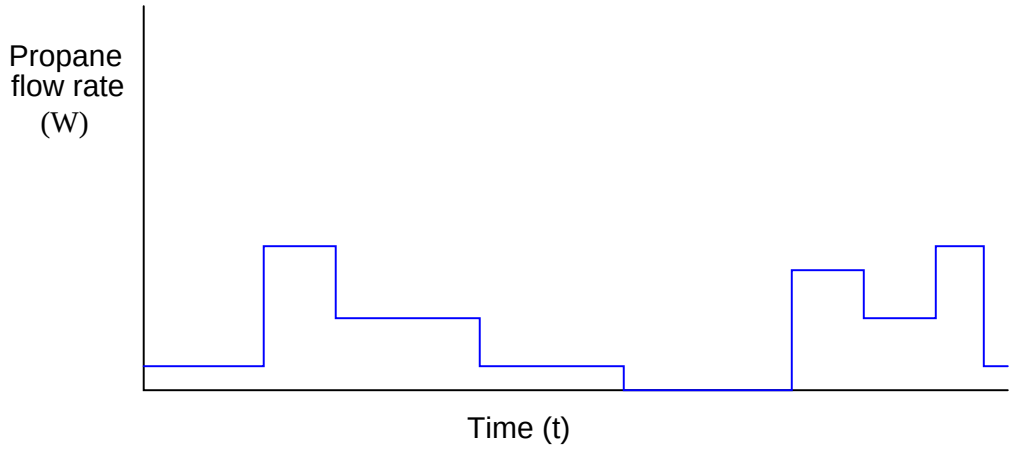
It is easy to see how this equation is nothing more than the quotient-of-differences equation used in the differential calculus section to define mass flow rate:

$$\overline{W} = \frac{\Delta m}{\Delta t} = \text{Average mass flow rate}$$

Inferring mass flow rate from changes in mass over intervals of time is a process of *division*. Inferring changes in mass from flow rate over time is a process of *multiplication*. The units of measurement used to express each of the variables makes this quite clear.

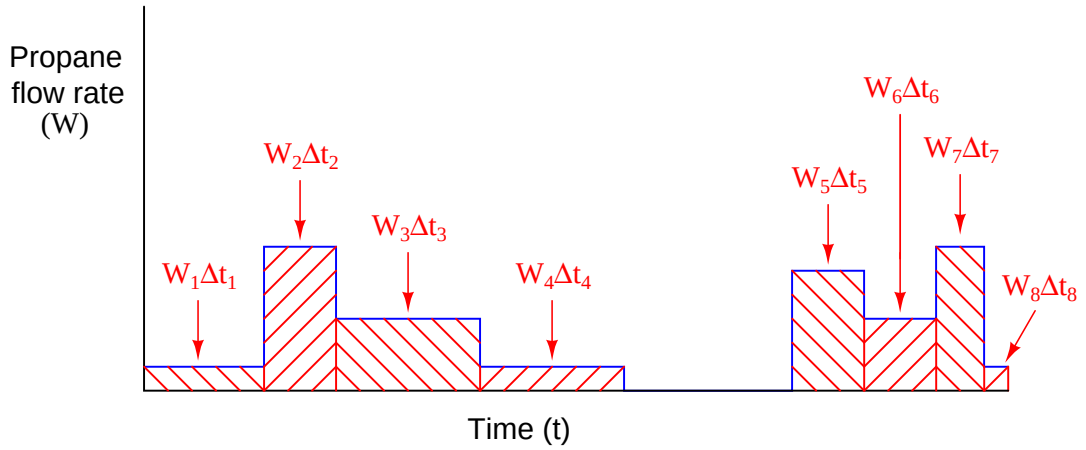
[3] Most likely a thermal mass flowmeter or a Coriolis flowmeter.

The task of inferring lost mass over time becomes much more complicated if the flow rate changes substantially over time. Consider the following graph, showing periods of increased and decreased flow rate due to different gas-fired appliances turning on and off inside the building:



Here, the propane gas flow rate does not stay constant throughout the entire time interval covered by the graph. This obviously complicates the task of calculating total propane mass used over that time.

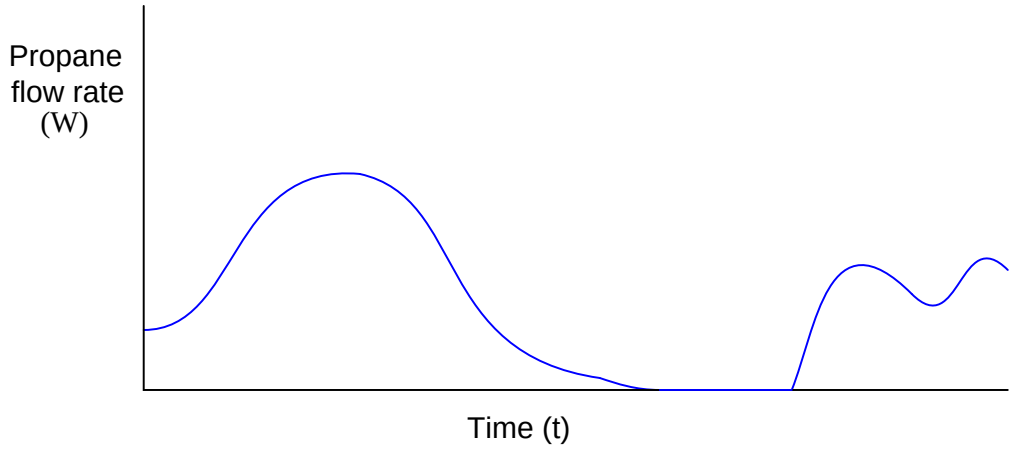
In order to accurately calculate the amount of propane mass consumed by the building over time, we must treat each period of constant flow as its own interval, calculating the mass lost in each interval, then summing those mass differences to arrive at a total mass for the entire time period covered by the graph. Since we know the difference (loss) in mass over a time interval is equal to the average flow rate for that interval multiplied by the interval time length ($\Delta m = W\,\Delta t$), we may calculate each interval's mass as an *area* underneath the graph line, each rectangular area being equal to height (W) times width (Δt):



Each rectangular area underneath the flow line on the graph ($W\,\Delta t$) represents a quantity of propane gas consumed in that time interval. To find the total amount of propane consumed in the time represented by the entire graph, we must sum these mass quantities. This sum may be mathematically expressed using the capital Greek letter sigma, summing repeated products (multiplication) of mass flow and time intervals:

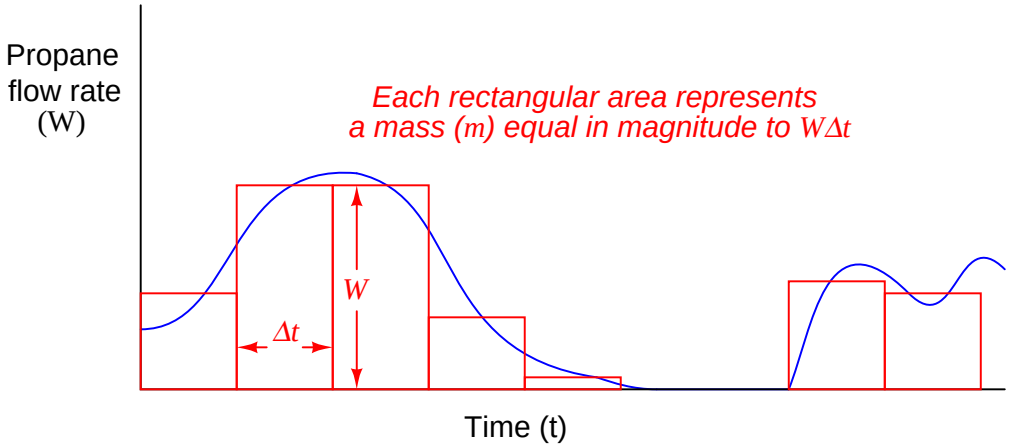
$$\sum_{n=1}^{8} W\,\Delta t_n$$

The task of inferring total propane mass consumed over time becomes even more complicated if the flow does not vary in stair-step fashion as it did in the previous example. Suppose the building were equipped with *throttling* gas appliances instead of on/off gas appliances, thus creating a continuously variable flow rate demand over time. A typical flow rate graph might look something like this:

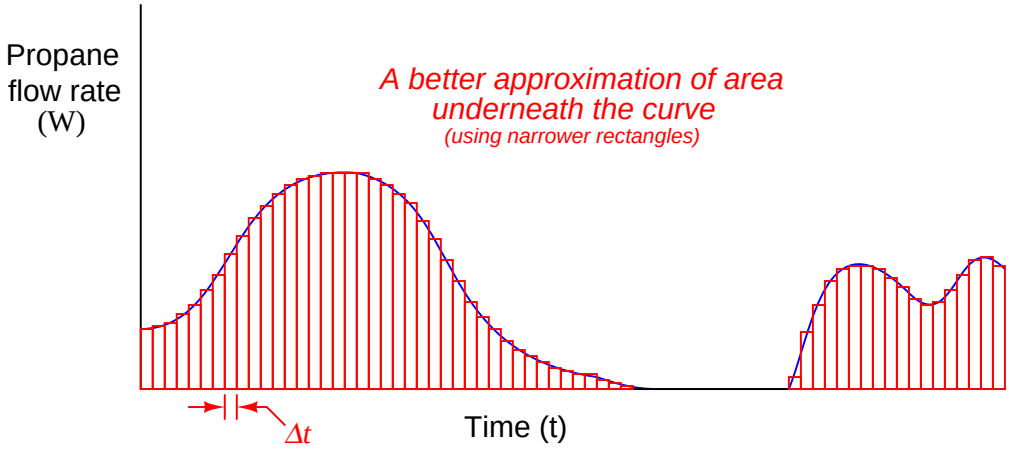


The physics of gas flow and gas mass over time has not changed: total propane mass consumed over time will still be the area contained underneath the flow curve on the graph. However, arbitrary curve shapes do not lend themselves well to calculation of geometric areas.

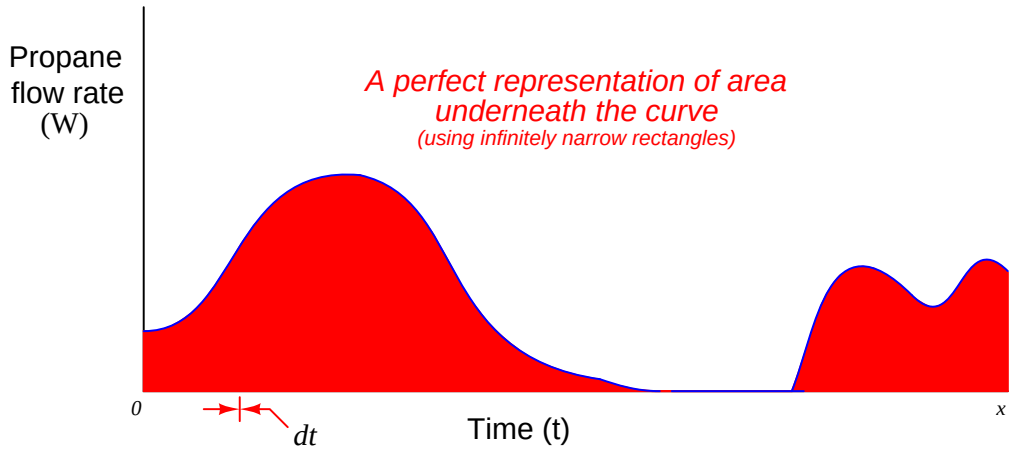
We can, however, approximate the area underneath this curve by overlaying a series of rectangles, the area of each rectangle being height (W) times width (Δt):



It should be intuitively evident that the strategy of using rectangles to approximate the area underneath a curve improves with the number of rectangles used. The more rectangles (the narrower each rectangle), the better approximation of area we will obtain:



Taking this idea to its ultimately realization, we could imagine a super-computer sampling mass flow rates at an infinite speed, then calculating the rectangular area covered by each flow rate (W) times each infinitesimal interval of time (dt). With time intervals of negligible width, the "approximation" of area underneath the graph found by the sum of all these rectangles would be perfect – indeed, it would not be an approximation at all:



If we represent infinitesimal intervals of time by the notation "dt" as opposed to the notation "Δt" used to represent discrete intervals of time, we must also use different notation to represent the mathematical sum of those quantities. Thus, we will dispense with the "sigma" symbol ($\sum$) for summation and replace it with the integral symbol ($\int$), which means a *continuous* summation of infinitesimal quantities:

$$\Delta m = \int_0^x W \, dt$$

This equation tells us the total change in mass (Δt) from time 0 to time x is equal to the continuous sum of all products (multiplication) of mass flow rate (W) over infinitesimal intervals of time (dt).

An extremely important detail to note is that this process of integration (multiplying flow rates by infinitesimal time intervals, then summing those products) only tells us how much propane mass was consumed – it does *not* tell us how much propane is left in the tank, which was the purpose of installing the mass flowmeter and performing all this math! The integral of mass flow and time ($\int W \, dt$) will always be a negative quantity[4], because a flow of propane gas out of the tank represents a *loss* of propane mass within the tank. In order to calculate the amount of propane mass left in the tank, we would need to know the initial value of propane in the tank before any of it flowed to

[4] Although we will measure time, and differentials of time, as positive quantities, the mass flowmeter should be configured to show a negative flow rate (W) when propane flows from the tank to the building. This way, the *integrand* (the product "inside" the integration symbol; $W \, dt$) will be a negative quantity, and thus the integral over a positive time interval (from 0 to x) will likewise be a negative quantity.

the building, then we would add this initial mass quantity (m_0) to the negative mass loss calculated by integration.

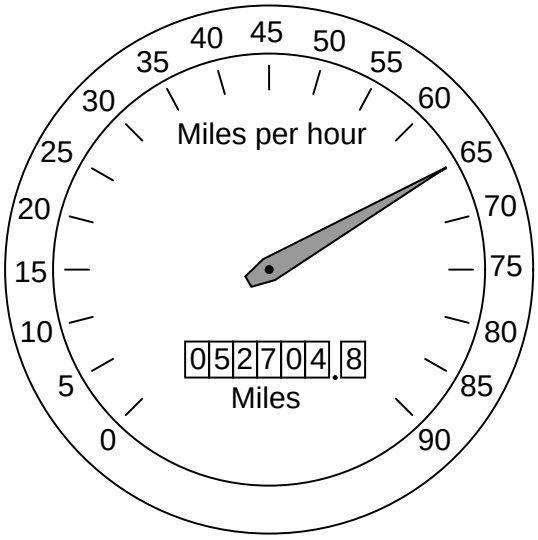
Thus, we would mathematically express the propane mass inside the tank at time x as such[5]:

$$m_x = \int_0^x W\, dt + m_0$$

This initial value must always be considered in problems of integration if we attempt to absolutely define some integral quantity. Otherwise, all the integral will yield is a relative quantity (how much something has *changed* over an interval).

The problem of initial values is very easy to relate to common experience. Consider the *odometer* indication in an automobile. This is an example of an integral function, the distance traveled (x) being the time-integral of speed (or velocity, v):

$$\Delta x = \int v\, dt$$



Although the odometer does accumulate to larger and larger values as I drive the automobile, its indication does not necessarily tell me how many miles I have driven it. If, for example, I purchased the automobile with 32,411.6 miles on the odometer, its current indication of 52,704.8 miles means that I have driven it 20,293.2 miles. The automobile's *total* distance traveled since manufacture is equal to the distance I have accumulated while driving it ($\int v\, dt$) *plus* the initial mileage accumulated at the time I took ownership of it (x_0):

$$x_{total} = \int v\, dt + x_0$$

[5] According to calculus convention, the differential dt represents the end of the integrand. This tells us m_0 is not part of the integrand, but rather comes after it. Using parentheses to explicitly declare the boundaries of the integrand, we may re-write the expression as $\int_0^x (W\, dt) + m_0$

1.4 How derivatives and integrals relate to one another

To review, let us consider some of the properties of *derivatives*:

- A derivative is always a quotient of two differential quantities – it is fundamentally a process of *division*
- The units of measurement for a derivative always reflect this process of division
- Geometrically, the derivative of a function is its *slope* on a graph

Let us also consider some of the properties of *integrals*:

- An integral is always a product of some variable and a differential quantity – it is fundamentally a process of *multiplication*
- The units of measurement for an integral always reflect this process of multiplication
- Geometrically, the integral of a function is the *area* bounded by a graph

Just as division and multiplication are *inverse* mathematical functions (i.e. one "un-does" the other), differentiation and integration are also inverse mathematical functions. The two examples of propane gas flow and mass measurement highlighted in the previous sections illustrates this complementary relationship. We may use differentiation with respect to time to convert a mass measurement (m) into a mass flow measurement (W, or $\frac{dm}{dt}$). Conversely, we may use integration with respect to time to convert a mass flow measurement (W, or $\frac{dm}{dt}$) into a measurement of mass gained or lost (Δm).

Likewise, the common examples of position (x), velocity (v), and acceleration (a) used to illustrate the principle of differentiation are also related to one another by the process of integration. Reviewing the derivative relationships:

$$v = \frac{dx}{dt} \qquad \text{Velocity is the derivative of position with respect to time}$$

$$a = \frac{dv}{dt} \qquad \text{Acceleration is the derivative of velocity with respect to time}$$

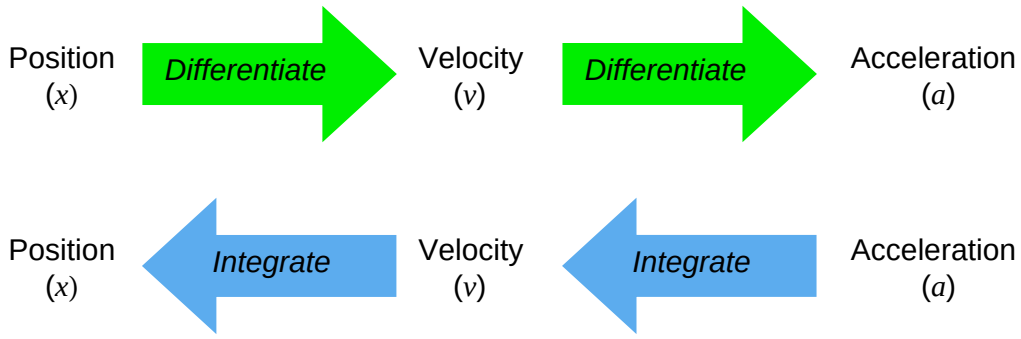
Now, expressing position and velocity as *integrals* of velocity and acceleration, respectively[6]:

$$x = \int v\, dt \qquad \text{Position is the integral of velocity with respect to time}$$

$$v = \int a\, dt \qquad \text{Velocity is the integral of acceleration with respect to time}$$

[6] To be perfectly accurate, we must also include initial values for position and velocity. In other words, $x = \int v dt + x_0$ and $v = \int a\, dt + v_0$

Differentiation and integration may be thought of as processes transforming these quantities into one another:



It is a relatively simple matter to build a computer (either analog or digital) capable of applying differentiation or integration to a real-world signal, which means these calculus techniques afford us the opportunity to infer multiple variables from a single measured variable. The measured position of a machine, for example, may be differentiated by a computer to yield the machine's velocity, and that velocity signal differentiated again to yield acceleration. Conversely, a signal taken from an accelerometer may be integrated to yield a velocity signal, and integrated again to yield a position signal. To be sure, there are practical limits to these processes[7], but they are at least possible.

References

Keisler, H. Jerome, *Elementary Calculus – An Infinitesimal Approach*, Second Edition, University of Wisconsin, 2000.

Stewart, James, *Calculus: Concepts and Contexts*, 2nd Edition, Brooks/Cole, Pacific Grove, CA, 2001.

Thompson, Silvanus P. and Gardner, Martin, *Calculus Made Easy*, St. Martin's Press, New York, NY, 1998.

[7] The major problem facing differentiation of real-world signals is *noise*. Noise superimposed on any measurement signal will be interpreted by a differentiator circuit as extremely high rates of change, thus becoming amplified at the output of that differentiator. Integration has its own unique problem: offset. The calibration of any sensor whose signal is to be integrated must be nearly perfect, for if the sensor's signal is offset by any amount at all (e.g. the sensor produces a slight signal when it should output no signal) the integrator circuit will continue to integrate this offset over time, producing an output that slowly ramps until saturation is reached.

Chapter 2

Physics

2.1 Terms and Definitions

Mass (m) is the opposition an object has to acceleration (changes in velocity). *Weight* is the force (F) imposed on a mass by a gravitational field. Mass is an intrinsic property of an object, regardless of the environment. Weight, on the other hand, depends on the strength of the gravitational field in which the object resides. A 20 kilogram slug of metal has the exact same mass whether it rests on Earth, or in the zero-gravity environment of outer space, or on the surface of the planet Jupiter. However, the *weight* of that mass depends on gravity: zero weight in outer space (where there is no gravity to act upon it), some weight on Earth, and a much greater amount of weight on the planet Jupiter (due to the much stronger gravitational field of that planet).

Since mass is the opposition of an object to changes in velocity (acceleration), it stands to reason force, mass, and acceleration for any particular object are directly related to one another:

$$F = ma$$

Where,
F = Force in newtons (metric) or pounds (British)
m = Mass in kilograms (metric) or slugs (British)
a = Acceleration in meters per second squared (metric) or feet per second squared (British)

If the force in question is the weight of the object, then the acceleration (a) in question is the acceleration constant of the gravitational field where the object resides. For Earth at sea level, $a_{gravity}$ is approximately 9.8 meters per second squared, or 32 feet per second squared. Earth's gravitational acceleration constant is usually represented in equations by the variable letter g instead of the more generic a.

Since acceleration is nothing more than the rate of velocity change with respect to time, the force/mass equation may be expressed using the calculus notation of the first derivative:

$$F = m\frac{dv}{dt}$$

Where,
F = Force in newtons (metric) or pounds (British)
m = Mass in kilograms (metric) or slugs (British)
v = Velocity in meters per second (metric) or feet per second (British)
t = Time in seconds

Since velocity is nothing more than the rate of position change with respect to time, the force/mass equation may be expressed using the calculus notation of the second derivative (acceleration being the derivative of velocity, which in turn is the derivative of position):

$$F = m\frac{d^2x}{dt^2}$$

Where,
F = Force in newtons (metric) or pounds (British)
m = Mass in kilograms (metric) or slugs (British)
x = Position in meters (metric) or feet (British)
t = Time in seconds

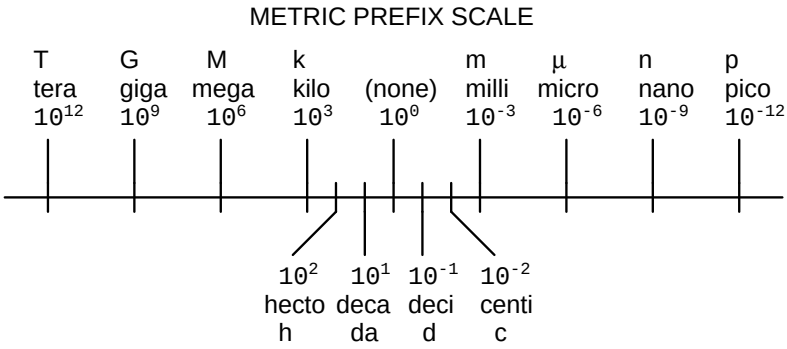
Mass density (ρ) for any substance is the proportion of mass to volume. *Weight density* (γ) for any substance is the proportion of weight to volume.

Just as weight and mass are related to each other by gravitational acceleration, weight density and mass density are also related to each other by gravity:

$$F_{weight} = mg \qquad \text{Weight and Mass}$$

$$\gamma = \rho g \qquad \text{Weight density and Mass density}$$

2.2 Metric prefixes



2.3 Unit conversions and physical constants

Converting between disparate units of measurement is the bane of many science students. The problem is worse for students of industrial instrumentation in the United States of America, who must work with British ("Customary") units such as the pound, the foot, the gallon, etc. Worldwide adoption of the metric system would go a long way toward alleviating this problem, but until then it is important for students of instrumentation to master the art of unit conversions[1].

It is possible to convert from one unit of measurement to another by use of tables designed expressly for this purpose. Such tables usually have a column of units on the left-hand side and an identical row of units along the top, whereby one can look up the conversion factor to multiply by to convert from any listed unit to any other listed unit. While such tables are undeniably simple to use, they are practically impossible to memorize.

The goal of this section is to provide you with a more powerful technique for unit conversion, which lends itself much better to memorization of conversion factors. This way, you will be able to convert between many common units of measurement while memorizing only a handful of essential conversion factors.

I like to call this the *unity fraction* technique. It involves setting up the original quantity as a fraction, then multiplying by a series of fractions having *physical* values of unity (1) so that by multiplication the original value does not change, but the units do. Let's take for example the conversion of quarts into gallons, an example of a fluid volume conversion:

$$35 \text{ qt} = ??? \text{ gal}$$

Now, most people know there are four quarts in one gallon, and so it is tempting to simply divide the number 35 by four to arrive at the proper number of gallons. However, the purpose of this example is to show you how the technique of unity fractions works, not to get an answer to a problem. First, we set up the original quantity as a fraction, in this case a fraction with 1 as the denominator:

$$\frac{35 \text{ qt}}{1}$$

Next, we multiply this fraction by another fraction having a *physical* value of unity, or 1. This means a fraction comprised of equal measures in the numerator and denominator, but with different units of measurement, arranged in such a way that the undesired unit cancels out leaving only the desired unit(s). In this particular example, we wish to cancel out quarts and end up with gallons, so we must arrange a fraction consisting of quarts and gallons having equal quantities in numerator and denominator, such that quarts will cancel and gallons will remain:

$$\left(\frac{35 \text{ qt}}{1}\right)\left(\frac{1 \text{ gal}}{4 \text{ qt}}\right)$$

[1] An interesting point to make here is the United States did get something right when they designed their monetary system of dollars and cents. This is essentially a *metric* system of measurement, with 100 cents per dollar. The founders of the USA wisely decided to avoid the utterly confusing denominations of the British, with their pounds, pence, farthings, shillings, etc. The denominations of penny, dime, dollar, and eagle ($10 gold coin) comprised a simple power-of-ten system for money. Credit goes to France for first adopting a metric system of general weights and measures as their national standard.

Now we see how the unit of "quarts" cancels from the numerator of the first fraction and the denominator of the second ("unity") fraction, leaving only the unit of "gallons" left standing:

$$\left(\frac{35 \text{ qt}}{1}\right)\left(\frac{1 \text{ gal}}{4 \text{ qt}}\right) = 8.75 \text{ gal}$$

The reason this conversion technique is so powerful is it allows one to do a large range of unit conversions while memorizing the smallest possible set of conversion factors.

Here is a set of six equal volumes, each one expressed in a different unit of measurement:

1 gallon (gal) = 231.0 cubic inches (in^3) = 4 quarts (qt) = 8 pints (pt) = 128 fluid ounces (fl. oz.) = 3.7854 liters (l)

Since all six of these quantities are physically equal, it is possible to build a "unity fraction" out of any two, to use in converting any of the represented volume units into any of the other represented volume units. Shown here are a few different volume unit conversion problems, using unity fractions built only from these factors:

40 gallons converted into fluid ounces:

$$\left(\frac{40 \text{ gal}}{1}\right)\left(\frac{128 \text{ fl. oz}}{1 \text{ gal}}\right) = 5120 \text{ fl. oz}$$

5.5 pints converted into cubic inches:

$$\left(\frac{5.5 \text{ pt}}{1}\right)\left(\frac{231 \text{ in}^3}{8 \text{ pt}}\right) = 158.8 \text{ in}^3$$

1170 liters converted into quarts:

$$\left(\frac{1170 \text{ l}}{1}\right)\left(\frac{4 \text{ qt}}{3.7854 \text{ l}}\right) = 1236 \text{ qt}$$

By contrast, if we were to try to memorize a 6×6 table giving conversion factors between *any two* of six volume units, we would have to commit 30 different conversion factors to memory! Clearly, the ability to set up "unity fractions" is a much more memory-efficient and practical approach.

But what if we wished to convert to a unit of volume measurement other than the six shown in the long equality? For instance, what if we wished to convert 5.5 pints into cubic *feet* instead of cubic *inches*? Since cubic feet is not a unit represented in the long string of quantities, what do we do?

We do know of another equality between inches and feet, though. Everyone should know that there are 12 inches in 1 foot. All we need to do is set up *another* unity fraction in the original problem to convert cubic inches into cubic feet:

5.5 pints converted into cubic feet (*our first attempt!*):

$$\left(\frac{5.5 \text{ pt}}{1}\right)\left(\frac{231 \text{ in}^3}{8 \text{ pt}}\right)\left(\frac{1 \text{ ft}}{12 \text{ in}}\right) = ???$$

Unfortunately, this will not give us the result we seek. Even though $\frac{1 \text{ ft}}{12 \text{ in}}$ is a valid unity fraction, it does not *completely* cancel out the unit of inches. What we need is a unity fraction relating *cubic* feet to *cubic* inches. We can get this, though, simply by *cubing* the $\frac{1 \text{ ft}}{12 \text{ in}}$ unity fraction:

5.5 pints converted into cubic feet (*our second attempt!*):

$$\left(\frac{5.5 \text{ pt}}{1}\right)\left(\frac{231 \text{ in}^3}{8 \text{ pt}}\right)\left(\frac{1 \text{ ft}}{12 \text{ in}}\right)^3$$

Distributing the third power to the interior terms of the last unity fraction:

$$\left(\frac{5.5 \text{ pt}}{1}\right)\left(\frac{231 \text{ in}^3}{8 \text{ pt}}\right)\left(\frac{1^3 \text{ ft}^3}{12^3 \text{ in}^3}\right)$$

Calculating the values of 1^3 and 12^3 inside the last unity fraction, then canceling units and solving:

$$\left(\frac{5.5 \text{ pt}}{1}\right)\left(\frac{231 \text{ in}^3}{8 \text{ pt}}\right)\left(\frac{1 \text{ ft}^3}{1728 \text{ in}^3}\right) = 0.0919 \text{ ft}^3$$

Once again, this unit conversion technique shows its power by minimizing the number of conversion factors we must memorize. We need not memorize how many cubic inches are in a cubic foot, or how many square inches are in a square foot, if we know how many linear inches are in a linear foot and we simply let the fractions "tell" us whether a power is needed for unit cancellation.

A major caveat to this method of converting units is that the units must be *directly proportional* to one another, since this multiplicative conversion method is really nothing more than an exercise in mathematical proportions. Here are some examples (but not an exhaustive list!) of conversions that cannot be performed using the "unity fraction" method:

- Absolute / Gauge pressures, because one scale is *offset* from the other by 14.7 PSI (atmospheric pressure).
- Celsius / Fahrenheit, because one scale is *offset* from the other by 32 degrees.
- Wire diameter / gauge number, because gauge numbers grow smaller as wire diameter grows larger (inverse proportion rather than direct) and because there is no proportion relating the two.
- Power / decibels, because the relationship is logarithmic rather than proportional.

The following subsections give sets of physically equal quantities, which may be used to create unity fractions for unit conversion problems. Note that only those quantities shown in the same line (separated by = symbols) are truly equal to each other, not quantities appearing in different lines!

2.3.1 Conversion formulae for temperature

Note: all of the conversion factors given for temperature are *exact*, not approximations.

^{o}F = (^{o}C)(9/5) + 32
^{o}C = (^{o}F - 32)(5/9)
^{o}R = ^{o}F + 459.67
K = ^{o}C + 273.15

2.3.2 Conversion factors for distance

Note: all of the conversion factors given for distance are *exact*, not approximations.

1 inch (in) = **2.54 centimeters** (cm)
1 foot (ft) = **12 inches** (in)
1 yard (yd) = **3 feet** (ft)
1 mile (mi) = **5280 feet** (ft)

2.3.3 Conversion factors for volume

Note: all conversion factors shown in **bold** type are *exact*, not approximations.

1 gallon (gal) = 231.0 cubic inches (in^3) = **4 quarts** (qt) = **8 pints** (pt) = **16 cups** = **128 fluid ounces** (fl. oz.) = 3.7854 liters (l)

1 milliliter (ml) = **1 cubic centimeter** (cm^3)

2.3.4 Conversion factors for velocity

Note: all conversion factors shown in **bold** type are *exact*, not approximations.

1 mile per hour (mi/h) = **88 feet per minute** (ft/m) = 1.46667 feet per second (ft/s) = 1.60934 kilometer per hour (km/h) = 0.44704 meter per second (m/s) = 0.868976 knot (knot – international)

2.3.5 Conversion factors for mass

1 pound-mass (lbm) = 0.4535924 kilogram (kg) = 0.031081 slugs

2.3.6 Conversion factors for force

1 pound-force (lbf) = 4.448222 newtons (N)

1 kilogram-force (kgf) = **9.80665 newtons** (N)

2.3.7 Conversion factors for area

Note: all conversion factors shown in **bold** type are *exact*, not approximations.

1 acre = 43560 square feet (ft^2) = 4840 square yards (yd^2) = 4046.86 square meters (m^2)

2.3.8 Conversion factors for pressure (either all gauge or all absolute)

Note: all conversion factors shown in **bold** type are *exact*, not approximations.

1 pound per square inch (PSI) = 2.03602 inches of mercury (in. Hg) = 27.6799 inches of water (in. W.C.) = **6.894757 kilo-pascals** (kPa) = **0.06894757 bar**

1 meter of water (m W.C.) = **9.80665 kilo-pascals** (kPa)

2.3.9 Conversion factors for pressure (absolute pressure units only)

Note: all conversion factors shown in **bold** type are *exact*, not approximations.

1 atmosphere (Atm) = 14.7 pounds per square inch absolute (PSIA) = **101.325 kilo-pascals absolute** (kPaA) = **1.01325 bar** = 760 millimeters of mercury absolute (mmHgA) = 760 torr (torr)

2.3.10 Conversion factors for energy or work

1 British thermal unit (Btu – "International Table") = 251.996 calories (cal – "International Table") = 1055.06 joules (J) = 1055.06 watt-seconds (W-s) = 0.293071 watt-hour (W-hr) = 1.05506 x 10^{10} ergs (erg) = 778.169 foot-pound-force (ft-lbf)

2.3.11 Conversion factors for power

Note: all conversion factors shown in **bold** type are *exact*, not approximations.

1 horsepower = 550 foot-pounds per second (ft-lbf/s) = 745.7 watts (W) = 2544.43 British thermal units per hour (Btu/hr) = 0.0760181 boiler horsepower (hp – boiler)

2.3.12 Terrestrial constants

Acceleration of gravity at sea level = 9.806650 meters per second per second (m/s^2) = 32.1740 feet per second per second (ft/s^2)

Atmospheric pressure = 14.7 pounds per square inch absolute (PSIA) = 760 millimeters of mercury absolute (mmHgA) = 760 torr (torr) = 1.01325 bar (bar)

Atmospheric gas concentrations (by volume, not mass):

- Nitrogen = 78.084 %
- Oxygen = 20.946 %
- Argon = 0.934 %
- Carbon Dioxide (CO_2) = 0.033 %
- Neon = 18.18 ppm
- Helium = 5.24 ppm
- Methane (CH_4) = 2 ppm
- Krypton = 1.14 ppm
- Hydrogen = 0.5 ppm
- Nitrous Oxide (N_2O) = 0.5 ppm
- Xenon = 0.087 ppm

Density of dry air at 20^{o}C and 760 torr = 1.204 mg/cm^3 = 1.204 kg/m^3 = 0.075 lb/ft^3 = 0.00235 slugs/ft^3

Absolute viscosity of dry air at 20^{o}C and 760 torr = 0.018 centipoise (cp) = 1.8×10^{-5} Pascal-seconds (Pa·s)

2.3.13 Properties of water

Freezing point at sea level = 32^{o}F = 0^{o}C

Boiling point at sea level = 212^{o}F = 100^{o}C

Density of water at 4^{o}C = 1000 kg/m^3 = 1 g/cm^3 = 1 kg/liter = 62.428 lb/ft^3 = 1.951 slugs/ft^3

Specific heat of water at 14^{o}C = 1.00002 calories/g·^{o}C = 1 BTU/lb·^{o}F = 4.1869 joules/g·^{o}C

Specific heat of ice $\approx$ 0.5 calories/g·^{o}C

Specific heat of steam $\approx$ 0.48 calories/g·^{o}C

Absolute viscosity of water at 20^{o}C = 1.0019 centipoise (cp) = 0.0010019 Pascal-seconds (Pa·s)

Surface tension of water (in contact with air) at 18^{o}C = 73.05 dynes/cm

pH of pure water at 25^o C = 7.0 (*pH scale = 0 to 14*)

2.3.14 Miscellaneous physical constants

Note: all constants shown in **bold** type are *exact*, not approximations. Parentheses show one standard deviation (σ) of uncertainty in the last digits: for example, Avogadro's number given as $6.02214179(30) \times 10^{23}$ means the center value ($6.02214179 \times 10^{23}$) plus or minus $0.00000030 \times 10^{23}$.

Avogadro's number (N_A) = $6.02214179(30) \times 10^{23}$ per mole (mol^{-1})

Boltzmann's constant (k) = $1.3806504(24) \times 10^{-23}$ joules per Kelvin (J/K)

Electronic charge (e) = $1.602176487(40) \times 10^{-19}$ Coulomb (C)

Faraday constant (F) = $9.64853399(24) \times 10^{4}$ Coulombs per mole (C/mol)

Gravitational constant (G) = $6.67428(67) \times 10^{-11}$ cubic meters per kilogram-seconds squared ($\text{m}^3/\text{kg-s}^2$)

Molar gas constant (R) = 8.314472(15) joules per mole-Kelvin (J/mol-K)

Planck constant (h) = $6.62606896(33) \times 10^{-34}$ joule-seconds (J-s)

Stefan-Boltzmann constant (σ) = $5.670400(40) \times 10^{-8}$ Watts per square meter-Kelvin^4 ($\text{W/m}^2\cdot\text{K}^4$)

Velocity of light in a vacuum (c) = **299792458 meters per second** (m/s) = 186,282.4 miles per second (mi/s)

All constants taken from NIST data "Fundamental Physical Constants – Extensive Listing", published 2006.

2.3.15 Weight densities of common materials

All density figures approximate for samples at standard temperature and pressure[2].

Liquids:

- Acetone: $\gamma = 49.4$ lb/ft^3
- Alcohol, ethyl (ethanol): $\gamma = 49.4$ lb/ft^3
- Alcohol, methyl (methanol): $\gamma = 50.5$ lb/ft^3
- Benzene: $\gamma = 56.1$ lb/ft^3
- Butane (liquid): $\gamma = 36.1$ lb/ft^3
- Carbon disulfide: $\gamma = 80.7$ lb/ft^3
- Carbon tetrachloride: $\gamma = 99.6$ lb/ft^3
- Chloroform: $\gamma = 93$ lb/ft^3
- Ethylene glycol (ethanediol): $\gamma = 69.22$ lb/ft^3
- Gasoline: $\gamma = 41$ lb/ft^3 to 43 lb/ft^3
- Glycerin: $\gamma = 78.6$ lb/ft^3
- Isobutane (liquid): $\gamma = 34.8$ lb/ft^3
- Kerosene: $\gamma = 51.2$ lb/ft^3
- Mercury: $\gamma = 849$ lb/ft^3
- Methanol (methyl alcohol): $\gamma = 50.5$ lb/ft^3
- Milk: $\gamma = 64.2$ lb/ft^3 to 64.6 lb/ft^3
- Naphtha, petroleum: $\gamma = 41.5$ lb/ft^3
- Oil, castor: $\gamma = 60.5$ lb/ft^3
- Oil, coconut: $\gamma = 57.7$ lb/ft^3
- Oil, linseed (boiled): $\gamma = 58.8$ lb/ft^3
- Oil, olive: $\gamma = 57.3$ lb/ft^3
- Propane (liquid): $\gamma = 31.2$ lb/ft^3
- Toluene: $\gamma = 54.1$ lb/ft^3

[2]Density figures taken or derived from tables in the *CRC Handbook of Chemistry and Physics*, 64th Edition. Most liquid densities taken from table on page F-3 and solid densities taken from table on page F-1. Some liquid densities taken from tables on pages E-27 through E-31. All temperatures at or near 20^{o}C.

- Turpentine: $\gamma = 54.3$ lb/ft^3
- Water, heavy: $\gamma = 68.97$ lb/ft^3
- Water, light (normal): $\gamma = 62.4$ lb/ft^3
- Water, sea: $\gamma = 63.99$ lb/ft^3

Solids:

- Beryllium: $\gamma = 115.37$ lb/ft^3
- Brass: $\gamma = 524.4$ lb/ft^3
- Calcium: $\gamma = 96.763$ lb/ft^3
- Carbon (diamond): $\gamma = 196.65$ lb/ft^3 to 220.37 lb/ft^3
- Cement (set): $\gamma = 170$ lb/ft^3 to 190 lb/ft^3
- Chromium: $\gamma = 448.86$ lb/ft^3
- Copper: $\gamma = 559.36$ lb/ft^3
- Cork: $\gamma = 14$ lb/ft^3 to 16 lb/ft^3
- Gold: $\gamma = 1178.6$ lb/ft^3
- Ice: $\gamma = 57.2$ lb/ft^3
- Iron: $\gamma = 490.68$ lb/ft^3
- Ivory: $\gamma = 114$ lb/ft^3 to 120 lb/ft^3
- Lead: $\gamma = 708.56$ lb/ft^3
- Leather: $\gamma = 54$ lb/ft^3
- Magnesium: $\gamma = 108.50$ lb/ft^3
- Molybdenum: $\gamma = 638.01$ lb/ft^3
- Quartz: $\gamma = 165$ lb/ft^3
- Rubber (soft): $\gamma = 69$ lb/ft^3
- Rubber (hard): $\gamma = 74$ lb/ft^3
- Salt, rock: $\gamma = 136$ lb/ft^3
- Sugar: $\gamma = 99$ lb/ft^3
- Tar: $\gamma = 66$ lb/ft^3
- Wood, balsa: $\gamma = 7$ lb/ft^3 to 9 lb/ft^3
- Wood, maple: $\gamma = 39$ lb/ft^3 to 47 lb/ft^3

2.4 Dimensional analysis

An interesting parallel to the "unity fraction" unit conversion technique is something referred to in physics as *dimensional analysis*. Performing dimensional analysis on a physics formula means to set it up with units of measurement in place of variables, to see how units cancel and combine to form the appropriate unit(s) of measurement for the result.

For example, let's take the familiar power formula used to calculate power in a simple DC electric circuit:

$$P = IV$$

Where,
P = Power (watts)
I = Current (amperes)
V = Voltage (volts)

Each of the units of measurement in the above formula (watt, ampere, volt) are actually comprised of more fundamental physical units. One watt of power is one joule of energy transferred per second. One ampere of current is one coulomb of electric charge moving by per second. One volt of potential is one joule of energy per coulomb of electric charge. When we write the equation showing these units in their proper orientations, we see that the result (power in watts, or joules per second) actually does agree with the units for amperes and volts because the unit of electric charge (coulombs) cancels out. In dimensional analysis we customarily distinguish unit symbols from variables by using non-italicized letters and surrounding each one with square brackets:

$$P = IV$$

$$[\text{Watts}] = [\text{Amperes}] \times [\text{Volts}] \qquad \text{or} \qquad [\text{W}] = [\text{A}][\text{V}]$$

$$\left[\frac{\text{Joules}}{\text{Seconds}}\right] = \left[\frac{\text{Coulombs}}{\text{Seconds}}\right] \times \left[\frac{\text{Joules}}{\text{Coulombs}}\right] \qquad \text{or} \qquad \left[\frac{\text{J}}{\text{s}}\right] = \left[\frac{\text{C}}{\text{s}}\right]\left[\frac{\text{J}}{\text{C}}\right]$$

Dimensional analysis gives us a way to "check our work" when setting up new formulae for physics- and chemistry-type problems.

2.5 The International System of Units

The very purpose of physics is to quantitatively describe and explain the physical world in as few terms as possible. This principle extends to units of measurement as well, which is why we usually find different units used in science actually defined in terms of more fundamental units. The *watt*, for example, is one joule of energy transferred per second of time. The joule, in turn, is defined in terms of three base units, the kilogram, the meter, and the second:

$$[J] = \frac{[\text{kg}][\text{m}^2]}{[\text{s}^2]}$$

Within the metric system of measurements, an international standard exists for which units are considered fundamental and which are considered "derived" from the fundamental units. The modern standard is called *SI*, which stands for *Système International*. This standard recognizes seven fundamental, or *base* units, from which all others are derived[3]:

Physical quantity	SI unit	SI symbol
Length	meter	m
Mass	kilogram	kg
Time	second	s
Electric current	ampere	A
Temperature	kelvin	K
Amount of substance	mole	mol
Luminous intensity	candela	cd

An older standard existed for base units, in which the *centimeter*, *gram*, and *second* comprised the first three base units. This standard is referred to as the *cgs* system, in contrast to the SI system[4]. You will still encounter some derived cgs units used in instrumentation, including the *poise* and the *stokes* (both used to express fluid viscosity). Then of course we have the *British engineering system* which uses such wonderful[5] units as feet, pounds, and (thankfully) seconds. Despite the fact that the majority of the world uses the metric (SI) system for weights and measures, the British system is sometimes referred to as the *Customary* system.

[3] The only exception to this rule being units of measurement for angles, over which there has not yet been full agreement whether the unit of the *radian* (and its solid counterpart, the *steradian*) is a base unit or a derived unit.

[4] The older name for the SI system was "MKS," representing meters, kilograms, and seconds.

[5] I'm noting my sarcasm here, just in case you are immune to my odd sense of humor.

2.6 Conservation Laws

The *Law of Mass Conservation* states that matter can neither be created nor destroyed. The *Law of Energy Conservation* states that energy can neither be created nor destroyed. However, both mass and energy may change forms, and even change into one another in the case of nuclear phenomena.

Conversion of mass into energy, or of energy into mass, is quantitatively described by Albert Einstein's famous equation:

$$E = mc^2$$

Where,
E = Energy (joules)
m = Mass (kilograms)
c = Speed of light (approximately 3×10^8 meters per second)

Conservation laws find practical context in many areas of science and life, but in the realm of process control we have the principles of *mass balance* and *energy balance* which are direct expressions of these Laws. "Mass balance" refers to the fact that the sum total of mass entering a process must equal the sum total of mass exiting the process, provided the process is in a steady-state condition (all variables remaining constant over time). To give a simple example of this, the mass flow rate of fluid entering a pipe *must* be equal to the mass flow rate of fluid exiting the pipe, provided the pipe is neither accumulating nor releasing mass within its internal volume. "Energy balance" is a parallel concept, stating that the sum total of energy entering a process must equal the sum total of energy exiting a process, provided a steady-state condition (no energy being stored or released from storage within the process).

2.7 Classical mechanics

Classical mechanics (often called *Newtonian* mechanics in honor of Isaac Newton) deal with forces and motions of objects in common circumstances. The vast majority of instrumentation applications deals with this realm of physics. Two other areas of physics, *relativistic* and *quantum*, will not be covered in this chapter because their domains lie outside the typical experience of industrial instrumentation[6].

[6] Relativistic physics deals with phenomena arising as objects travel near the velocity of light. Quantum physics deals with phenomena at the atomic level. Neither is germane to the vast majority of industrial instrument applications.

2.7.1 Newton's Laws of Motion

These laws were formulated by the great mathematician and physicist Isaac Newton (1642-1727). Much of Newton's thought was inspired by the work of an individual who died the same year Newton was born, Galileo Galilei (1564-1642).

1. An object at rest tends to stay at rest; an object in motion tends to stay in motion
2. The acceleration of an object is directly proportional to the net force acting upon it and inversely proportional to the object's mass
3. Forces between objects always exist in equal and opposite pairs

Newton's first law may be thought of as the *law of inertia*, because it describes the property of inertia that all objects having mass exhibit: resistance to change in velocity.

Newton's second law is the verbal equivalent of the force/mass/acceleration formula: $F = ma$

Newton's third law describes how forces always exist in *pairs* between two objects. The rotating blades of a helicopter, for example, exert a downward force on the air (accelerating the air), but the air in turn exerts an upward force on the helicopter (suspending it in flight). These two forces are equal in magnitude but opposite in direction. Such is always the case when forces exist between objects.

2.7.2 Work, energy, and power

Work is the expenditure of energy resulting from exerting a force over a parallel displacement (motion)[7]:

$$W = Fx$$

Where,

W = Work, in joules (metric) or foot-pounds (British)

F = Force doing the work, in newtons (metric) or pounds (British)

x = Displacement over which the work was done, in meters (metric) or feet (British)

Potential energy is energy existing in a stored state, having the potential to do useful work. If we perform work in lifting a mass vertically against the pull of Earth's gravity, we store potential energy which may later be released by allowing the mass to return to its previous altitude. The equation for potential energy in this case is just a special form of the work equation ($W = Fx$), where work is now expressed as potential energy ($W = E_p$), force is now expressed as a weight caused by gravity acting on a mass ($F = mg$), and displacement is now expressed as a height ($x = h$):

$$W = Fx$$

$$E_p = mgh$$

Where,

E_p = Potential energy in joules (metric) or foot-pounds (British)

m = Mass of object in kilograms (metric) or slugs (British)

g = Acceleration of gravity in meters per second squared (metric) or feet per second squared (British)

h = Height of lift in meters (metric) or feet (British)

Kinetic energy is energy in motion. The kinetic energy of a moving mass is equal to:

$$E_k = \frac{1}{2}mv^2$$

Where,

E_k = Potential energy in joules (metric) or foot-pounds (British)

m = Mass of object in kilograms (metric) or slugs (British)

v = Velocity of mass in meters per second (metric) or feet per second (British)

The Law of Energy Conservation is extremely useful in projectile mechanics problems, where we typically assume a projectile loses no energy and gains no energy in its flight. The velocity of

[7] Technically, the best way to express work resulting from force and displacement is in the form of a vector dot-product: $W = \vec{F} \cdot \vec{x}$. The result of a dot product is always a scalar quantity (neither work nor energy possesses a direction, so it cannot be a vector), and the result is the same magnitude as a scalar product only if the two vectors are pointed in the same direction.

a projectile, therefore, depends on its height above the ground, because the sum of potential and kinetic energies must remain constant:

$$E_p + E_k = \text{constant}$$

In free-fall problems, where the only source of energy for a projectile is its initial height, the initial potential energy must be equal to the final kinetic energy:

$$E_p \text{ (initial)} = E_k \text{ (final)}$$

$$mgh_i = \frac{1}{2}mv_f^2$$

We can see from this equation that mass cancels out of both sides, leaving us with this simpler form:

$$gh_i = \frac{1}{2}v_f^2$$

It also leads to the paradoxical conclusion that the mass of a free-falling object is irrelevant to its velocity. That is, both a heavy object and a light object in free fall will hit the ground with the same velocity, and fall for the same amount of time, if released from the same height under the influence of the same gravity[8].

Dimensional analysis confirms the common nature of energy whether in the form of potential, kinetic, or even mass (as described by Einstein's equation). First, we will set these three energy equations next to each other for comparison of their variables:

$E_p = mgh$ Potential energy due to elevation

$E_k = \frac{1}{2}mv^2$ Kinetic energy due to velocity

$E = mc^2$ Mass-to-energy equivalence

[8] In practice, we usually see heavy objects fall faster than light objects due to the resistance of air. Energy losses due to air friction nullify our assumption of constant total energy during free-fall. Energy lost due to air friction never translates to velocity, and so the heavier object ends up hitting the ground faster (and sooner) because it had much more energy than the light object did to start.

Next, we will dimensionally analyze them using standard SI metric units (kilogram, meter, second). Following the SI convention, mass (m) is always expressed in kilograms [kg], distance (h) in meters [m], and time (t) in seconds [s]. This means velocity (v, or c for the velocity of light) in the SI system will be expressed in meters per second [m/s] and acceleration (a, or g for gravitational acceleration) in meters per second squared [m/s^2]:

$$\frac{[\text{kg}][\text{m}^2]}{[\text{s}^2]} = [\text{kg}]\left[\frac{\text{m}}{\text{s}^2}\right][\text{m}] \qquad \text{Potential energy due to elevation}$$

$$\frac{[\text{kg}][\text{m}^2]}{[\text{s}^2]} = [\text{kg}]\left[\frac{\text{m}}{\text{s}}\right]^2 \qquad \text{Kinetic energy due to velocity}$$

$$\frac{[\text{kg}][\text{m}^2]}{[\text{s}^2]} = [\text{kg}]\left[\frac{\text{m}}{\text{s}}\right]^2 \qquad \text{Mass-to-energy equivalence}$$

In all three cases, the unit for energy is the same: kilogram-meter squared per second squared. This is the fundamental definition of a "joule" of energy, and it is the same result given by all three formulae.

Power is defined as the rate at which work is being done, or the rate at which energy is transferred. Mathematically expressed, power is the first time-derivative of work (W):

$$P = \frac{dW}{dt}$$

The metric unit of measurement for power is the *watt*, defined as one joule of work performed per second of time. The British unit of measurement for power is the *horsepower*, defined as 550 foot-pounds of work performed per second of time.

Although the term "power" is often colloquially used as a synonym for force or strength, it is in fact a very different concept. A "powerful" machine is not necessarily a machine capable of doing a great amount of work, but rather (more precisely) a great amount of work *in a short amount of time*. Even a "weak" machine is capable of doing a great amount of work given sufficient time to complete the task. The "power" of any machine is the measure of *how rapidly* it may perform work.

An interesting exercise in dimensional analysis for people familiar with Joule's Law in electric circuits shows just how work and power relate. Power, as you may recall, is defined in electric circuits as the product of voltage and current:

$$P = IV$$

Showing the common units of measurement for each of these variables:

$$[\text{Watts}] = [\text{Amperes}] \times [\text{Volts}] \qquad \text{or} \qquad [\text{W}] = [\text{A}][\text{V}]$$

Now we will substitute more fundamental units of measurement to show how the units comprising "power" really do come from the units comprising "volts" and "amps". We know for example that the unit of the "ampere" is really coulombs of charge flowing per second, and that the unit of the "volt" is really joules of energy (or joules of work) per coulomb of charge. Thus, we may make the unit substitutions and prove to ourselves that the "watt" is really joules of energy (or joules of work) per second of time:

$$\left[\frac{\text{Joules}}{\text{Seconds}}\right] = \left[\frac{\text{Coulombs}}{\text{Seconds}}\right] \times \left[\frac{\text{Joules}}{\text{Coulombs}}\right] \qquad \text{or} \qquad \left[\frac{\text{J}}{\text{s}}\right] = \left[\frac{\text{C}}{\text{s}}\right]\left[\frac{\text{J}}{\text{C}}\right]$$

In summary, voltage is a measure of how much potential energy is infused in every coulomb of charge in an electric circuit, and current is a measure of how quickly those charges flow through the circuit. Multiplying those two quantities tells us the rate at which energy is transferred by those moving charges in a circuit: the rate of charge flow multiplied by the energy value of each charge unit.

2.7.3 Mechanical springs

Many instruments make use of springs to translate force into motion, or visa-versa. The basic "Ohm's Law" equation for a mechanical spring relating applied force to spring motion (displacement) is called *Hooke's Law*[9]:

$$F = -kx$$

Where,

F = Force generated by the spring in newtons (metric) or pounds (British)

k = Constant of elasticity, or "spring constant" in newtons per meter (metric) or pounds per foot (British)

x = Displacement of spring in meters (metric) or feet (British)

Hooke's Law is a linear function, just like Ohm's Law is a linear function: doubling the displacement (either tension or compression) doubles the spring's force. At least this is how springs behave when they are displaced a small percentage of their total length. If you displace a spring more substantially, the spring material will become strained beyond its elastic limit and either yield (permanently deform) or fail (break).

The amount of potential energy stored in a tensed spring may be predicted using calculus. We know that potential energy stored in a spring is the same as the amount of work done on the spring, and work is equal to the product of force and displacement (assuming parallel lines of action for both):

$$E_p = Fx$$

Thus, the amount of work done on a spring is the force applied to the spring ($F = kx$) multiplied by the displacement (x). The problem is, the force applied to a spring varies with displacement and therefore is not constant as we compress or stretch the spring. Thus, in order to calculate the amount of potential energy stored in the spring ($E_p = Fx$), we must calculate the amount of energy stored over infinitesimal amounts of displacement ($F\,dx$, or $kx\,dx$) and then add those bits of energy up ($\int$) to arrive at a total:

$$E_p = \int kx\,dx$$

[9] Hooke's Law may be written as $F = kx$ without the negative sign, in which case the force (F) is the force *applied* on the spring from an external source. Here, the negative sign represents the spring's reaction force to being displaced (the *restoring* force). A spring's reaction force always opposes the direction of displacement: compress a spring, and it pushes back on you; stretch a spring, and it pulls back. A negative sign is the mathematically symbolic way of expressing the opposing direction of a vector.

We may evaluate this integral using the power rule (x is raised to the power of 1 in the integrand):

$$E_p = \frac{1}{2}kx^2 + E_0$$

Where,

E_p = Energy stored in the spring in joules (metric) or foot-pounds (British)

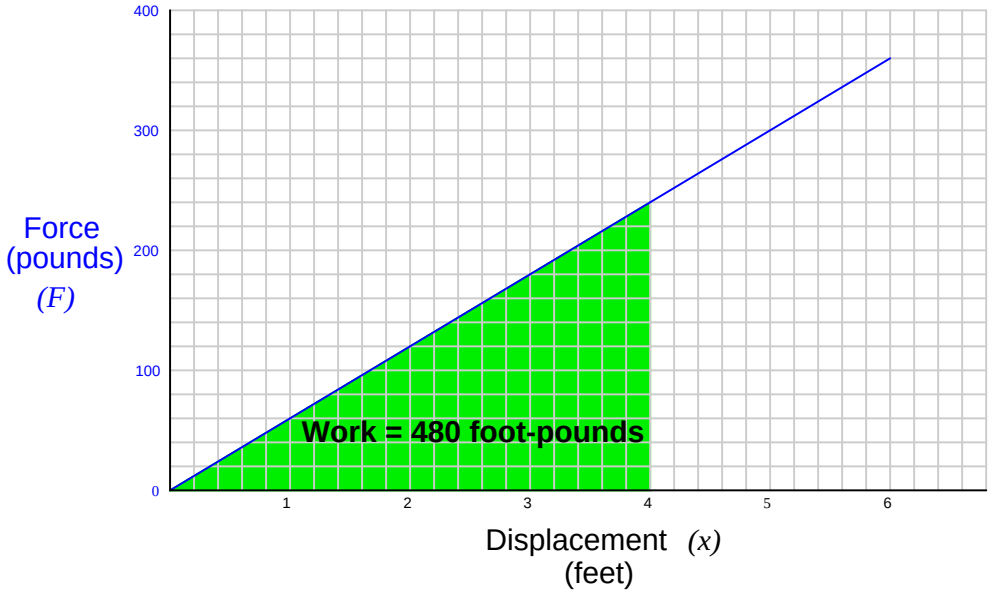
k = Constant of elasticity, or "spring constant" in newtons per meter (metric) or pounds per foot (British)

x = Displacement of spring in meters (metric) or feet (British)

E_0 = The constant of integration, representing the amount of energy initially stored in the spring prior to our displacement of it

For example, if we take a very large spring with a constant k equal to 60 pounds per foot and displace it by 4 feet, we will store 480 foot-pounds of potential energy in that spring (i.e. we will do 480 foot-pounds of work on the spring).

Graphing the force-displacement function on a graph yields a straight line (as we would expect, because Hooke's Law is a linear function). The area accumulated underneath this line from 0 feet to 4 feet represents the integration of that function over the interval of 0 to 4 feet, and thus the amount of potential energy stored in the spring:

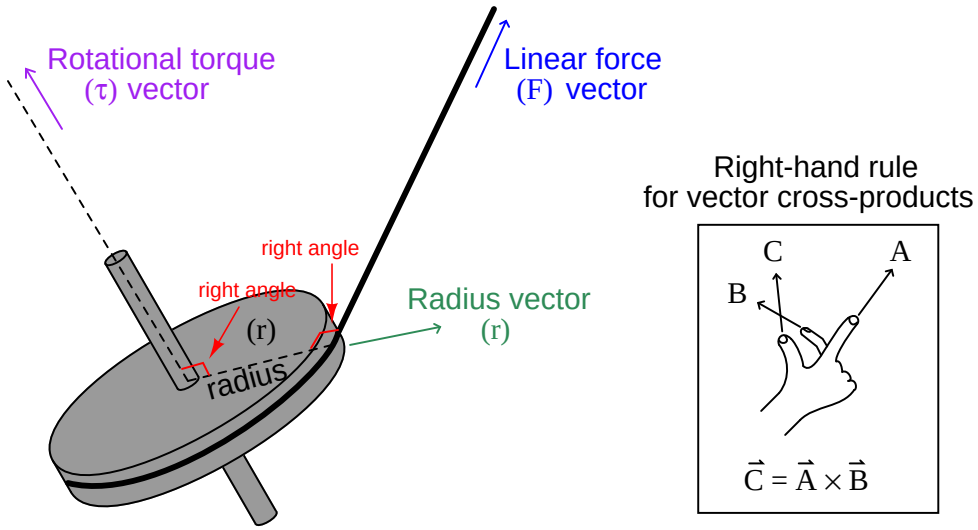


Note how the geometric interpretation of the shaded area on the graph exactly equals the result predicted by the equation $E_p = \frac{1}{2}kx^2$: the area of a triangle is one-half times the base times the height. One-half times 4 feet times 240 pounds is 480 foot-pounds.

2.7.4 Rotational motion

Rotational motion may be quantified in terms directly analogous to linear motion, using different symbols and units.

The rotational equivalent of linear *force* (F) is *torque* (τ). Linear force and rotational torque are both vector quantities, mathematically related to one another by the radial distance separating the force vector from the centerline of rotation. To illustrate with a string pulling on the circumference of a wheel:



This relationship may be expressed mathematically as a *vector cross-product*, where the vector directions are shown by the *right-hand rule* (the first vector $\vec{r}$ is the direction of the index finger, the second vector $\vec{F}$ is the direction of the middle finger, and the product vector $\vec{\tau}$ is the direction of the thumb, with all three vectors perpendicular to each other):

$$\vec{\tau} = \vec{r} \times \vec{F}$$

The proper unit of measurement for torque is the product of the force unit and distance unit. In the metric system, this is customarily the *Newton-meter* (N-m). In the British system, this is customarily the *foot-pound* (ft-lb) or alternatively the *pound-foot* (lb-ft). Note that while these are the exact same *units* as those used to express work, they are not the same types of *quantities*. Torque is a vector cross-product, while work is a *dot-product* ($W = \vec{F} \cdot \vec{x}$). The cross-product of two vectors is always another vector[10], while the dot-product of two vectors is always a scalar (direction-less) quantity. Thus, torque always has a direction, whereas work or energy does not.

[10] Technically, it is a *pseudovector*, because it does not exhibit all the same properties of a true vector, but this is a mathematical abstraction far beyond the scope of this book!

Every quantity of force and motion which may be expressed in linear form has a rotational equivalent. As we have seen, torque (τ) is the rotational equivalent of force (F). The following table contrasts equivalent quantities for linear and rotational motion (all units are metric, shown in *italic* font):

Linear quantity, symbol, and unit	**Rotational quantity, symbol, and unit**
Force (F) *N*	Torque (τ) *N-m*
Linear displacement (x) *m*	Angular displacement (θ) *radian*
Linear velocity (v) *m/s*	Angular velocity (ω) *rad/sec*
Linear acceleration (a) m/s^2	Angular acceleration (α) rad/s^2
Mass (m) *kg*	Moment of Inertia (I) $kg\text{-}m^2$

Familiar equations for linear motion have rotational equivalents as well. For example, Newton's Second Law of motion states, "The acceleration of an object is directly proportional to the net force acting upon it and inversely proportional to the object's mass." We may modify this law for rotational motion by saying, "The angular acceleration of an object is directly proportional to the net torque acting upon it and inversely proportional to the object's moment of inertia." The mathematical expressions of both forms of Newton's Second Law are as follows:

$$F = ma \qquad\qquad \tau = I\alpha$$

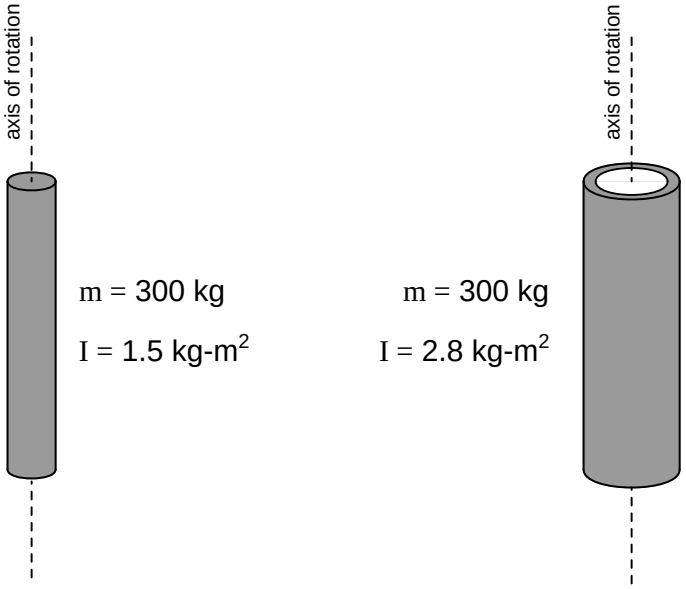
The calculus-based relationships between displacement (x), velocity (v), and acceleration (a) find parallels in the world of angular motion as well. Consider the following formula pairs, linear motion on the left and angular motion on the right:

$$v = \frac{dx}{dt} \quad \text{(Velocity as the time-derivative of displacement)} \quad \omega = \frac{d\theta}{dt}$$

$$a = \frac{dv}{dt} \quad \text{(Acceleration as the time-derivative of velocity)} \quad \alpha = \frac{d\omega}{dt}$$

$$a = \frac{d^2x}{dt^2} \quad \text{(Acceleration as the second time-derivative of displacement)} \quad \alpha = \frac{d^2\theta}{dt^2}$$

An object's "moment of inertia" represents its angular inertia (opposition to changes in rotational velocity), and is proportional to the object's mass and to the square of its radius. Two objects having the same mass will have different moments of inertia if there is a difference in the distribution of their mass relative to radius. Thus, a hollow tube will have a greater moment of inertia than a solid rod of equal mass, assuming an axis of rotation in the center of the tube/rod length:



This is why *flywheels*[11] are designed to be as wide as possible, to maximize their moment of inertia with a minimum of total mass.

The formula describing the amount of work done by a torque acting over an angular displacement is remarkably similar to the formula describing the amount of work done by a force acting over a linear displacement:

$$W = Fx \qquad\qquad W = \tau\theta$$

The formula describing the amount of kinetic energy possessed by a spinning object is also similar to the formula describing the amount of energy possessed by a linearly-traveling object:

$$E_k = \frac{1}{2}mv^2 \qquad\qquad E_k = \frac{1}{2}I\omega^2$$

[11] A "flywheel" is a disk on a shaft, designed to maintain rotary motion in the absence of a motivating torque for the function of machines such as piston engines. The rotational kinetic energy stored by an engine's flywheel is necessary to give the pistons energy to compress the gas prior to the power stroke, during the times the other pistons are not producing power.

2.8 Elementary thermodynamics

Thermodynamics is the study of heat, temperature, and their related effects in physical systems. As a subject, thermodynamics is quite complex and expansive, usually taught as a course in itself at universities. The coverage in this book is limited to some of the more elementary and immediately practical facets of thermodynamics rather than a comprehensive overview.

2.8.1 Heat versus Temperature

Most people use the words *heat* and *temperature* interchangeably, as though they meant the exact same thing. This is unfortunate for every student of thermodynamics, who must first deconstruct this false conception and replace it with one more scientifically accurate before any progress may be made.

When people say something is "hot," what they really mean is that the object has a high temperature. Temperature is a direct function of molecular motion within an object or a fluid sample. This is usually easier to visualize for a gas, where the individual molecules have great freedom of motion.

Heat, by contrast, is an expression of thermal energy transfer. By placing a pot of water over a fire, we are *adding heat* to that pot (transferring thermal energy to the water), the effect of which is to *raise its temperature* (make the water molecules' motions more vigorous). If that same pot is taken away from the fire and allowed to cool, its *loss of heat* (transferring energy out of the water to the surrounding air) will result in its *temperature lowering* (the individual water molecules slow down).

Heat gain or loss often results in temperature change, but not always. In some cases heat may be gained or lost with negligible temperature change – here, the gain or loss of heat manifests as physical changes to the substance other than temperature. One example of this is the boiling of water at constant pressure: no matter how much heat is transferred to the water, its temperature will remain constant at the boiling point (100 degrees Celsius at sea level) until all the water has boiled to vapor. The addition of thermal energy to the boiling water does not raise its temperature, but rather goes into the work of breaking molecules apart from each other so that the liquid turns into vapor.

Heat transfer can *only* happen, though, where there is a difference of temperature between two objects. Thermal energy (heat) flows from the "hotter" (higher-temperature) substance to the "colder" (lower-temperature) substance. To use the boiling water example, the only way to get heat transfer into the water is to subject the water to a hotter substance (e.g., a flame, or a hot electric heating element).

Much more attention will be directed to the concepts of heat and temperature in subsequent subsections.

2.8.2 Temperature

In an ideal, monatomic[12] gas (one atom per molecule), the mathematical relationship between average molecular velocity and temperature is as follows:

$$\frac{1}{2}m\overline{v^2} = \frac{3}{2}kT$$

Where,

m = Mass of each molecule
v = Velocity of molecule in the sample
$\overline{v^2}$ = Mean-squared molecular velocities in the sample
k = Boltzmann's constant (1.38×10^{-23} J / K)
T = Absolute temperature (Kelvin)

Non-ideal gases, liquids, and solids are more complex than this. Not only can the atoms of complex molecules move to and fro, but they may also twist and oscillate with respect to each other. No matter how complex the particular substance may be, however, the basic principle remains unchanged: temperature is an expression of how vigorously molecules are moving within a substance.

There is a temperature at which all molecular motion ceases. At that temperature, the substance cannot possibly become "colder," because there is no more motion to halt. This temperature is called *absolute zero*, equal to -273.15 degrees Celsius, or -459.67 degrees Fahrenheit. Two temperature scales based on this absolute zero point, *Kelvin* and *Rankine*, express temperature relative to absolute zero. A sample of freezing water at sea level, stable in temperature at 0 degrees Celsius (32 degrees Fahrenheit), is also at 273.15 Kelvin[13] or 488.67 degrees Rankine.

A set of common melting and boiling points (at sea-level atmospheric pressure) appears in this table, labeled in these four different units of temperature measurement. Note how degrees Celsius and Kelvin for each point on the table differ by a constant (offset) of 273.15, while each corresponding degree Fahrenheit and degree Rankine value differs by 459.67 (note that many of the figures in this table are slightly rounded, so the offset might not be *exactly* that much). You might think of Kelvin as nothing more than the Celsius scale zero-shifted by 273.15 degrees, and likewise degrees Rankine as nothing more than the Fahrenheit scale zero-shifted by 459.67 degrees:

Melting or boiling substance	^{o}C	^{o}F	K	^{o}R
Melting point of water (H_2O)	0	32	273.15	491.67
Boiling point of water (H_2O)	100	212	373.15	671.67
Melting point of ammonia (NH_3)	-77.7	-107.9	195.5	351.8
Boiling point of ammonia (NH_3)	-33.6	-28.5	239.6	431.2
Melting point of gold (Au)	1063	1945	1336	2405
Melting point of magnesium (Mg)	651	1203.8	924.2	1663.5
Boiling point of acetone (C_3H_6O)	56.5	133.7	329.65	593.37
Boiling point of propane (C_3H_8)	-42.1	-43.8	231.1	415.9
Boiling point of ethanol (C_2H_6O)	78.4	173.1	351.6	632.8

[12] Helium at room temperature is a close approximation of an ideal, monatomic gas, and is often used as an example for illustrating the relationship between temperature and molecular velocity.

[13] Kelvin is typically expressed without the customary "degree" label, unlike the three other temperature units: (degrees) Celsius, (degrees) Fahrenheit, and (degrees) Rankine.

2.8.3 Heat

Heat, being the transfer of energy in thermal (molecular motion) form, may measured in the same units as energy is measured: *joules* (metric) and *foot-pounds* (British). However, alternate units of measurement are often used specifically for heat instead:

- calorie
- kilocalorie (or "dietary calorie")
- British Thermal Unit (BTU)

A *calorie* or heat is defined as the amount of thermal energy transfer required to change the temperature of one gram of water by one degree Celsius (same temperature change as one Kelvin). One calorie is equivalent to 4.186 joules.

A *British Thermal Unit*, or *BTU* is defined as the amount of thermal energy transfer required to change the temperature of one pound of water by one degree Fahrenheit (same temperature change as one degree Rankine). One BTU is equivalent to 778.2 foot-pounds.

The unit of "dietary" calories is used to express the amount of thermal energy available in a sample of food by combustion[14]. Since the official unit of the "calorie" is so small compared to the typical amounts of energy contained in a meal, nutritionists use the unit of the kilocalorie (1000 calories, or 4186 joules) and call it "Calorie" (with a capital letter "C").

Just as "Calories" are used to rate the energy content of food, the heat units of "calories" and "BTU" are very useful in describing the potency of various industrial fuels. The following table shows the *heat of combustion* for a few common fuels, in units of kilocalories per gram and BTU per pound:

Fuel	Combustion heat (kcal/g)	Combustion heat (BTU/lb)
Methane (CH_4)	13.3	23,940
Methanol (CH_4O)	5.43	9,767
Ethanol (C_2H_6O)	7.10	12,783
Propane (C_3H_8)	12.1	21,700
Carbon monoxide (CO)	2.415	4,347

For example, if exactly one gram of methane gas were completely burnt, the resulting heat transfer to water would be sufficient to warm 13.3 kilograms of water from 20 degrees Celsius to 21 degrees Celsius (a temperature rise of one degree Celsius).

If a meal rated at 900 Calories (900 "dietary calories," or 900 kilocalories) of energy is metabolized, the resulting heat would be sufficient to warm a pool of water 900 kilograms in mass (900 liters, or about 237 gallons) by one degree Celsius. This same amount of heat could raise half the amount of water twice the temperature rise: 450 liters of water warmed two degrees Celsius.

[14] Animals process food by performing a very slow version of combustion, whereby the carbon and hydrogen atoms in the food join with oxygen atoms inhaled to produce water and carbon dioxide gas (plus energy). Although it may seem strange to rate the energy content of food by measuring how much heat it gives off when *burnt*, burning is just a faster method of energy extraction than the relatively slow processes of biological metabolism.

2.8.4 Heat transfer

Heat spontaneously flows from higher-temperature substances to lower-temperature substances. This is the phenomenon you experience standing next to a fire on a cold day. Your body is cold (low temperature), but the fire is much hotter (high temperature), and your proximity to the fire aids in heat transfer from the fire to you.

Three principal methods exist for heat to transfer from one substance to another:

- Radiation[15] (by light waves)
- Conduction (by direct contact)
- Convection (by intermediate contact with a moving fluid)

Radiation

If you have ever experienced the immediate sensation of heat from a large fire or explosion some distance away, you know how *radiation* works to transfer thermal energy. Radiation is also the method of heat transfer experienced in the Earth's receiving of heat from the Sun (and also the mechanism of heat loss from Earth to outer space). Radiation is the least efficient of the three heat transfer mechanisms. It may be quantified by the Stefan-Boltzmann Law, which states the rate of heat lost by an object ($\frac{dQ}{dt}$) is proportional to the *fourth power* of its absolute temperature, and directly proportional to its radiating area:

$$\frac{dQ}{dt} = e\sigma AT^4$$

Where,
$\frac{dQ}{dt}$ = Radiant heat loss rate (watts)
e = Emissivity factor (unitless)
σ = Stefan-Boltzmann constant (5.67×10^{-8} W / m^2 $\cdot$ K^4)
A = Surface area (square meters)
T = Absolute temperature (Kelvin)

The emissivity factor varies with surface finish and color, ranging from one (ideal) to zero (no radiation possible). Dark-colored, rough surfaces offer the best emissivity factors, which is why heater elements and radiators are usually painted black.

[15] In this context, we are using the word "radiation" in a very general sense, to mean thermal energy radiated away from the hot source via photons. This is quite different from nuclear radiation, which is what some may assume this term means upon first glance.

Conduction

If you have ever accidently touched a hot iron or stove heating element, you possess a very vivid recollection of heat transfer through *conduction*. In conduction, fast-moving molecules in the hot substance transfer some of their kinetic energy to slower-moving molecules in the cold substance. Since this transfer of energy requires collisions between molecules, it only applies when the hot and cold substances directly contact each other.

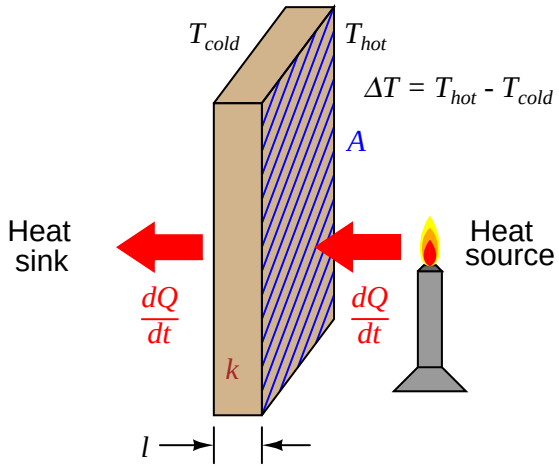
Perhaps the most common application of heat conduction in industrial processes is heat conduction through the walls of a furnace or some other enclosure. In such applications, the desire is usually to *minimize* heat loss through the walls, so those walls will be "insulated" with a substance having poor thermal conductivity.

Conductive heat transfer rate is proportional to the difference in temperature between the hot and cold points, the area of contact, the distance of heat travel from hot to cold, and the thermal conductivity of the substance:

$$\frac{dQ}{dt} = \frac{kA\Delta T}{l}$$

Where,

$\frac{dQ}{dt}$ = Conductive heat transfer rate
k = Thermal conductivity
A = Surface area
ΔT = Difference of temperature between "hot" and "cold" sides
l = Length of heat flow path from "hot" to "cold" side



In the United States, a common measure of insulating ability used for the calculation of conductive heat loss in shelters is the *R-value*. The greater the R-value of a thermally insulating material, the less conductive it is to heat (lower k value). "R-value" mathematically relates to k and l by the following equation:

$$R = \frac{l}{k}$$

Rearranging this equation, we see that $l = kR$, and this allows us to substitute kR for l in the conduction heat equation, then cancel the k terms:

$$\frac{dQ}{dt} = \frac{kA\Delta T}{kR}$$

$$\frac{dQ}{dt} = \frac{A\Delta T}{R}$$

R is always expressed in the compound unit of square feet · hours · degrees Fahrenheit per BTU. This way, with a value for area expressed in square feet and a temperature difference expressed in degrees Fahrenheit, the resulting heat transfer rate ($\frac{dQ}{dt}$) will naturally be in units of BTU per

hour, which is the standard unit in the United States for expressing heat output for combustion-type heaters.

The usefulness of R-value ratings may be shown by a short example. Consider a contractor trailer, raised up off the ground on a mobile platform, with a total skin surface area of 2400 square feet (walls, floor, and roof) and a uniform R-value of 4 for all surfaces. If the trailer's internal temperature must be maintained at 70 degrees Fahrenheit while the outside temperature averages 40 degrees Fahrenheit, the required output of the trailer's heater will be:

$$\frac{dQ}{dt} = \frac{(2400 \text{ ft}^2)(30^o \text{ F})}{4} = 18,000 \text{ BTU per hour}$$

If the trailer's heater is powered by propane and rated at 80% efficiency (requiring 22,500 BTU per hour of fuel heating value to produce 18,000 BTU per hour of heat transfer into the trailer), the propane usage will be just over one pound per hour, since propane fuel has a heating value of 21,700 BTU per pound.

Convection

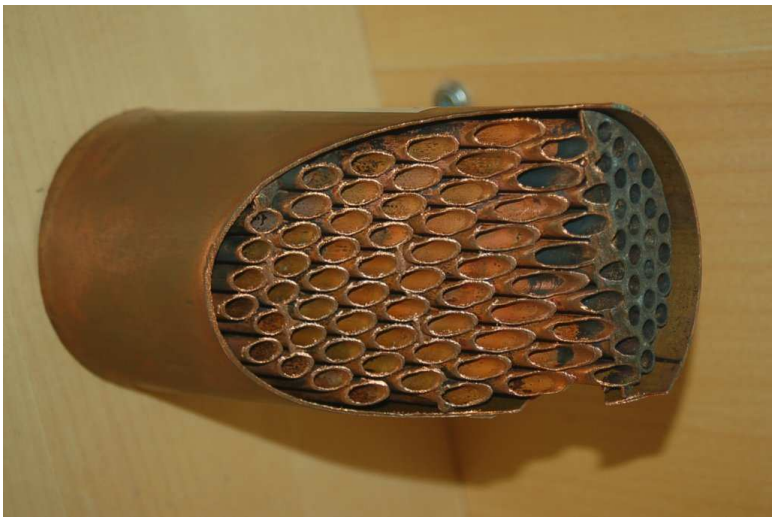
Most industrial heat-transfer processes occur through *convection*, where a moving fluid acts as an intermediary substance to transfer heat from a hot substance (heat *source*) to a cold substance (heat *sink*). Convection may be thought of as two-stage heat conduction on a molecular scale: fluid molecules come into contact with a hot object and pick up heat from that object through conduction, then later come into contact with a cold(er) object and release that heat energy to it through conduction. If the fluid is recycled in a piping loop, the two-stage conduction process repeats indefinitely, individual molecules heating up as they absorb heat from the heat source and then cooling down as they release heat to the heat sink.

Special process devices called *heat exchangers* perform this heat transfer function between two different fluids, the two fluids circulating past each other on different sides of tube walls. A simple example of a heat exchanger is the radiator connected to the engine of an automobile, being a water-to-air exchanger, the engine's hot water transferring heat to cooling air entering the grille of the car as it moves.

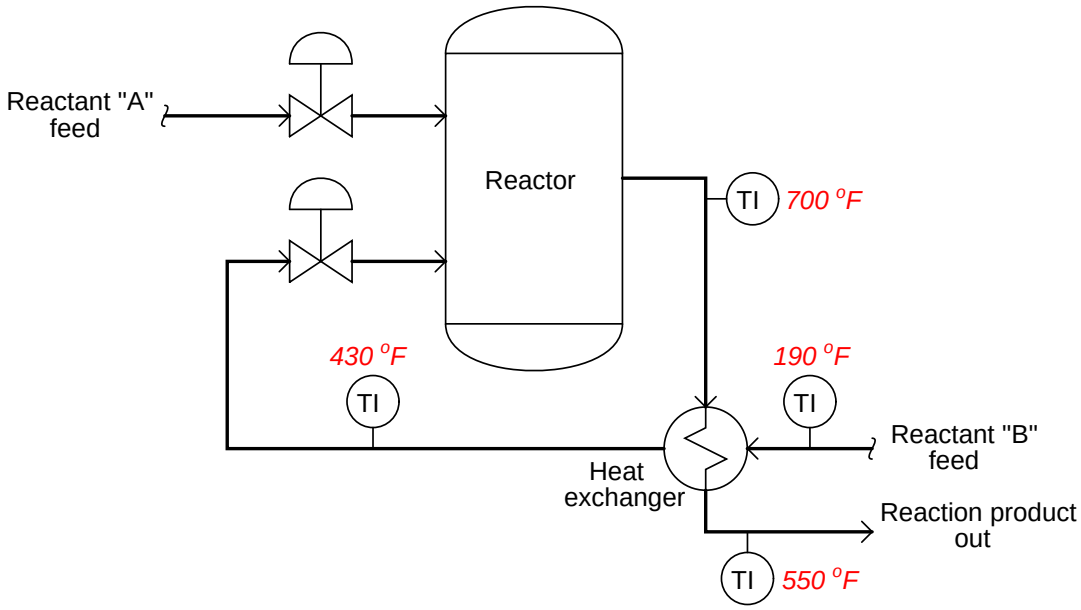
Another example of a liquid-to-air heat exchanger is the *condenser* on a heat pump, refrigerator, or air conditioner, a photograph appearing here:

Another common style of heat exchanger works to transfer heat between two liquids. A small example of this design used to transfer heat from a boat engine is shown here:

This is an example of a *shell-and-tube exchanger*, where one fluid passes inside small tubes and a second fluid passes outside those same tubes, the tube bundle being contained in a shell. The interior of such an exchanger looks like this when cut away:

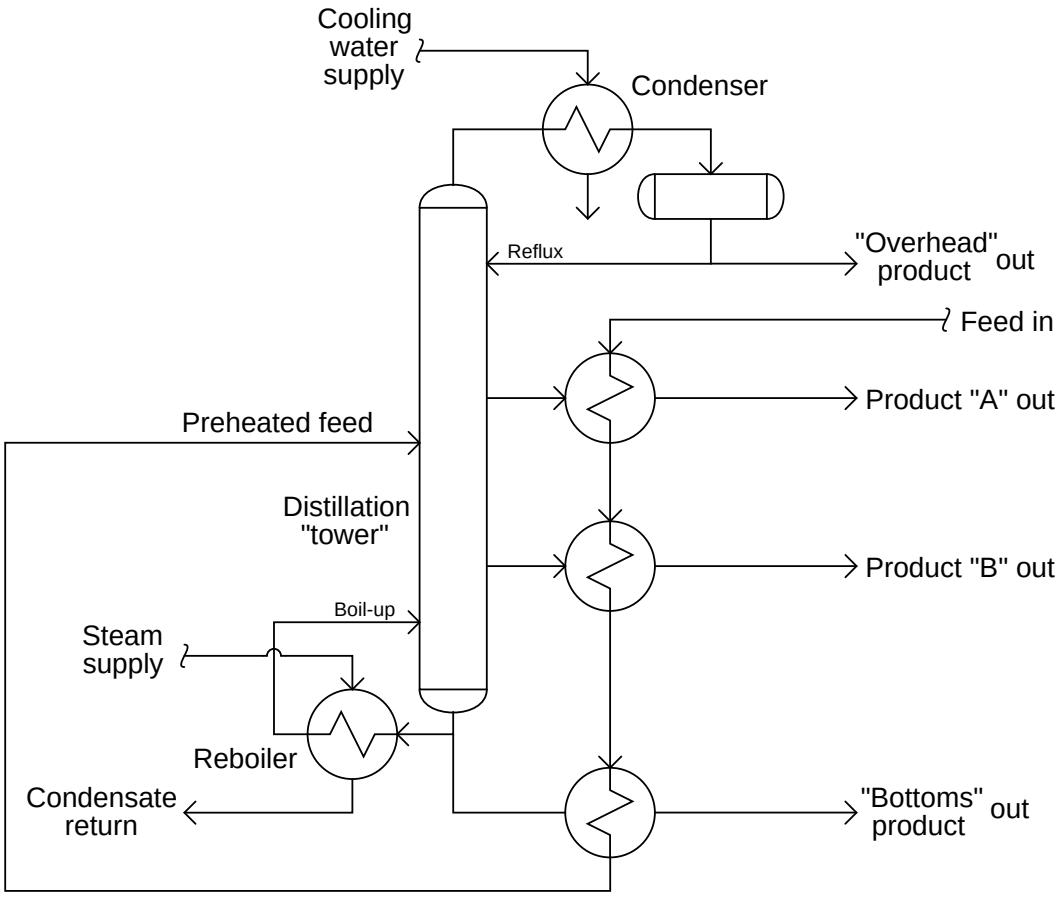


A common application of liquid-to-liquid heat exchangers is in exothermic (heat-releasing) chemical reaction processes where the reactants must be pre-heated before entering a reaction vessel ("reactor"). Since the chemical reaction is exothermic, the reaction itself may be used as the heat source for pre-heating the incoming feed. A simple P&ID shows how a heat exchanger accomplishes this transfer of heat:



Another industrial application of heat exchangers is in *distillation* processes, where mixed components are separated from each other by a continuous process of boiling and condensation. Alcohol purification is one example of distillation, where a mixture of alcohol and water are separated to yield a purer (higher-percentage) concentration of alcohol. Distillation (also called *fractionation*) is a very energy-intensive process, requiring great inputs of heat to perform the task of separation. Any method of energy conservation typically yields significant cost savings in a distillation process, and so we often find heat exchangers used to transfer heat from outgoing (distilled, or fractionated) products to the incoming feed mixture, pre-heating the feed so that less heat need be added to the distillation process from an external source.

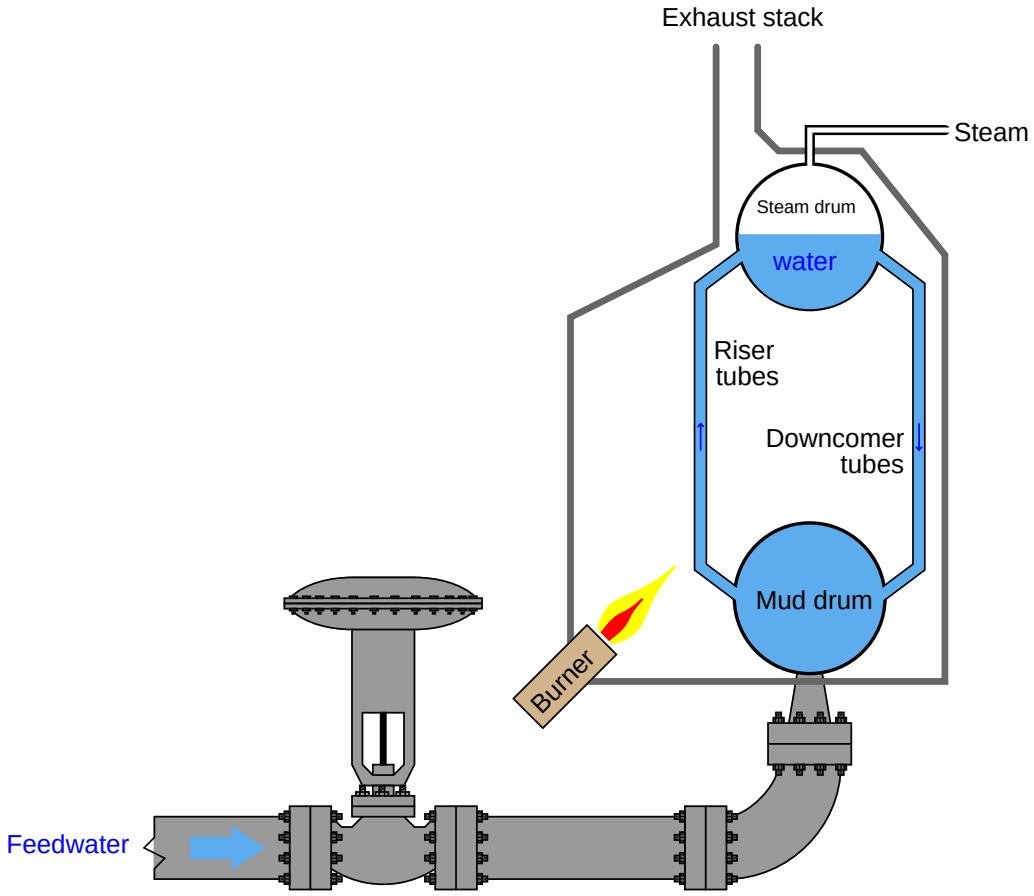
The following P&ID shows a simple distillation process complete with heat exchangers for reboiling (adding heat to the bottom of the distillation column), condensing (extracting heat from the "overhead" product at the top of the column), and energy conservation (transferring heat from the hot products to the incoming feed):



Distillation "columns" (also called *towers* in the industry) are tall vessels containing sets of "trays" where rising vapors from the boiling process contact falling liquid from the condensing process. Temperatures increase toward the bottom of the column, while temperatures decrease toward the top. In this case, steam through a "reboiler" drives the boiling process at the bottom of the column (heat input), and cold water through a "condenser" drives the condensing process at the top of the column (heat extraction). Products coming off the column at intermediate points are hot enough to serve as pre-heating flows for the incoming feed. Note how the "economizing" heat exchangers expose the cold feed flow to the cooler Product A before exposing it to the warmer Product B, and then finally the warmest "Bottoms" product. This sequence of cooler-to-warmer maximizes the efficiency of the heat exchange process, with the incoming feed flowing past products of increasing temperature as it warms up to the necessary temperature for distillation entering the

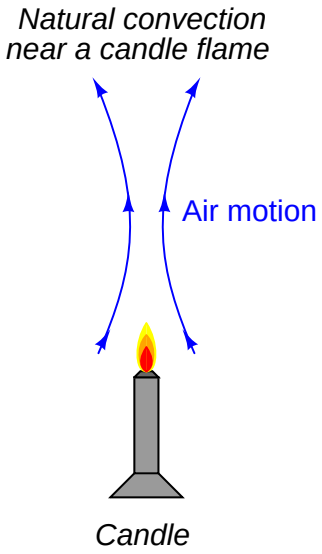
column.

Some heat exchangers transfer heat from hot gases to cool(er) liquids An example of this type of heat exchanger is the construction of a steam boiler, where hot combustion gases transfer heat to water flowing inside metal tubes:



Here, hot gases from the combustion burners travel past the metal "riser" tubes, transferring heat to the water within those tubes. This also serves to illustrate an important convection phenomenon: a *thermal siphon* (often written as *thermosiphon*). As water heats in the "riser" tubes, it becomes less dense, producing less hydrostatic pressure at the bottom of those tubes than the colder water in the "downcomer" tubes. This difference of pressure causes the colder water in the downcomer tubes to flow down to the mud drum, and hot water in the riser tubes to flow up to the steam drum. This natural *convection current* will continue as long as heat is applied to the riser tubes by the burners, and an unobstructed path exists for water to flow in a loop.

Natural convection also occurs in heated air, such as in the vicinity of a lit candle:



This thermally forced circulation of air helps *convect* heat from the candle to all other points within the room it is located, by carrying heated air molecules to colder objects.

Liquid-to-liquid heat exchangers are quite common in industry, where a set of tubes carry one process liquid while a second process liquid circulates on the outside of those same tubes. The metal walls of the tubes act as heat transfer areas for conduction to occur. Metals such as copper with very high k values (very low R values) encourage heat transfer, while long lengths of tube ensure ample surface area for heat exchange.

2.8.5 Specific heat and enthalpy

Earlier, we saw how units of heat measurement were defined in terms of the amount of energy gain or loss required to alter the temperature of a water sample by one degree. In the case of the *calorie*, it was the amount of heat gain/loss required to heat/cool one gram of water one degree Celsius. In the case of the *BTU*, it was the amount of heat gain/loss required to heat/cool one pound of water one degree Fahrenheit.

As one might expect, one heat unit might be similarly defined as the amount of heat gain or loss to alter the temperature one-half of a degree for twice as much water, or two degrees for half as much water. We could express this as a proportionality:

$$Q \propto m\Delta T$$

Where,
Q = Heat gain or loss
m = Mass of sample
ΔT = Temperature change (rise or fall) over time

The next logical question to ask is, "How does the relationship between heat and temperature change work for substances other than water?" Does it take the same amount of heat to change one gram of *iron* by one degree Celsius as it does water? The answer to this question is a resounding *no!* Different substances require vastly different amounts of heat gain/loss to alter their temperature by the same amount, even when the masses of those substances are identical.

We have a term for this ability to absorb or release heat, called *heat capacity* or *specific heat*, symbolized by the variable c. Thus, our heat/mass/temperature change relationship may be described as a true formula instead of a mere proportionality:

$$Q = mc\Delta T$$

Where,
Q = Heat gain or loss (metric calories or British BTU)
m = Mass of sample (metric grams or British pounds)
c = Specific heat of substance
ΔT = Temperature change (metric degrees Celsius or British degrees Fahrenheit)

Pure water, being the standard by which all other substances are measured, has a specific heat value of 1. The smaller the value for c, the less heat gain or loss is required to alter the substance's temperature by a set amount. That substance (with a low value of c) has a low "heat capacity" because each degree of temperature rise or fall represents a relatively small amount of energy gained or lost. Substances with low c values are easy to heat and cool, while substances having high c values require much heat in order to alter their temperatures, assuming equal masses.

A table of specific heat values (at room temperature, 25 degrees Celsius[16]) for common substances appears here:

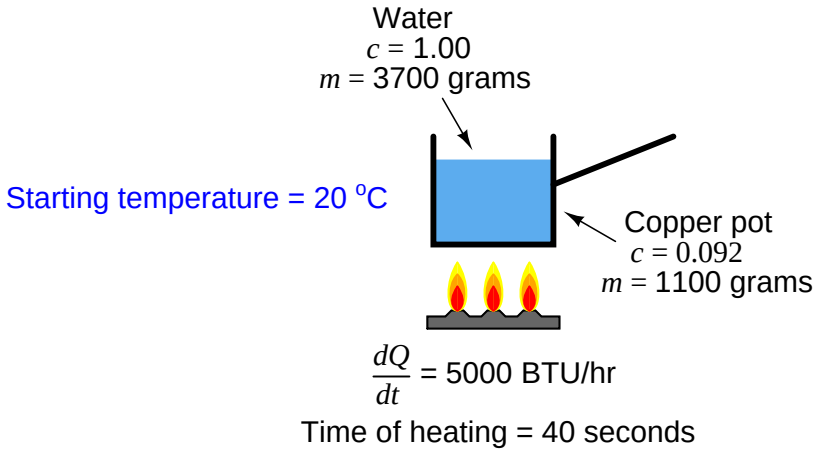
Substance	Specific heat value (c) cal/g·^{o}C or BTU/lb·^{o}F
Aluminum (solid)	0.215
Iron (solid)	0.108
Copper (solid)	0.092
Lead (solid)	0.031
Water (liquid)	1.0
Methanol (liquid)	0.609
Ethanol (liquid)	0.587
Acetone (liquid)	0.521
Hydrogen (gas)	3.41
Helium (gas)	1.24
Nitrogen (gas)	0.249
Oxygen (gas)	0.219

If a liquid or a gas is chosen for use as a coolant (a substance to efficiently convect heat away from an object), greater values of c are better. Water is one of the best liquid coolants with its relatively high c value of one: it has more capacity to absorb heat than other liquids, for the same rise in temperature. The ideal coolant would have an infinite c value, being able to absorb an infinite amount of heat without itself rising in temperature at all.

As you can see from the table, the light gases (hydrogen and helium) have extraordinarily high c values. Consequently, they function as excellent gas coolants. This is why large electric power generators often use hydrogen gas as a coolant: that gas has an amazing ability to absorb heat from the wire windings of a generator without suffering a large rise in temperature. Helium, although not as good a coolant as hydrogen, has the distinct advantage of being chemically inert (non-reactive), in stark contrast to hydrogen's extreme flammability. Some nuclear reactors use helium as a coolant rather than a liquid such as water or liquefied sodium metal.

[16] An important detail to note is that specific heat does *not* remain constant over wide temperature changes. This complicates calculations of heat required to change the temperature of a sample: instead of simply multiplying the temperature change by mass and specific heat ($Q = mc\Delta T$ or $Q = mc[T_2 - T_1]$), we must *integrate* specific heat over the range of temperature ($Q = m \int_{T_1}^{T_2} c\, dT$), summing up infinitesimal products of specific heat and temperature change ($c\, dT$) over the range starting from temperature T_1 through temperature T_2 then multiplying by the mass to calculate total heat required.

Numerical examples are helpful in better understanding specific heat. Consider a case where a copper pot filled with water receives heat from a small gas burner operating at an output of 5,000 BTU per hour (350 calories per second):



A reasonable question to ask would be, "How much will the temperature of this water-filled pot rise after 40 seconds of heating?" With the burner's heat output of 350 calories per second and a heating time of 40 seconds, we may assume[17] the amount of heat absorbed by the water-filled pot will be the simple product of heat rate times time:

$$Q = \left(\frac{dQ}{dt}\right) t = \left(\frac{350 \text{ cal}}{\text{sec}}\right) 40 \text{ sec} = 14000 \text{ calories}$$

This amount of heat not only goes into raising the temperature of the water, but it also raises the temperature of the copper pot. Each substance (water, copper) has its own specific heat and mass values (c and m), but they will share the same temperature rise (ΔT), so we must sum their heats as follows:

$$Q_{total} = Q_{pot} + Q_{water}$$

$$Q_{total} = m_{pot} c_{pot} \Delta T + m_{water} c_{water} \Delta T$$

$$Q_{total} = (m_{pot} c_{pot} + m_{water} c_{water}) \Delta T$$

[17] In reality, the amount of heat actually absorbed by the pot will be less than this, because there will be heat losses from the warm pot to the surrounding (cooler) air. However, for the sake of simplicity, we will assume *all* the burner's heat output goes into the pot and the water it holds.

Solving this equation for temperature rise, we get:

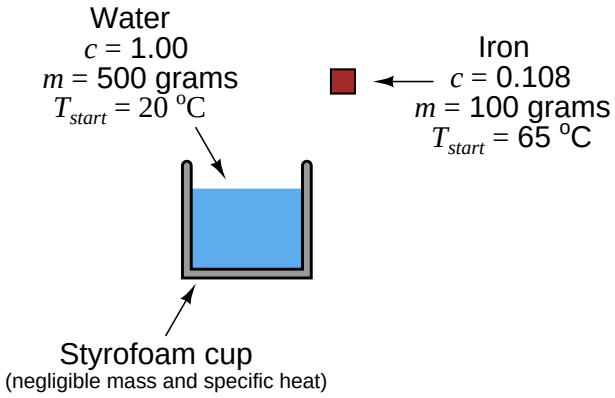
$$\Delta T = \frac{Q_{total}}{m_{pot}c_{pot} + m_{water}c_{water}}$$

$$\Delta T = \frac{14000 \text{ cal}}{(1100 \text{ g})(0.092\frac{\text{cal}}{\text{g}^o\text{C}}) + (3700 \text{ g})(1\frac{\text{cal}}{\text{g}^o\text{C}})}$$

$$\Delta T = 3.68\,^o\text{C}$$

So, if the water and pot began at a temperature of 20 degrees Celsius, they will be at a temperature of 23.68 degrees Celsius after 40 seconds of heating over this small burner.

Another example involves the mixing of two substances at different temperatures. Suppose a heated mass of iron drops into a cool container[18] of water. Obviously, the iron will lose heat energy to the water, causing the iron to decrease in temperature while the water rises in temperature. Suppose the iron's mass is 100 grams, and its original temperature is 65 degrees Celsius. Suppose the water's mass is 500 grams, and its original temperature is 20 degrees Celsius:



[18] We will assume for the sake of this example that the container holding the water is of negligible mass, such as a Styrofoam cup. This way, we do not have to include the container's mass or its specific heat into the calculation.

What will the equilibrium temperature be after the iron falls into the water and both their temperatures equalize? We may solve this by setting two heat equations equal to each other[19]: the heat lost by the iron and the heat gained by the water, with the final equilibrium temperature being T:

$$Q_{iron-loss} = Q_{water-gain}$$

$$m_{iron}c_{iron}(65\,^{o}\text{C} - T) = m_{water}c_{water}(T - 20\,^{o}\text{C})$$

Note how the ΔT term is carefully set up for each side of the equation. In order to make the iron's heat loss a positive value and the water's heat gain a positive value, we must ensure the quantity within each set of parentheses is positive. For the iron, this means ΔT will be 65 degrees minus the final temperature. For the water, this means ΔT will be the final temperature minus its starting temperature of 20 degrees.

$$m_{iron}c_{iron}(65) - m_{iron}c_{iron}T = m_{water}c_{water}T - m_{water}c_{water}(20)$$

$$m_{iron}c_{iron}(65) + m_{water}c_{water}(20) = m_{iron}c_{iron}T + m_{water}c_{water}T$$

$$m_{iron}c_{iron}(65) + m_{water}c_{water}(20) = T(m_{iron}c_{iron} + m_{water}c_{water})$$

$$T = \frac{m_{iron}c_{iron}(65) + m_{water}c_{water}(20)}{m_{iron}c_{iron} + m_{water}c_{water}}$$

$$T = \frac{(100)(0.108)(65) + (500)(1)(20)}{(100)(0.108) + (500)(1)}$$

$$T = 20.95\,^{o}\text{C}$$

Thus, the iron's temperature falls from 65 degrees Celsius to 20.95 degrees Celsius, while the water's temperature rises from 20 degrees Celsius to 20.95 degrees Celsius. The water's tremendous specific heat value compared to the iron (nearly 10 times as much!), as well as its superior mass (5 times as much) results in a much larger temperature change for the iron than for the water.

[19] An alternative way to set up the problem would be to calculate ΔT for each term as $T_{final} - T_{start}$, making the iron's heat loss a negative quantity and the water's heat gain a positive quantity, in which case we would have to set up the equation as a zero-sum balance, with zero equal to $Q_{iron} + Q_{water}$. I find this approach less intuitive than simply saying the iron's heat loss will be equal to the water's heat gain, and setting up the equation as two positive values equal to each other.

An analogy to help grasp the concept of specific heat is to imagine heat as a fluid[20] that may be "poured" into vessels of different size, those vessels being objects or substances to be heated. The amount of liquid held by any vessel represents the total amount of thermal energy, while the *height* of the liquid inside any vessel represents its temperature:

Fluid analogy for heat and temperature

The factor determining the relationship between liquid volume (heat) and liquid height (temperature) is of course the cross-sectional area of the vessel. The wider the vessel, the more heat will be required to "fill" it up to any given temperature. In this analogy, the area of the vessel is analogous to the term mc: the product of mass and specific heat. Objects with larger mass require more heat to raise their temperature to any specific point, specific heats being equal. Likewise, objects with large specific heat values require more heat to raise their temperature to any specific point, masses being equal.

In the first numerical calculation example where we determined the temperature of a pot of water after 40 seconds of heating, the analogous model would be to determine the height of liquid in a vessel after pouring liquid into it for 40 seconds at a fixed rate. A model for the second numerical example would be to calculate the equilibrium height (of liquid) after connecting two vessels together at their bottoms with a tube. Although the liquid heights of those vessels may be different at first, the levels will equalize after time by way of liquid passing through the tube from the higher-level vessel to the lower-level vessel.

[20] This is not far from the reality of eighteenth-century science, where heat was thought to be an invisible fluid called *caloric*.

Many industrial processes use fluids to convectively transfer heat from one object (or fluid) to another. In such applications, it is important to know how much heat will be carried by a specific quantity of that fluid over a specified temperature drop. One common way to express this heat quantity is called *enthalpy*. Enthalpy is the amount of heat lost by a unit mass (one gram metric, or one pound British) of the fluid as it cools from a given temperature all the way down to the freezing point of water (0 degrees Celsius, or 32 degrees Fahrenheit). A sample of water at a temperature of 125 degrees Fahrenheit, for example, has an enthalpy of 93 BTU per pound (or 93 calories per gram):

$$Q = mc\Delta T$$

$$Q = (1 \text{ lb}) \left(1 \frac{\text{BTU}}{\text{lb}^o\text{F}} \right) (125\ ^o\text{F} - 32\ ^o\text{F})$$

$$Q = 93 \text{ BTU}$$

Even if the process in question does not cool the heat transfer fluid down to water's freezing point, enthalpy is a useful figure for estimating the thermal energy "content" of hot fluids (per unit mass). Enthalpy is especially useful when dealing with heat transfer fluids as they change phase from vapor to liquid, as the following subsection will discuss.

2.8.6 Phase changes

Scientists often speak of four *phases* of matter: *solid*, *liquid*, *gas* (or *vapor*), and *plasma*. Of these four, the first three are common to everyday life. Plasma is a phase of matter where the atoms of a gas a superheated to the point where they become electrically ionized, such as neon gas in an electric tube light, or the gas comprising stars in space.

Phase changes are very important in thermodynamics, principally because energy transfer (heat loss or heat gain) must occur for a substance to change states, often with negligible change in temperature. To cite an example, consider the case of water (a liquid) turning into steam (a vapor) at atmospheric pressure. At sea level, this phase change will occur at a temperature of 100 degrees Celsius, or 212 degrees Fahrenheit. The amount of energy required to increase the temperature of water up to its boiling point is a simple function of the sample's mass and its original temperature. For instance, a sample of water 70 grams in mass starting at 24 degrees Celsius will require 5320 calories of heat to reach the boiling point:

$$Q = mc\Delta T$$

$$Q = (70\text{ g})\left(1\frac{\text{cal}}{\text{g}^o\text{C}}\right)(100\,^o\text{F} - 24\,^o\text{F})$$

$$Q = 5320\text{ cal}$$

However, actually boiling the 70 grams of water into 70 grams of steam (both at 100 degrees Celsius) requires a comparatively enormous input of heat: *37,734 calories* – over seven times as much heat to turn the water to steam than required to warm the water to its boiling point. Furthermore, this additional input of 37,734 calories does not increase the temperature of the water one bit: the resulting steam is still at (only) 100 degrees Celsius. If further heat is added to the 70 gram steam sample, its temperature will rise, albeit at a rate proportional to the value of steam's specific heat (0.48 calories per gram degree Celsius, or BTU per pound degree Fahrenheit).

What we see here is a fundamentally different phenomenon than we did with specific heat. Here, we are looking at the thermal energy required to transition a substance from one phase to another, not to change its temperature. We call this quantity *latent heat* rather than *specific heat*, because no temperature change is manifest[21]. Conversely, if we allow the steam to condense back into liquid water, it must release the same 37,734 calories of heat energy we invested in it turning the water into steam before it may cool at all below the boiling point (100 degrees Celsius).

[21] The word "latent" refers to something with potential that is not yet realized. Here, heat exchange takes place without there being any realized change in temperature.

As with specific heat, there is a formula relating mass, latent heat, and heat exchange:

$$Q = mL$$

Where,

Q = Heat of transition required to completely change the phase of a sample (metric calories or British BTU)

m = Mass of sample (metric grams or British pounds)

L = Latent heat of substance

Each substance has its own set of latent heat values, one[22] for each phase-to-phase transition. Water, for example, exhibits a latent heat of vaporization (boiling/condensing) of 539.1 calories per gram, or 970.3 BTU per pound. Water also exhibits a latent heat of fusion (melting/freezing) of 79.7 calories per gram, or 143.5 BTU per pound. Both figures are enormous compared to water's specific heat value of 1 calorie per gram-degree Celsius (or 1 BTU per pound-degree Fahrenheit[23]): it takes only one calorie of heat to warm one gram of water one degree Celsius, but it takes *539.1 calories* of heat to boil that same gram of water into one gram of steam, and *79.7 calories* of heat to melt one gram of ice into one gram of water. The lesson here is simple: phase changes involve huge amounts of heat.

A table showing various latent heats of vaporization (all at room temperature, 70 degrees Fahrenheit) for common industrial fluids appears here, contrasted against their specific heat values (as liquids). In each case you will note how much larger L is than c:

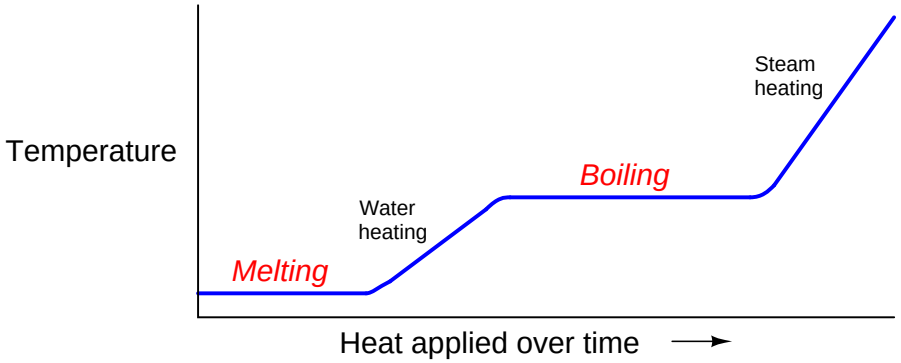
Fluid @ 70 ^{o}F	$L_{vaporization}$, BTU/lb	$L_{vaporization}$, cal/g	c_{liquid}
Water	970.3	539.1	1
Ammonia	508.6	282.6	1.1
Carbon dioxide	63.7	35.4	0.66
Butane	157.5	87.5	0.56
Propane	149.5	83.06	0.6

One of the most important, and also non-intuitive, consequences of latent heat is the relative stability of temperature during the phase-change process. Referencing the table of latent heats of vaporization, we see how much more heat is needed to boil a liquid into a vapor than is needed to warm that same liquid one degree of temperature. During the process of boiling, all heat input to the liquid goes into the task of phase change (latent heat) and none of it goes into increased temperature. In fact, until all the liquid has been vaporized, the liquid's temperature *cannot* rise above its boiling point! The requirement of heat input to vaporize a liquid forces temperature to stabilize (not rise further) until *all* the liquid has evaporated from the sample.

[22] Latent heat of vaporization also varies with pressure, as different amounts of heat are required to vaporize a liquid depending on the pressure that liquid is subject to. Generally, increased pressure (increased boiling temperature) results in less latent heat of vaporization.

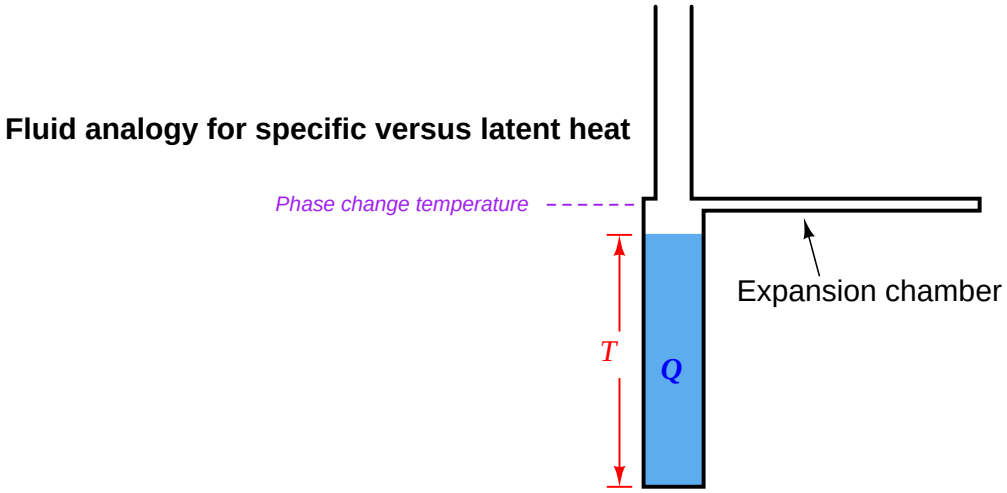
[23] The reason specific heat values are identical between metric and British units, while latent heat values are not, is because latent heat does not involve temperature change, and therefore there is one less unit conversion taking place between metric and British when translating latent heats. Specific heat in both metric and British units is *defined* in such a way that the three different units for heat, mass, and temperature all cancel each other out. With latent heat, we are only dealing with mass and heat, and so we have a proportional conversion of $\frac{5}{9}$ or $\frac{9}{5}$ left over, just the same as if we were converting between degrees Celsius and Fahrenheit alone.

If we take a sample of ice and add heat to it over time until it melts, warms, boils, and then becomes steam, we will notice a temperature profile that looks something like this:



The flat areas of the graph during the melting and boiling periods represents times where the sample's temperature does not change at all, but where all heat input goes into the work of changing the sample's phase. Only where we see the curve rising does the temperature change.

To use our liquid-filled vessel analogy again, it is as if at some point along the vessel's height there is a pipe connection leading to a large, relatively flat expansion chamber, so that the vessel "acts" as if its area were much larger at one point, requiring *much* more fluid volume (heat) to change height (temperature):



Liquid poured into this vessel will fill it at a rate proportional to the volume added and inversely proportional to the vessel's cross-sectional area at the current liquid height. As soon as the liquid level reaches the expansion chamber, a great deal more liquid must be added to cause level to increase, since this chamber must fill before the liquid level may rise above it. Once that happens, the liquid level rises at a different rate with addition introduced volume, because now the phase is different (with a different specific heat value).

Remember that the filling of a vessel with liquid is merely an analogy for heat and temperature, intended to provide an easily visualized process mimicking another process not so easily visualized. The important concept to realize with latent heat and phase change is that it constitutes a discontinuity in the temperature/heat function of any given substance.

A vivid demonstration of this phenomenon is to take a paper[24] cup filled with water and place it in the middle of a roaring fire[25]. "Common sense" might tell you the paper will burn through with the fire's heat, so that the water runs out of the cup through the burn-hole. This does not happen, however. Instead, the water in the cup will rise in temperature until it boils, and there it will maintain that temperature no matter how hot the fire burns. The boiling point of water happens to be substantially below the burning point of paper, and so the boiling water keeps the paper cup too cool to burn. As a result, the paper cup remains intact so long as water remains in the cup. The *rim* of the cup above the water line will burn up because the steam does not have the same temperature-stabilizing effect as the water, leaving a rimless cup that grows shorter as the water boils away.

[24] Styrofoam and plastic cups work as well, but paper exhibits the furthest separation between the boiling point of water and the burning point of the cup material, and it is usually thin enough to ensure good heat transfer from the outside (impinging flame) to the inside (water).

[25] This is a lot of fun to do while camping!

The point at which a pure substances changes phase not only relates to temperature, but to pressure as well. We may speak casually about the boiling point of water being 100 degrees Celsius (212 degrees Fahrenheit), but that is only if we assume the water and steam are at atmospheric pressure (at sea level). If we reduce the ambient air pressure[26], water will boil at a lesser temperature. Anyone familiar with cooking at high altitudes knows you must generally cook for longer periods of time at altitude, because the decreased boiling temperature of water is not as effective for cooking. Conversely, anyone familiar with *pressure cooking* (where the cooking takes place inside a vessel pressurized by steam) knows how comparatively little cooking time is required because the pressure raises water's boiling temperature. In either of these scenarios, where pressure influences boiling temperature, the latent heat of water acts to hold the boiling water's temperature stable until all the water has boiled away. The only difference is the temperature at which the water begins to boil (or when the steam begins to condense).

Many industrial processes use boiling liquids to convectively transfer heat from one object (or fluid) to another. In such applications, it is important to know how much heat will be carried by a specific quantity of the vapor as it condenses into liquid over a specified temperature drop. The quantity of *enthalpy* (heat content) used for rating the heat-carrying capacity of liquids applies to condensing vapors as well. Enthalpy is the amount of heat lost by a unit mass (one gram metric, or one pound British) of the fluid as it cools from a given temperature all the way down to the freezing point of water (0 degrees Celsius, or 32 degrees Fahrenheit). When the fluid's initial state is vapor, and it condenses into liquid as it cools down to the reference temperature (32 ^{o}F), the heat content (enthalpy) is not just a function of specific heat, but also of latent heat. Steam at atmospheric pressure and 212 degrees Fahrenheit (the boiling point of water) has an enthalpy of about 1150 BTU per pound. 970 BTU is released due to the phase change from vapor to liquid, while the rest is due to the water's specific heat of (approximately) one as it cools from 212 degrees Fahrenheit to 32 degrees Fahrenheit (approximately 180 BTU released).

If the vapor in question is at a temperature greater than its boiling point at that pressure, the vapor is said to be *superheated*. The enthalpy of superheated vapor comes from three different heat-loss mechanisms:

- Cooling the vapor down to its condensing temperature (specific heat of vapor)
- Phase-changing from vapor to liquid (latent heat of phase change)
- Cooling the liquid down to the reference temperature (specific heat of liquid)

[26] This may be done in a vacuum jar, or by traveling to a region of high altitude where the ambient air pressure is less than at sea level.

Using steam as the example once more, a sample of superheated steam at 500 degrees Fahrenheit and atmospheric pressure (boiling point = 212 degrees Fahrenheit) has an enthalpy of approximately 1287 BTU per pound. We may calculate the heat lost by one pound of this superheated steam as it cools from 500 ^{o}F to 32 ^{o}F in each of the three steps previously described. Here, we will assume a specific heat for steam of 0.476, a specific heat for water of 1, and a latent heat of vaporization for water of 970:

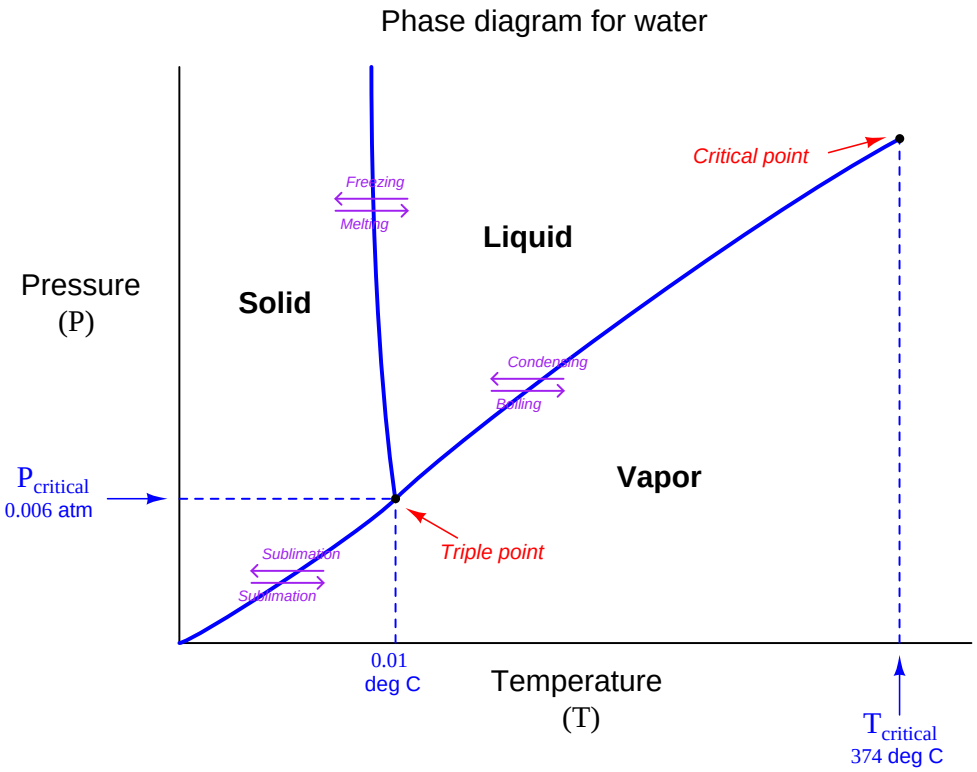
Heat loss mechanism	Formula	Quantity
Cooling vapor	$Q = mc\Delta T$	(1)(0.476)(500-212) = 137 BTU
Phase change	$Q = mL$	(1)(970) = 970 BTU
Cooling liquid	$Q = mc\Delta T$	(1)(1)(212-32) = 180 BTU
TOTAL		1287 BTU

Enthalpy values are very useful in steam engineering to predict the amount of thermal energy delivered to a load given the steam's initial temperature, its final (cooled) temperature, and the mass flow rate. Although the definition of enthalpy – where we calculate heat value by supposing the vapor cools all the way down to the *freezing point* of water – might seem a bit strange and impractical (how often does steam lose so much heat to a process that it reaches freezing temperature?), it is not difficult to shift the enthalpy value to reflect a more practical final temperature. Since we know the specific heat of liquid water is very nearly one, all we have to do is offset the enthalpy value by the amount that the final temperature differs from freezing in order to calculate how much heat the steam will lose (per pound) to its load.

For example, suppose we were to use the same 500 ^{o}F superheated steam used in the previous example to heat a flow of oil through a heat exchanger, with the steam condensing to water and then cooling down to 170 degrees Fahrenheit as it delivers heat to the flowing oil. Here, the steam's enthalpy value of 1287 BTU per pound may simply be offset by 138 (170 degrees minus 32 degrees) to calculate how much heat (per pound) this steam will deliver to the oil: 1287 - 138 = 1149 BTU per pound. If we require a heat transfer rate of 45,000 BTU per hour to the flowing oil, the steam flow rate will have to be just over 39 pounds per hour.

2.8.7 Phase diagrams and critical points

A comprehensive way of describing the relationship between pressure, temperature, and substance phase is with something called a *phase diagram*. With pressure shown on one axis, and temperature on the other, a phase diagram describes the various phases of a substance in possible equilibrium at certain pressure/temperature combinations.



This phase diagram (for water) illustrates some of the features common to all phase diagrams: curved lines define the boundaries between solid, liquid, and vapor phases; the point of intersection of these three curves is where the substance may exist in all three phases simultaneously (called the *triple point* of water); and points where a curve simply ends within the span of the graph indicate critical points, where the certain phases cease to exist.

The curved line from the triple point up and to the right defines the boundary between liquid water and water vapor. Each point on that line represents a set of unique pressure and temperature conditions for boiling (changing phase from liquid to vapor) or for condensation (changing phase from vapor to liquid). As you can see, increased pressure results in an increased boiling point (i.e. at higher pressures, water must be heated to greater temperatures before boiling may take place). In fact, the whole concept of a singular boiling *point* for water becomes quaint in the light of a phase diagram: boiling is seen to occur over a wide range of temperatures[27], the exact temperature

[27] Anywhere between the triple-point temperature and the critical temperature, to be exact.

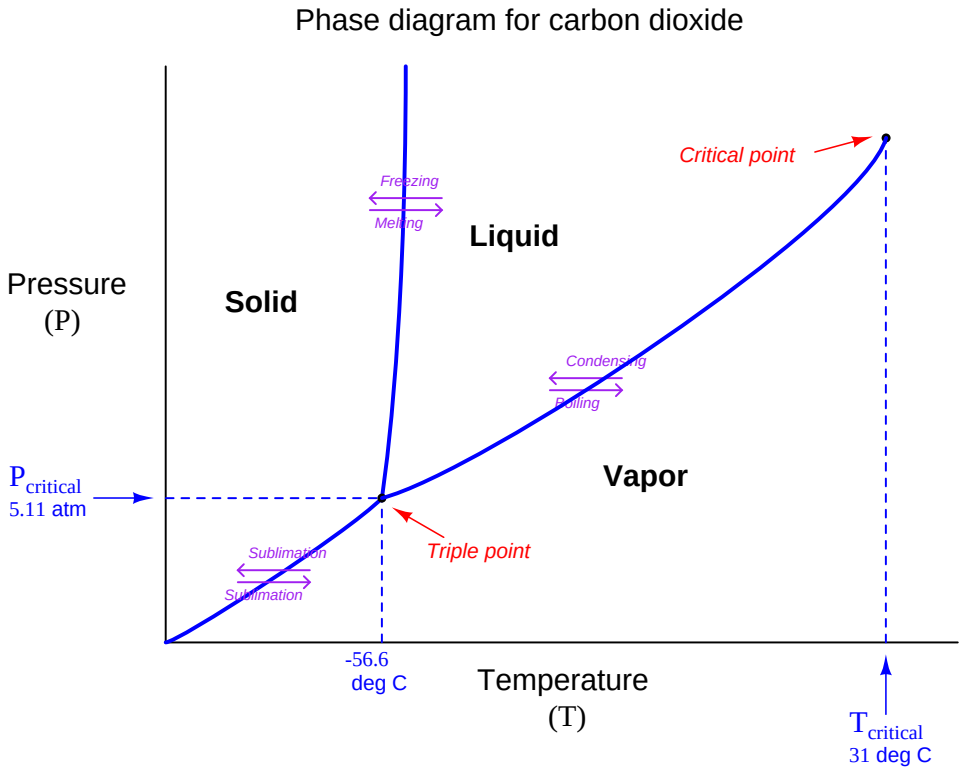
varying with pressure.

Something interesting happens when the temperature is raised above a certain value called the *critical temperature*. At this value (approximately 374 degrees Celsius for water), no amount of pressure will maintain it in a liquid state. Water heated to 374 degrees Celsius or above can only exist in a stable condition as a vapor.

The slightly curved line from the triple point up and to the left defines the boundary between solid ice and liquid water. As you can see, the near-vertical pitch of this curve suggests the freezing temperature of water is quite stable over a wide pressure range.

Below a certain pressure, called the *critical pressure*, the possibility of a stable liquid phase disappears. The substance may exist in solid or gaseous forms, but not liquid, if the pressure is below the critical pressure value.

Carbon dioxide exhibits a different set of curves than water on its phase diagram, with its own unique critical temperature and pressure values:



Note how the critical pressure of carbon dioxide is well above ambient conditions on Earth. This means carbon dioxide is not stable in its liquid state unless put under substantial pressure (at least 60.4 PSIG). This is why solid carbon dioxide is referred to as *dry ice*: it does not liquefy with the application of heat, rather it *sublimates* directly into its vapor phase.

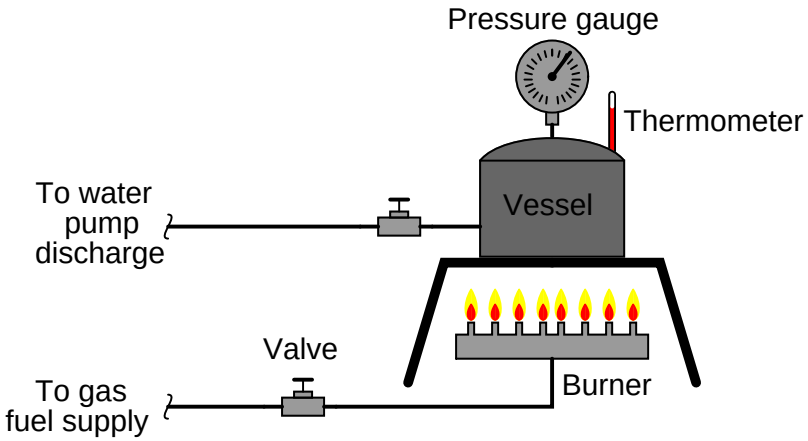
Another interesting difference between the carbon dioxide and water phase diagrams is the slope of the solid/liquid boundary line. With water, this boundary drifts to the left (lower temperature) as pressure increases. With carbon dioxide, this boundary drifts to the right (higher temperature) as pressure increases. Whether the fusion temperature increases or decreases with increasing pressure marks whether that substance contracts or expands as it transitions from liquid to solid. Carbon dioxide, like most pure substances, contracts to a smaller volume when it goes from liquid to solid, and its fusion curve drifts to the right as pressure increases. Water is unusual in this regard, expanding to a larger volume when freezing, and its fusion curve drifts to the left as pressure increases.

2.8.8 Thermodynamic degrees of freedom

If we look at the areas bounded by phase transition curves in a phase diagram (solid area, liquid area, and vapor area), we see that both pressure and temperature may change independent of one another. A vessel filled with liquid water, for instance, may be at 30 degrees Celsius and 2 atmospheres, or at 50 degrees Celsius and 2 atmospheres, or at 50 degrees Celsius and 1 atmosphere, all equally stable. A more technical way to state this is to say the liquid water has *two degrees of freedom*. Here, the word "degree" has a completely different meaning than it does when used to denote a unit of temperature or angle. In this context, "degree" may be thought of loosely as "dimension." A cube has three physical dimensions, a square has two and a line has one. A point within a cube has three degrees of freedom (motion), while a point within a square only has two, and a point along a line only has one. Here, we use the word "degree" to denote the number of independent ways a system may change. For areas bounded by phase transition curves in a phase diagram, *pressure* and *temperature* are the two "free" variables, because within those bounded areas we may freely alter pressure without altering temperature, and visa-versa.

Such is not the case at any point lying on one of the phase transition curves. Any point along a curve is geometrically defined by a pair of coordinates, which means that for a two-phase mixture in equilibrium there will be exactly one temperature value valid for each unique pressure value. At any point along a phase transition curve, pressure and temperature are not independent variable, but rather are *related*. For any single substance, there is only one degree of freedom along any point of a phase transition curve.

To illustrate this concept, suppose we equip a closed vessel containing water with both a thermometer and a pressure gauge. The thermometer measures the temperature of this water, while the pressure gauge measures the pressure of the water. A burner beneath the vessel adds heat to alter the water's temperature, and a pump adds water to the vessel to alter the pressure inside:



So long as the water is all liquid (one phase), we may adjust its pressure and temperature independently. In this state, the system has two thermodynamic degrees of freedom.

However, if the water becomes hot enough to boil, creating a system of two phases in direct contact with each other (equilibrium), we will find that pressure and temperature become linked: one cannot alter one without altering the other. For a steam boiler, operation at a given steam

pressure thus *defines* the temperature of the water, and visa-versa. In a single-component, two-phase system, there is only one degree of thermodynamic freedom.

Our freedom to alter pressure and temperature becomes even more restricted if we ever reach the *triple point* of the substance. For water, this occurs (only) at a pressure of -14.61 PSIG (0.006 atmospheres) and a temperature of 0.01 degrees Celsius: the coordinates where all three phase transition curves intersect on the phase diagram. In this state, where solid (ice), liquid (water), and vapor (steam) coexist, there are zero degrees of thermodynamic freedom. Both the temperature and pressure are *locked* at these values until one or more of the phases disappears.

The relationship between degrees of freedom and phases is expressed neatly by *Gibbs' Phase Rule* – the sum of phases and degrees of freedom equals the number of substances ("components") plus two:

$$n_{freedom} + n_{phase} = n_{substance} + 2$$

We may simplify Gibbs' rule for systems of just one substance (1 "component") by saying the number of degrees of freedom plus phases in direct contact with each other is always equal to three. So, a vessel filled with nothing but liquid water (one component, one phase) will have two thermodynamic degrees of freedom: we may change pressure or temperature independently of one another. A vessel containing nothing but boiling water (two phases – water and steam, but still only one component) has just one thermodynamic degree of freedom: we may change pressure and temperature, but just not independently of one another. A vessel containing water at its triple point (three phases, one component) has no thermodynamic freedom at all: both temperature and pressure are fixed[28] so long as all three phases coexist in equilibrium.

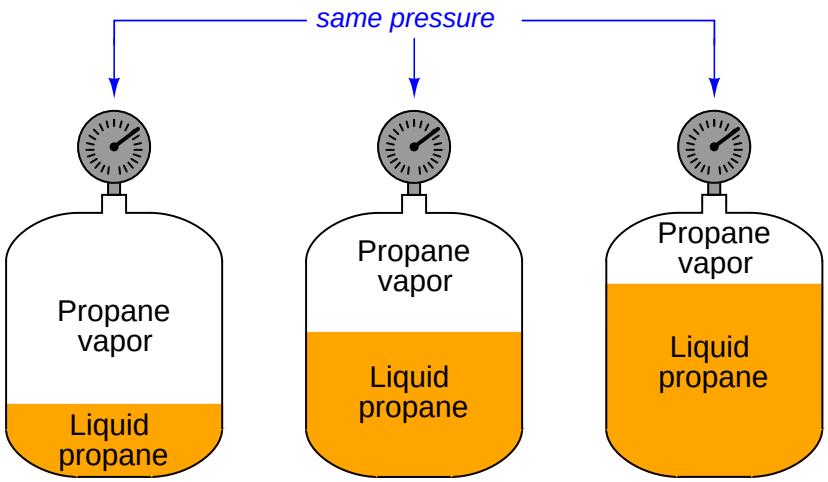
2.8.9 Applications of phase changes

Applications of phase changes abound in industrial and commercial processes. Some of these applications exploit phase changes for certain production goals, such as the storage and transport of energy. Others merely serve to illustrate certain phenomena such as latent heat and degrees of thermodynamic freedom. This subsection will highlight several different processes for your learning benefit.

[28] The non-freedom of both pressure and temperature for a pure substance at its triple point means we may exploit different substances' triple points as *calibration standards* for both pressure and temperature. Using suitable laboratory equipment and samples of sufficient purity, anyone in the world may force a substance to its triple point and calibrate pressure and/or temperature instruments against that sample.

Propane storage tanks

A common example of a saturated liquid/vapor (two-phase) system is the internal environment of a propane storage tank, such as the kind commonly used to store propane fuel for portable stoves and gas cooking grills. If multiple propane storage tanks holding different volumes of liquid propane are set side by side, pressure gauges attached to each tank will all register the exact same pressure:



This is counter-intuitive, as most people tend to think the fullest tank should register the highest pressure (having the least space for the vapor to occupy). However, since the interior of each tank is a liquid/vapor system in equilibrium, the pressure is defined by the point on the liquid/vapor transition curve on the phase diagram for pure propane matching the tanks' *temperature*. Thus, the pressure gauge on each tank actually functions as a *thermometer*, since pressure is a direct function of temperature for a saturated liquid/vapor system and therefore cannot change without a corresponding change in temperature. This is a thermodynamic system with just *one* degree of freedom.

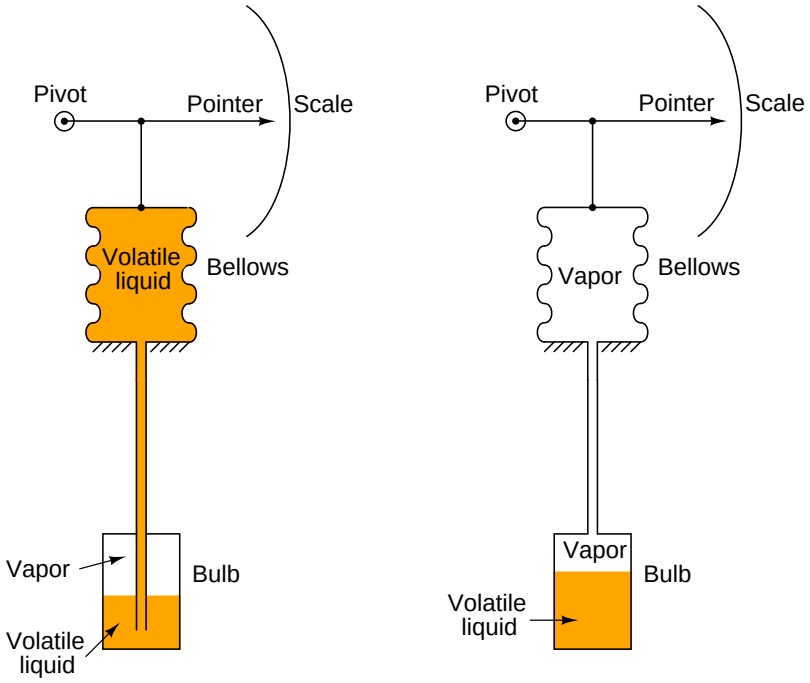
Storage tanks containing liquid/vapor mixtures in equilibrium present unique safety hazards. If ever a rupture were to occur in such a vessel, the resulting decrease in pressure causes the liquid to spontaneously boil, halting any further decrease in pressure. Thus, a punctured propane tank does not lose pressure in the same manner than a punctured compressed air tank loses pressure. This gives the escaping vapor more "power" to worsen the rupture, as its pressure does not fall off over time the way it would in a simple compressed-gas application. As a result, relatively small punctures can and often do grow into catastrophic ruptures, where all liquid previously held inside the tank escapes and flashes into vapor, generating a vapor cloud of surprisingly large volume[29].

Compounding the problem of catastrophic tank rupture is the fact that propane happens to be highly flammable. The thermodynamic properties of a boiling liquid combined with the chemical property of flammability in air makes propane tank explosions particularly violent. Fire fighters often refer to this as a *BLEVE*: a *Boiling Liquid Expanding Vapor Explosion*.

[29] Steam boilers exhibit this same explosive tendency. The expansion ratio of water to steam is on the order of a thousand to one (1000:1), making steam boiler ruptures very violent even at relatively low operating pressures.

Class II Filled-bulb thermometers

This same pressure-temperature interdependence finds application in a type of temperature measurement instrument called a *Class II filled-bulb*, where a metal bulb, tube, and pressure-sensing element are all filled with a saturated liquid/vapor mixture:



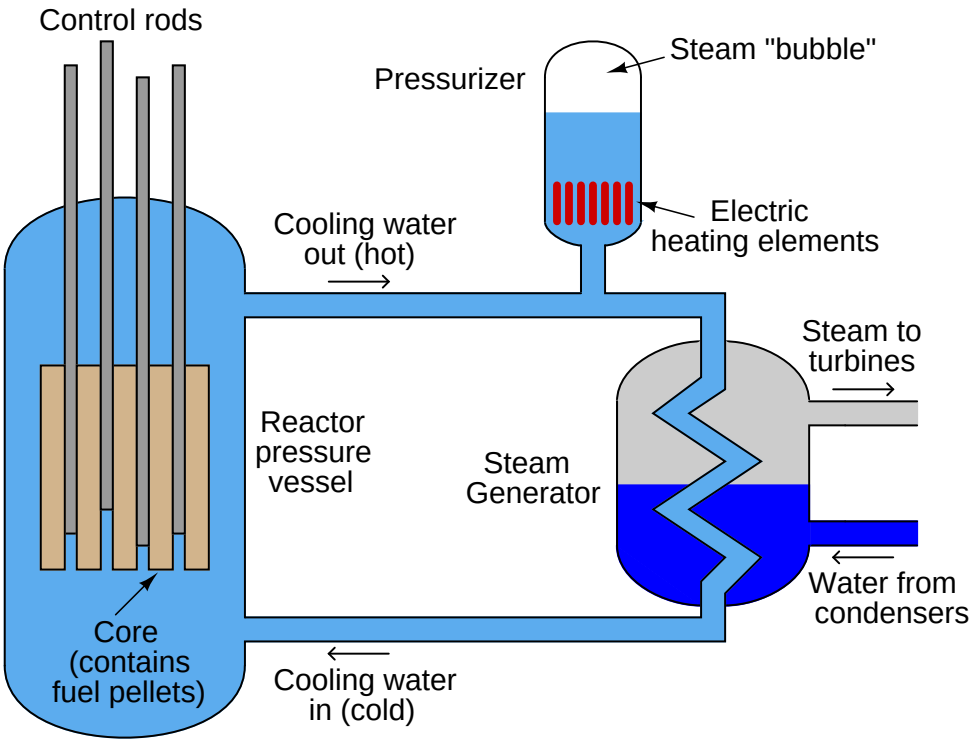
Heat applied to the bulb literally "boils" the liquid inside until its pressure reaches the equilibrium point with temperature. As the bulb's temperature increases, so does the pressure throughout the sealed system, indicating at the operator display where a bellows (or some other pressure-sensing element) moves a pointer across a calibrated scale.

The only difference between the two filled-bulb thermometers shown in the illustration is which end of the instrument is warmer. The Class IIA system on the left (where liquid fills the pressure-indicating element) is warmer at the bulb than at the indicating end. The Class IIB system on the right (where vapor fills the indicating bellows) has a cooler bulb than the indicating bellows. The long length and small internal diameter of the connecting tube prevents any substantial heat transfer from one end of the system to the other, allowing the sensing bulb to easily be at a different temperature than the indicating bellows. Both types of Class II thermometers work the same[30], the indicated pressure being a strict function of the bulb's temperature where the liquid and vapor coexist in equilibrium.

[30] Class IIA systems do suffer from *elevation error* where the indicator may read a higher or lower temperature than it should due to hydrostatic pressure exerted by the column of liquid inside the tube connecting the indicator to the sensing bulb. Class IIB systems do not suffer from this problem, as the gas inside the tube exerts no pressure over an elevation.

Nuclear reactor pressurizers

Nuclear reactors using pressurized water as the moderating and heat-transfer medium must maintain the water coolant in liquid form despite the immense heat output of the reactor core, to avoid the formation of steam bubbles which could lead to destructive "hot spots" inside the reactor. The following diagram shows a simplified[31] pressurized water reactor (PWR) cooling system:



In order to maintain a liquid-only cooling environment for the reactor core, the water is held at a pressure too high for boiling to occur inside the reactor vessel. Referencing the phase diagram for water, the operating point of the reactor core is maintained *above* the liquid/vapor phase transition line by an externally supplied pressure.

This excess pressure comes from a device in the primary coolant loop called a *pressurizer*. Inside the pressurizer is an array of immersion-style electric heater elements. The pressurizer is essentially an electric boiler, purposely boiling the water inside at a temperature greater[32] than that reached by the reactor core itself.

[31] Circulation pumps and a multitude of accessory devices are omitted from this diagram for the sake of simplicity.

[32] This is another example of an important thermodynamic concept: the distinction between *heat* and *temperature*. While the temperature of the pressurizer heating elements exceeds that of the reactor core, the total heat output of course does not. Typical comparative values for pressurizer power versus reactor core power are 1800 kW versus 3800 MW, respectively. The pressurizer heating elements don't have to dissipate much power (compared to the reactor core) because the pressurizer is not being cooled by a forced convection of water like the reactor core is.

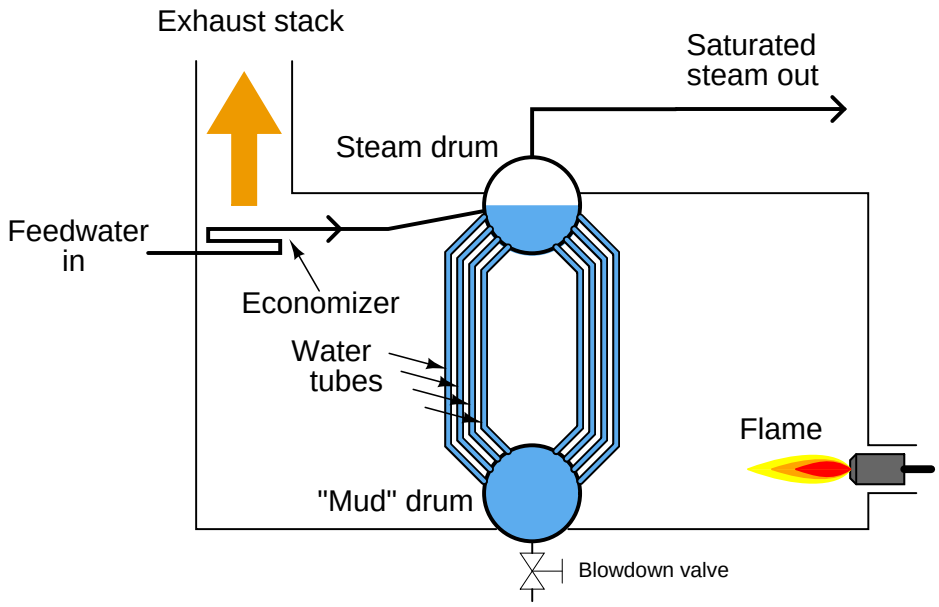
By maintaining the water temperature inside the pressurizer greater than at the reactor core, the water flowing through the reactor core literally *cannot* boil. The water/vapor equilibrium inside the pressurizer is a system with one degree of freedom (pressure and temperature linked), while the water-only environment inside the reactor core has two degrees of freedom (temperature may vary to any amount below the pressurizer's temperature without water pressure changing at all). Thus, the pressurizer functions like the temperature-sensing bulb of a *gigantic* Class IIA filled-bulb thermometer, with a liquid/vapor equilibrium inside the pressurizer vessel and liquid only inside the reactor vessel and all other portions of the primary coolant loop. Reactor pressure is then controlled by the temperature inside the pressurizer, which is in turn controlled by the amount of power applied to the heating element array[33].

Steam boilers

Boilers in general (the nuclear reactor system previously described being just one example of a large "power" boiler) are outstanding examples of phase change applied to practical use. The purpose of a boiler is to convert water into steam, sometimes for heating purposes, sometimes as a means of producing mechanical power (through a steam engine), sometimes for chemical processes requiring pressurized steam as a reactant, sometimes for utility purposes (maintenance-related cleaning, process vessel purging, sanitary disinfection, fire suppression, etc.) or all of the above. Steam is a tremendously useful substance in many industries, so you will find boilers in use at almost every industrial facility.

[33] In this application, the heaters are the *final control element* for the reactor pressure control system.

A simplified diagram of a basic *water tube* boiler appears here:



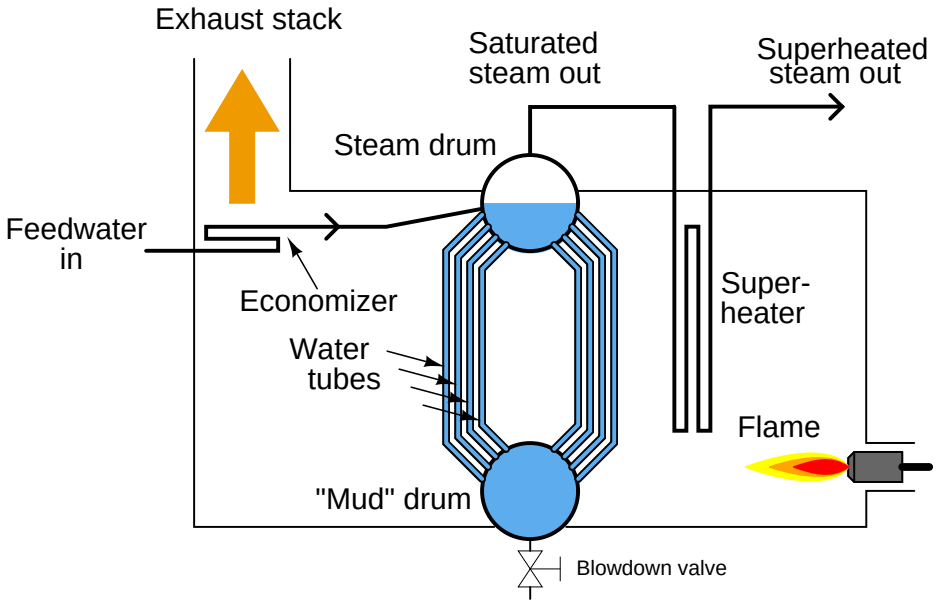
Water enters the boiler through a heat exchanger in the stack called an *economizer*. This allows cold water to be pre-heated by the relatively cool exhaust gases before they exit the stack. After pre-heating in the economizer, the water enters the boiler itself, where water circulates by natural convection ("thermosiphon") through a set of tubes exposed to high-temperature fire. Steam collects in the "steam drum," where it is drawn off through a pipe at the top. Since this steam is in direct contact with the boiling water, it will be at the same temperature as the water, and the steam/water environment inside the steam drum represents a two-phase system with only one degree of freedom. With just a single degree of freedom, steam temperature and pressure are direct functions of each other – coordinates at a single point along the liquid/vapor phase transition line of water's phase diagram. One cannot change one variable without changing the other.

Consulting a steam table[34], you will find that the temperature required to boil water at a pressure of 120 PSIG is approximately 350 degrees Fahrenheit. Thus, a steam boiler operating at that pressure will have its temperature fixed at 350 degrees. The only way to increase pressure in that boiler is to increase its temperature, and visa-versa.

When steam is at the same temperature as the boiling water it came from, it is referred to as *saturated* steam. Steam in this form is very useful for heating and cleaning, but not as much for operating mechanical engines or for process chemistry. If saturated steam loses any temperature at all (by losing its latent heat), it immediately condenses back into water. Liquid water can cause major mechanical problems inside steam engines (although "wet" steam works wonderfully well as a cleaning agent!), and so steam must be made completely "dry" for some process applications.

[34]Since the relationship between saturated steam pressure and temperature does not follow a simple mathematical formula, it is more practical to consult published tables of pressure/temperature data for steam. A great many engineering manuals contain steam tables, and in fact entire books exist devoted to nothing but steam tables.

The way this is done is by a process known as *superheating*. If steam exiting the steam drum of a boiler is fed through another heat exchanger inside the firebox so it may receive more heat, its temperature will rise beyond the saturation point. This steam is now said to be *superheated*:



Superheated steam is absolutely dry, containing no liquid water at all. It is therefore safe to use as a fluid medium for engines (piston and turbine alike) and as a process reactant where liquid water is not tolerable. The difference in temperature between superheated steam and saturated steam at any given pressure is the amount of *superheat*. For example, if saturated steam at 350 degrees Fahrenheit and 120 PSI drawn from the top of the steam drum in a boiler is heated to a higher temperature of 380 degrees Fahrenheit (at the same pressure of 120 PSI), it is said to have 30 degrees (Fahrenheit) of superheat.

Fruit crop freeze protection

An interesting application of phase changes and latent heat occurs in agriculture. Fruit growers, needing to protect their budding crop from the damaging effects of a late frost, will spray water over the fruit trees to maintain the sensitive buds above freezing temperature. As cold air freezes the water, the water's latent heat of fusion prevents the temperature at the ice/water interface from dropping below 32 degrees Fahrenheit. So long as liquid water continues to spray over the trees, the buds' temperature *cannot* fall below freezing. Indeed, the buds cannot even freeze in this condition, because once they cool down to the freezing point, there will be no more temperature difference between the freezing water and the buds. With no difference of temperature, no heat will transfer out of the buds. With no heat loss, water inside the buds cannot change phase from liquid to solid (ice) even if held at the freezing point for long periods of time, thus preventing freeze damage[35]. Only if the buds are exposed to cold air (below the freezing point), or the water turns completely to ice and no longer holds stable at the freezing point, can the buds themselves ever freeze solid.

2.9 Fluid mechanics

A *fluid* is any substance having the ability to *flow*: to freely change shape and move under the influence of a motivating force. Fluid motion may be analyzed on a microscopic level, treating each fluid molecule as an individual projectile body. This approach is extraordinarily tedious on a practical level, but still useful as a simple model of fluid motion.

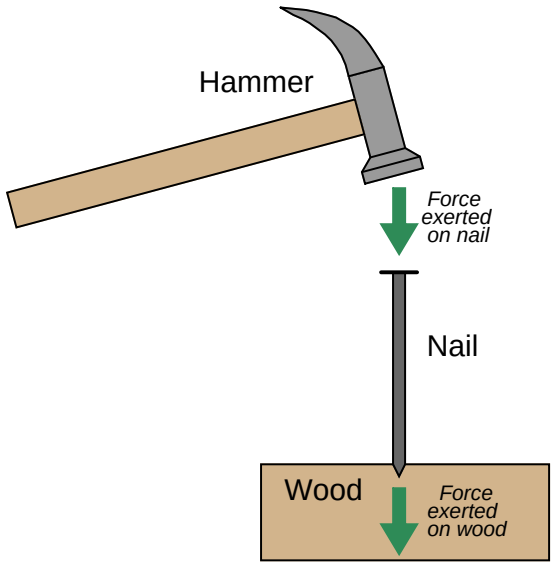
Some fluid properties are accurately predicted by this model, especially predictions dealing with potential and kinetic energies. However, the ability of a fluid's molecules to independently move give it unique properties that solids do not possess. One of these properties is the ability to effortlessly transfer *pressure*, defined as force applied over area.

[35] An experiment illustrative of this point is to maintain an ice-water mixture in an open container, then to insert a sealed balloon containing liquid water into this mixture. The water inside the balloon will eventually equalize in temperature with the surrounding ice-water mix, but it will not itself freeze. Once the balloon's water reaches 0 degrees Celsius, it stops losing heat to the surrounding ice-water mix, and therefore cannot make the phase change to solid form.

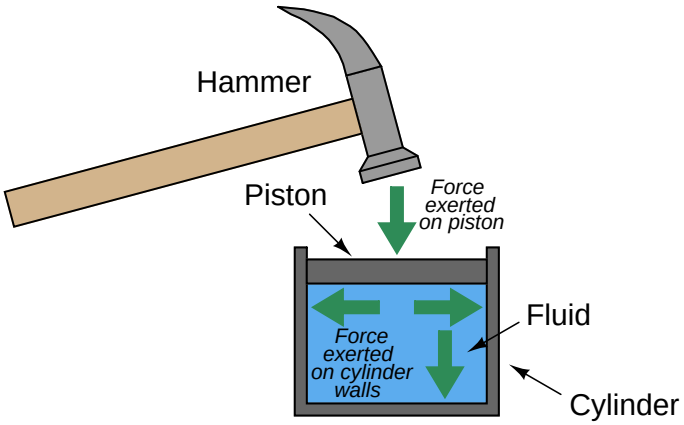
2.9.1 Pressure

The common phases of matter are *solid*, *liquid*, and *gas*. Liquids and gases are fundamentally distinct from solids in their intrinsic inability to maintain a fixed shape. In other words, liquids and gases tend to fill whatever solid containers they are held in. Similarly, both liquids and gases both have the ability to flow, which is why they are collectively called *fluids*.

Due to their lack of definite shape, fluids tend to disperse any force applied to them. This stands in marked contrast to solids, which tend to transfer force with the direction unchanged. Take for example the force transferred by a nail, from a hammer to a piece of wood:



The impact of the hammer's blow is directed straight through the solid nail into the wood below. Nothing surprising here. But now consider what a fluid would do when subjected to the same hammer blow:



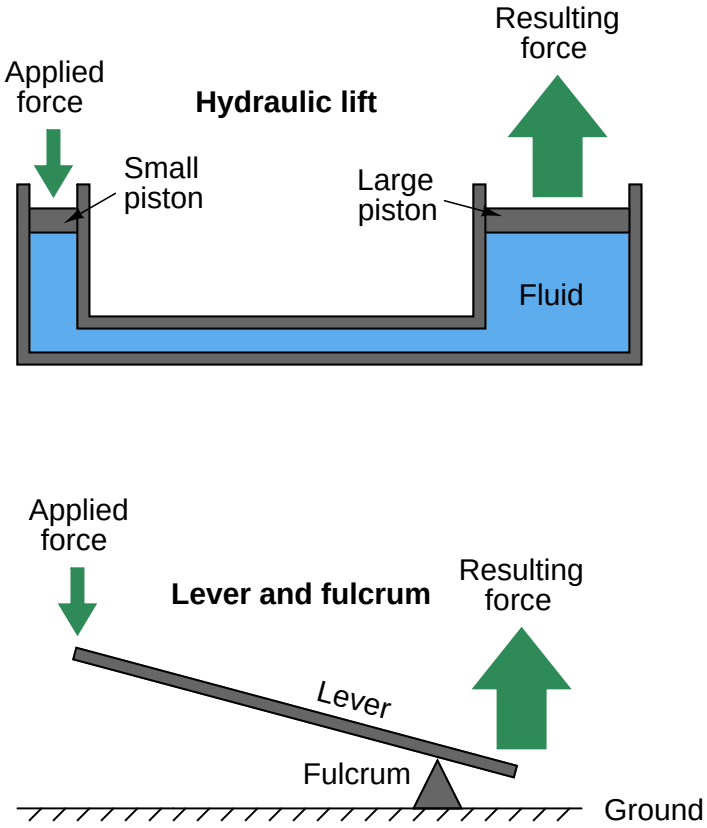
Given the freedom of a fluid's molecules to move about, the impact of the hammer blow becomes directed *everywhere* against the inside surface of the container (the cylinder). This is true for all fluids: liquids and gases alike. The only difference between the behavior of a liquid and a gas in the same scenario is that the gas will compress (i.e. the piston will move down as the hammer struck it), whereas the liquid will not compress (i.e. the piston will remain in its resting position). Gases yield under pressure, liquids do not.

It is very useful to quantify force applied to a fluid in terms of force per unit area, since the force applied to a fluid becomes evenly dispersed in all directions to the surface containing it. This is the definition of *pressure* (P): how much force (F) is distributed across how much area (A).

$$P = \frac{F}{A}$$

In the metric system, the standard unit of pressure is the *Pascal* (Pa), defined as one Newton (N) of force per square meter (m^2) of area. In the British system of measurement, the standard unit of pressure is the *PSI*: pounds (lb) of force per square inch (in^2) of area. Pressure is often expressed in units of kilo-pascals (kPa) when metric units are used because one pascal is a rather low pressure in most engineering applications.

The even distribution of force throughout a fluid has some very practical applications. One application of this principle is the *hydraulic lift*, which functions somewhat like a fluid lever:



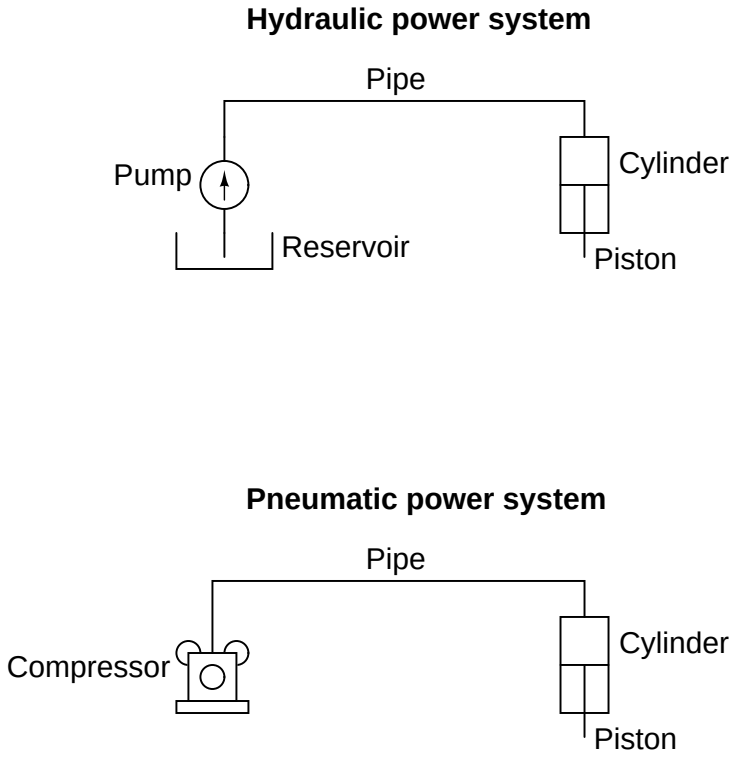
Force applied to the small piston creates a pressure throughout the fluid. That pressure exerts a greater force on the large piston than what is exerted on the small piston, by a factor equal to the ratio of piston areas. If the large piston has five times the area of the small piston, force will be multiplied by five. Just like with the lever, however, there must be a trade-off so we do not violate the Conservation of Energy. The trade-off for increased force is decreased distance, whether in the lever system or in the hydraulic lift system. If the large piston generates a force five times greater than what was input at the small piston, it will move only one-fifth the distance that the small piston does. In this way, energy in equals energy out (remember that *work*, which is equivalent to energy, is calculated by multiplying force by parallel distance traveled).

For those familiar with electricity, what you see here in either the lever system or the hydraulic lift is analogous to a *transformer*: we can step AC voltage up, but only by reducing AC current. Being a passive device, a transformer cannot boost power. Therefore, power out can never be greater than power in, and given a perfectly efficient transformer, power out will always be precisely equal to power in:

$$\text{Power} = (\text{Voltage in})(\text{Current in}) = (\text{Voltage out})(\text{Current out})$$

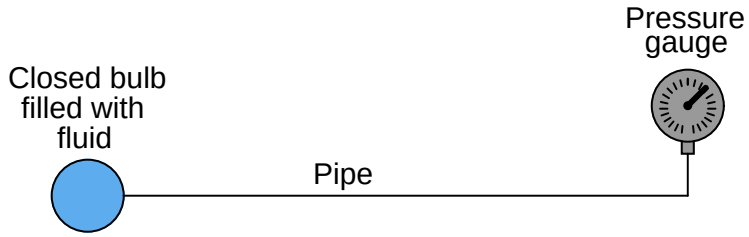
$$\text{Work} = (\text{Force in})(\text{Distance in}) = (\text{Force out})(\text{Distance out})$$

Fluid may be used to transfer power just as electricity is used to transfer power. Such systems are called *hydraulic* if the fluid is a liquid (usually oil), and *pneumatic* if the fluid is a gas (usually air). In either case, a machine (pump or compressor) is used to generate a continuous fluid pressure, pipes are used to transfer the pressurized fluid to the point of use, and then the fluid is allowed to exert a force against a piston or a set of pistons to do mechanical work:



To learn more about fluid power systems, refer to section 10.2 on page 392.

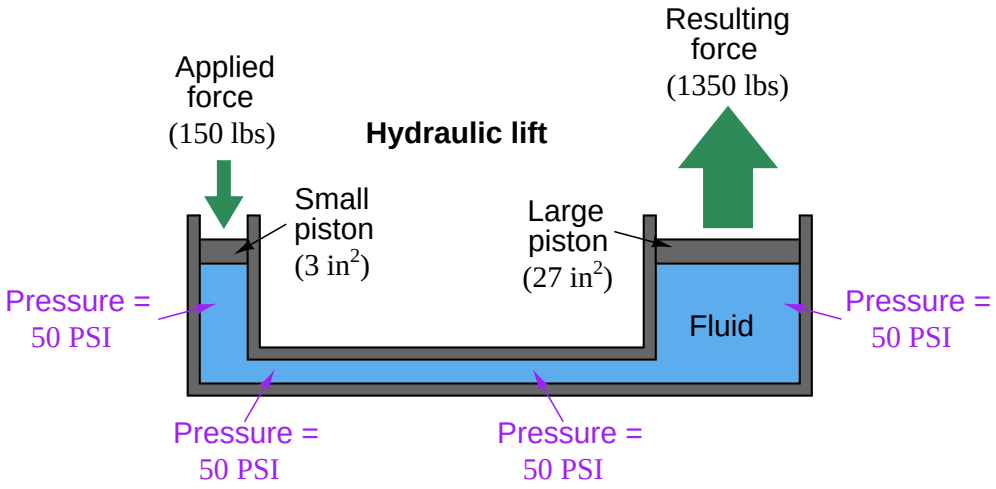
An interesting use of fluid we see in the field of instrumentation is as a *signaling medium*, to transfer information between places rather than to transfer power between places. This is analogous to using electricity to transmit voice signals in telephone systems, or digital data between computers along copper wire. Here, fluid pressure represents some other quantity, and the principle of force being distributed equally throughout the fluid is exploited to transmit that representation to some distant location, through piping or tubing:



This illustration shows a simple temperature-measuring system called a *filled bulb*, where an enclosed bulb filled with fluid is exposed to a temperature that we wish to measure. A rise in temperature causes the fluid pressure to increase, which is sent to the gauge far away through the pipe, and registered at the gauge. The purpose of the fluid here is two-fold: first to sense temperature, and second to relay this temperature measurement a long distance away to the gauge. The principle of even pressure distribution allows the fluid to act as a signal medium to convey the information (bulb temperature) to a distant location.

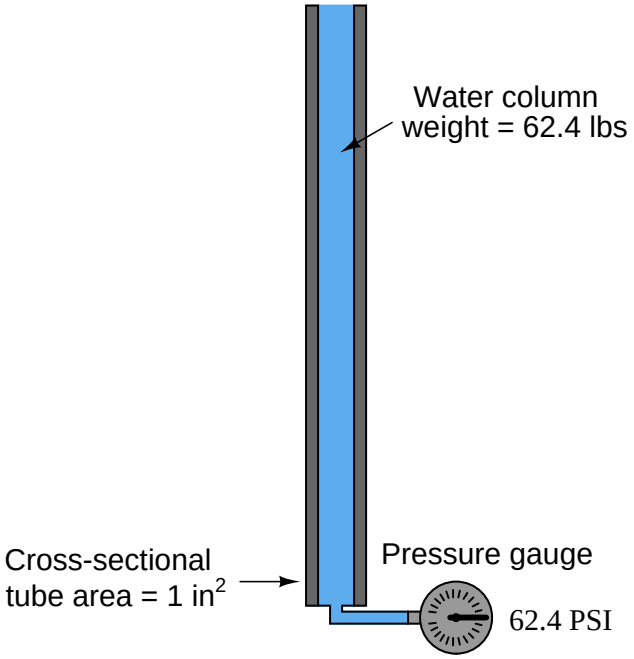
2.9.2 Pascal's Principle and hydrostatic pressure

We learned earlier that fluids tend to evenly distribute the force applied to them. This tendency is known as *Pascal's principle*, and it is the fundamental principle upon which fluid power and fluid signaling systems function. In the example of a hydraulic lift given earlier, we assume that the pressure throughout the fluid pathway is equal:



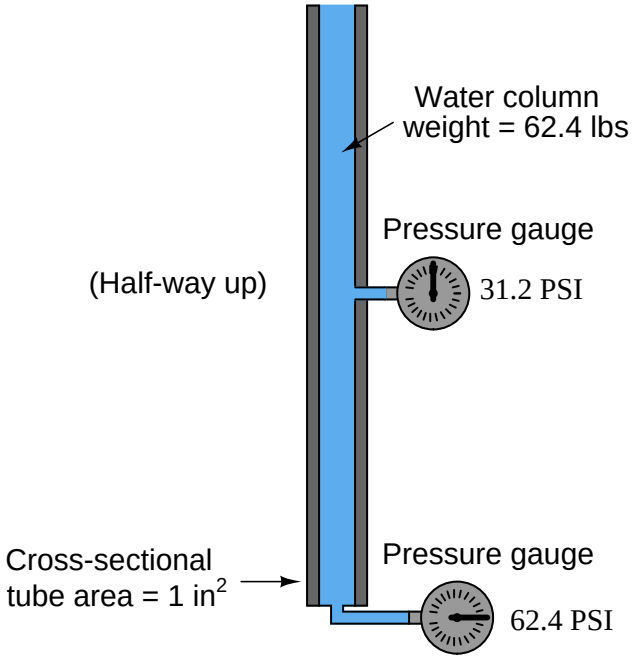
The key assumption we make here is that the only force we need to consider on the fluid is the force exerted on the small piston (150 pounds). If this is truly the only force acting on the fluid, then it will likewise be the only source of fluid pressure, and pressure will simply be equal to force divided by area (150 pounds ÷ 3 square inches = 50 PSI).

However, when we are dealing with tall columns of fluid, and/or dense fluids, there is another force we must consider: the weight of the fluid itself. Suppose we took a cubic foot of water which weighs approximately 62.4 pounds, and poured it into a tall, vertical tube with a cross-sectional area of 1 square inch:



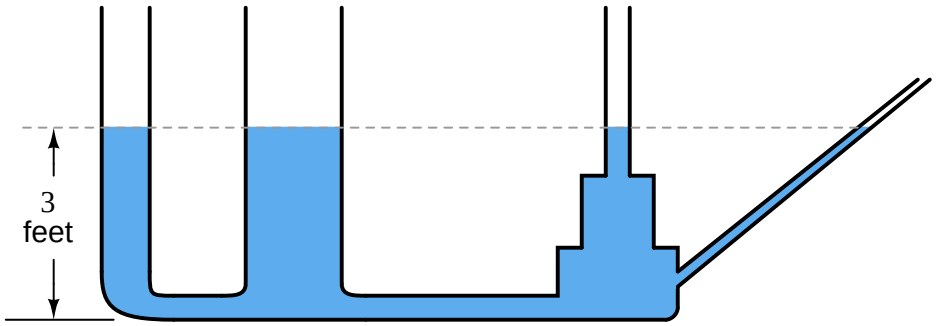
Naturally, we would expect the pressure measured at the bottom of this tall tube to be 62.4 pounds per square inch, since the entire column of water (weighing 62.4 pounds) has its weight supported by one square inch of area.

If we placed another pressure gauge mid-way up the tube, though, how much pressure would it register? At first you might be inclined to say 62.4 PSI as well, because you learned earlier in this lesson that fluids naturally distribute force throughout their bulk. However, in this case the pressure is *not* the same mid-way up the column as it is at the bottom:



The reason for this apparent discrepancy is that the source of pressure in this fluid system comes from the weight of the water column itself. Half-way up the column, the water only experiences half the total weight (31.2 pounds), and so the pressure is half of what it is at the very bottom. We never dealt with this effect before, because we assumed the force exerted by the piston in the hydraulic lift was so large it "swamped" the weight of the fluid itself. Here, with our very tall column of water (144 feet tall!), the effect of gravity upon the water's mass is quite substantial. Indeed, without a piston to exert an external force on the water, weight is the *only* source of force we have to consider when calculating pressure.

An interesting fact about pressure generated by a column of fluid is that the width or shape of the containing vessel is irrelevant: the *height* of the fluid column is the only dimension we need to consider. Examine the following tube shapes, all connected at the bottom:



Since the force of fluid weight is generated only along the axis of gravitational attraction (straight down), that is the only axis of measurement important in determining "hydrostatic" fluid pressure.

The fixed relationship between the vertical height of a water column and pressure is such that sometimes water column height is used as a unit of measurement for pressure. That is, instead of saying "30 PSI," we could just as correctly quantify that same pressure as 830.4 inches of water ("W.C. or "H_2O), the conversion factor being approximately 27.68 inches of vertical water column per PSI.

As one might guess, the *density* of the fluid in a vertical column has a significant impact on the hydrostatic pressure that column generates. A liquid twice as dense as water, for example, will produce twice the pressure for a given column height. For example, a column of this liquid (twice as dense as water) 14 inches high will produce a pressure at the bottom equal to 28 inches of water (28 "W.C.), or just over 1 PSI. An extreme example is liquid mercury, which is over 13.5 times as dense as water. Due to its exceptional density and ready availability, the height of a mercury column is also used as a standard unit of pressure measurement. For instance, 25 PSI could be expressed as 50.9 inches of mercury ("Hg), the conversion factor being approximately 2.036 inches of vertical mercury column per PSI.

The mathematical relationship between vertical liquid height and hydrostatic pressure is quite simple, and may be expressed by either of the following formulae:

$$P = \rho g h$$

$$P = \gamma h$$

Where,

P = Hydrostatic pressure in units of weight per square area unit: Pascals (N/m^2) or lb/ft^2

ρ = Mass density of liquid in kilograms per cubic meter (metric) or slugs per cubic foot (British)

g = Acceleration of gravity (9.8 meters per second squared or 32 feet per second squared)

γ = Weight density of liquid in newtons per cubic meter (metric) or pounds per cubic foot (British)

h = Vertical height of liquid column

Dimensional analysis vindicates these formulae in their calculation of hydrostatic pressure. Taking the second formula as an example:

$$P = \gamma h$$

$$\left[\frac{\text{lb}}{\text{ft}^2}\right] = \left[\frac{\text{lb}}{\text{ft}^3}\right]\left[\frac{\text{ft}}{1}\right]$$

As you can see, the unit of "feet" in the height term cancels out one of the "feet" units in the denominator of the density term, leaving an answer for pressure in units of pounds per *square* foot. If one wished to set up the problem so the answer presented in a more common pressure unit such as pounds per square *inch*, both the liquid density and height would have to be expressed in appropriate units (pounds per cubic *inch* and *inches*, respectively).

Applying this to a realistic problem, consider the case of a tank filled with 8 feet (vertical) of castor oil, having a weight density of 60.5 pounds per cubic foot. This is how we would set up the formula to calculate for hydrostatic pressure at the bottom of the tank:

$$P = \gamma h$$

$$P = \left(\frac{60.5 \text{ lb}}{\text{ft}^3}\right)(8 \text{ ft})$$

$$P = \frac{484 \text{ lb}}{\text{ft}^2}$$

If we wished to convert this result into a more common unit such as PSI (pounds per square inch), we could do so using an appropriate fraction of conversion units:

$$P = \left(\frac{484 \text{ lb}}{\text{ft}^2}\right)\left(\frac{1 \text{ ft}^2}{144 \text{ in}^2}\right)$$

$$P = \frac{3.36 \text{ lb}}{\text{in}^2} = 3.36 \text{ PSI}$$

2.9.3 Fluid density expressions

Fluid density is commonly expressed as a ratio in comparison to pure water at standard temperature[36]. This ratio is known as *specific gravity*. For example, the specific gravity of glycerin may be determined by dividing the density of glycerin by the density of water:

$$\text{Specific gravity of any liquid} = \frac{D_{liquid}}{D_{water}}$$

$$\text{Specific gravity of glycerin} = \frac{D_{glycerin}}{D_{water}} = \frac{78.6 \text{ lb/ft}^3}{62.4 \text{ lb/ft}^3} = 1.26$$

The density of gases may also be expressed in ratio form, except the standard of comparison is ambient air instead of water. Chlorine gas, for example, has a specific gravity of 2.47 (each volumetric unit of chlorine having 2.47 times the mass of the same volume of air under identical temperature and pressure conditions). Specific gravity values for gases are sometimes called *relative gas densities* to avoid confusion with "specific gravity" values for liquids.

As with all ratios, specific gravity is a unitless quantity. In our example with glycerine, we see how the identical units of pounds per cubic foot cancel out of both numerator and denominator, to leave a quotient with no unit at all.

An alternative to expressing fluid density as a ratio of mass (or weight) to volume, or to compare it against the density of a standard fluid such as pure water or air, is to express it as the ratio of volume to mass. This is most commonly applied to vapors such as steam, and it is called *specific volume*. The relationship between specific volume and density is one of mathematical reciprocation: the reciprocal of density (e.g. pounds per cubic foot) is specific volume (e.g. cubic feet per pound). For example, consulting a table of saturated steam properties, we see that saturated steam at a pressure of 60 PSIA has a specific volume of 7.175 cubic feet per pound. Translating this into units of pounds per cubic feet, we reciprocate the value 7.175 to arrive at 0.1394 pounds per cubic foot.

Industry-specific units of measurement do exist for expressing the relative density of a fluid. These units of measurement all begin with the word "degree" much the same as for units of temperature measurement, for example:

- Degrees API *(used in the petroleum industries)*
- Degrees Baumé *(used in a variety of industries including paper manufacture and alcohol production)*
- Degrees Twaddell

The mathematical relationships between each of these "degree" units of density versus specific gravity[37] is as follows:

$$\text{Degrees API} = \frac{141.5}{\text{Specific gravity}} - 131.5$$

[36] Usually, this standard temperature is 4 degrees Celsius, the point of maximum density for water. However, sometimes the specific gravity of a liquid will be expressed in relation to the density of water at some other temperature.

[37] For each of these calculations, specific gravity is defined as the ratio of the liquid's density at 60 degrees Fahrenheit to the density of pure water, also at 60 degrees Fahrenheit.

$$\text{Degrees Twaddell} = 200 \times (\text{Specific gravity} - 1)$$

Two different formulae exist for the calculation of degrees Baumé, depending on whether the liquid in question is heavier or lighter than water. For lighter-than-water liquids:

$$\text{Degrees Baumé (light)} = \frac{140}{\text{Specific gravity}} - 130$$

Note that pure water would measure 10^o Baumé on the light scale. As liquid density decreases, the light Baumé value increases. For heavier-than-water liquids:

$$\text{Degrees Baumé (heavy)} = 145 - \frac{145}{\text{Specific gravity}}$$

Note that pure water would measure 0^o Baumé on the heavy scale. As liquid density increases, the heavy Baumé value increases.

Just to make things confusing, there are different standards for the heavy Baumé scale. Instead of the constant value 145 shown in the above equation (used throughout the United States of America), an older Dutch standard used the same formula with a constant value of 144. The *Gerlach* heavy Baumé scale uses a constant value of 146.78:

$$\text{Degrees Baumé (heavy, old Dutch)} = 144 - \frac{144}{\text{Specific gravity}}$$

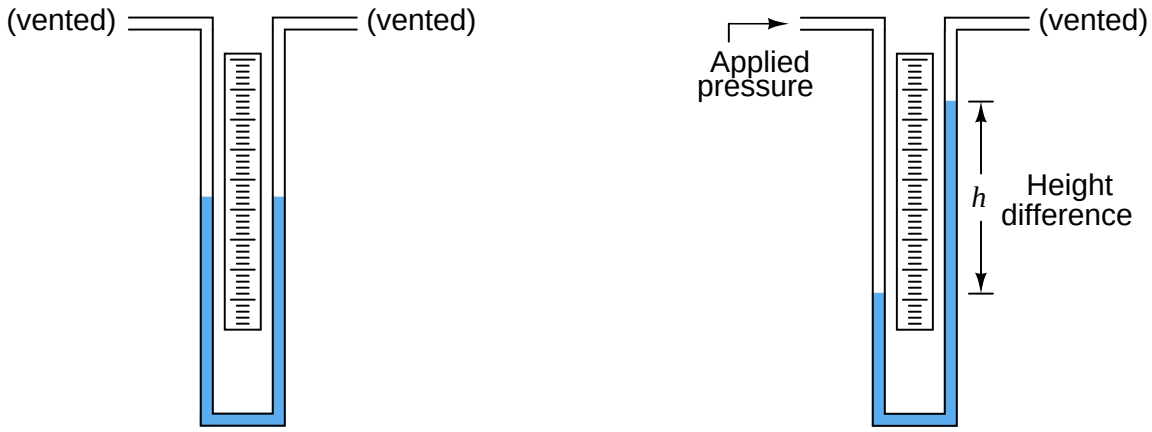
$$\text{Degrees Baumé (heavy, Gerlach scale)} = 146.78 - \frac{146.78}{\text{Specific gravity}}$$

There exists a seemingly endless array of "degree" scales used to express liquid density, scattered throughout the pages of history. For the measurement of sugar concentrations in the food industries, the unit of degrees *Balling* was invented. This scale was later revised to become the unit of degrees *Brix*, which directly corresponds to the percent concentration of sugar in the liquid. The density of tanning liquor may be measured in degrees *Bark*. Milk density may be measured in degrees *Soxhlet*. Vegetable oil density (and in older times, the density of oil extracted from sperm whales) may be measured in degrees *Oleo*.

2.9.4 Manometers

Expressing fluid pressure in terms of a vertical liquid column makes perfect sense when we use a very simple kind of motion-balance pressure instrument called a *manometer*. A manometer is nothing more than a piece of clear (glass or plastic) tubing filled with a liquid of known density, situated next to a scale for measuring distance. The most basic form of manometer is the *U-tube* manometer, shown here:

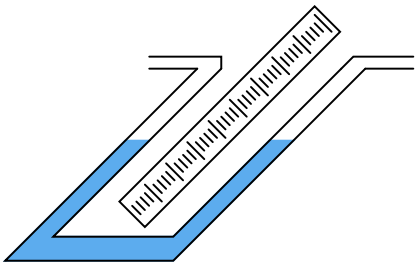
U-tube manometer



Pressure is read on the scale as the difference in height (h) between the two liquid columns. One nice feature of a manometer is it really cannot become "uncalibrated" so long as the fluid is pure and the assembly is maintained in an upright position. If the fluid used is water, the manometer may be filled and emptied at will, and even rolled up for storage if the tubes are made of flexible plastic.

We may build even more sensitive manometers by purposely inclining one or more of the tubes, so that distance read along the tube length is a fractional proportion of distance measured along the vertical:

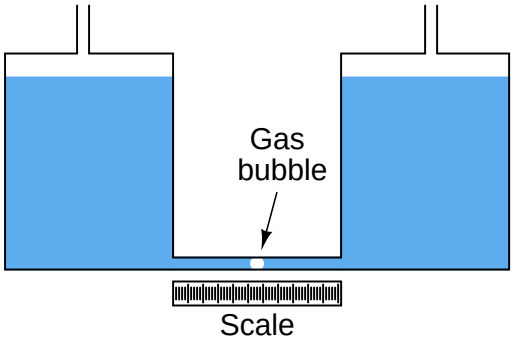
Inclined manometer



This way, a greater motion of liquid is required to generate the same hydrostatic pressure (vertical liquid displacement) than in an upright manometer, making the inclined manometer more sensitive.

If even more sensitivity is desired, we may build something called a *micromanometer*, consisting of a gas bubble trapped in a clear horizontal tube between two large vertical manometer chambers:

A simple micromanometer



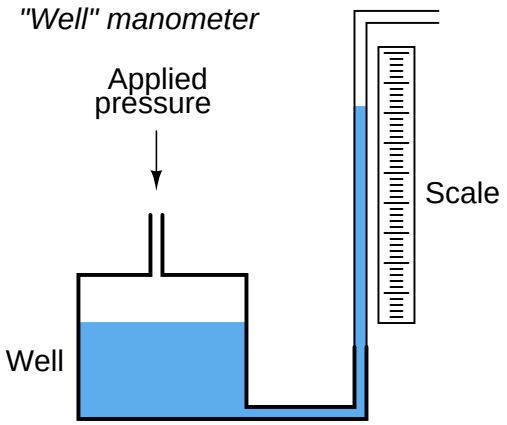
Pressure applied to the top of either vertical chamber will cause the vertical liquid columns to shift just the same as any U-tube manometer. However, the bubble trapped in the clear horizontal tube will move much further than the vertical displacement of either liquid column, owing to the huge difference in cross-sectional area between the vertical chambers and the horizontal tube. This amplification of motion makes the micromanometer exceptionally sensitive to small pressures.

Using water as the working liquid in a standard U-tube manometer, 1 PSI of applied gas pressure results in approximately 27.7 inches of vertical liquid column displacement (i.e. 27.7 inches of height *difference* between the two water columns). This relatively large range of motion limits the usefulness of water manometers to modest pressures only. If we wished to use a water manometer to measure the pressure of compressed air in an industrial pneumatic supply system at approximately 100 PSI, the manometer would have to be in excess of 230 feet tall! Clearly, a water manometer would not be the proper instrument to use for such an application.

However, water is not the only viable liquid for use in manometers. We could take the exact same clear U-tube and fill it partially full of liquid *mercury* instead, which is substantially denser than water. In a mercury manometer, 1 PSI of applied gas pressure results in very slightly more than 2 inches of liquid column displacement. A mercury manometer applied to the task of measuring air pressure in an industrial pneumatic system would only have to be 17 feet tall – still quite large and cumbersome[38] for a measuring instrument, but not impossible to construct or to use.

A common form of manometer seen in calibration laboratories is the *well* type, consisting of a single vertical tube and a relatively large reservoir (called the "well") acting as the second column:

[38] A colleague of mine told me once of working in an industrial facility with a very old steam boiler, where boiler steam pressure was actually indicated by tall mercury manometers reaching from floor to ceiling. Operations personnel had to climb a ladder to accurately read pressure indicated by these manometers!



Due to the well's much larger cross-sectional area, liquid motion inside of it is negligible compared to the motion of liquid inside the clear viewing tube. For all practical purposes, the only liquid motion is inside the smaller tube. Thus, the well manometer provides an easier means of reading pressure: no longer does one have to measure the difference of height between *two* liquid columns, only the height of a single column.

2.9.5 Systems of pressure measurement

Pressure measurement is often a relative thing. What we mean when we say there is 35 PSI of air pressure in an inflated car tire is that the pressure inside the tire is 35 pounds per square inch *greater than* the surrounding, ambient air pressure. It is a fact that we live and breathe in a pressurized environment. Just as a vertical column of liquid generates a hydrostatic pressure, so does a vertical column of gas. If the column of gas is very tall, the pressure generated by it will be substantial enough to measure. Such is the case with Earth's atmosphere, the pressure at sea level caused by the weight of the atmosphere is approximately 14.7 PSI.

You and I do not perceive this constant air pressure around us because the pressure inside our bodies is equal to the pressure outside our bodies. Thus our skin, which serves as a differential pressure-sensing diaphragm, detects no *difference* of pressure between the inside and outside of our bodies. The only time the Earth's air pressure becomes perceptible to us is if we rapidly ascend or descend in a vehicle, where the pressure inside our bodies does not have time to equalize with the pressure outside, and we feel the force of that differential pressure on our eardrums.

If we wish to speak of a fluid pressure in terms of how it compares to a perfect vacuum (absolute zero pressure), we specify it in terms of *absolute* units. For example, when I said earlier that the atmospheric pressure at sea level was 14.7 PSI, what I really meant is it is 14.7 PSIA (pounds per square inch *absolute*), meaning 14.7 pounds per square inch *greater than a perfect vacuum*. When I said earlier that the air pressure inside an inflated car tire was 35 PSI, what I really meant is it was 35 PSIG (pounds per square inch *gauge*), meaning 35 pounds per square inch *greater than ambient air pressure*. When units of pressure measurement are specified without a "G" or "A" suffix, it is usually (but not always!) assumed that *gauge* pressure (relative to ambient pressure) is meant.

This offset of 14.7 PSI between *absolute* and *gauge* pressures can be confusing if we must convert between different pressure units. Suppose we wished to express the tire pressure of 35 PSIG in units of inches of water column ("W.C.). If we stay in the gauge-pressure scale, all we have to do is multiply by 27.68:

$$\frac{35\ \text{PSI}}{1} \times \frac{27.68\ \text{"W.C.}}{1\ \text{PSI}} = 968.8\ \text{"W.C.}$$

Note how the fractions have been arranged to facilitate cancellation of units. The "PSI" unit in the numerator of the first fraction cancels with the "PSI" unit in the denominator of the second fraction, leaving inches of water column ("W.C.) as the only unit standing. Multiplying the first fraction (35 PSI over 1) by the second fraction (27.68 "W.C. over 1 PSI) is "legal" to do since the second fraction has a *physical* value of unity (1): being that 27.68 inches of water column is the same physical pressure as 1 PSI, the second fraction is really the number "1" in disguise. As we know, multiplying any quantity by unity does not change its value, so the result of 968.8 "W.C. we get has the exact same physical meaning as the original figure of 35 PSI. This technique of unit conversion is sometimes known as *unity fractions*, and it is discussed in more general terms in another section of this book (refer to section 2.3 beginning on page 28).

If, however, we wished to express the car's tire pressure in terms of inches of water column *absolute* (in reference to a perfect vacuum), we would have to include the 14.7 PSI offset in our calculation, and do the conversion in two steps:

$$35\ \text{PSIG} + 14.7\ \text{PSI} = 49.7\ \text{PSIA}$$

$$\frac{49.7\text{ PSIA}}{1} \times \frac{27.68\text{ "W.C.A}}{1\text{ PSIA}} = 1375.7\text{ "W.C.A}$$

The proportion between inches of water column and pounds per square inch is still the same (27.68) in the absolute scale as it is in the gauge scale. The only difference is that we included the 14.7 PSI offset in the very beginning to express the tire's pressure on the absolute scale rather than on the gauge scale. From then on, all conversions were in absolute units.

There are some pressure units that are *always* in absolute terms. One is the unit of *atmospheres*, 1 atmosphere being 14.7 PSIA. There is no such thing as "atmospheres gauge" pressure. For example, if we were given a pressure as being 4.5 atmospheres and we wanted to convert that into pounds per square inch gauge (PSIG), the conversion would be a two-step process:

$$\frac{4.5\text{ atm}}{1} \times \frac{14.7\text{ PSIA}}{1\text{ atm}} = 66.15\text{ PSIA}$$

$$66.15\text{ PSIA} - 14.7\text{ PSI} = 51.45\text{ PSIG}$$

Another unit of pressure measurement that is always absolute is the *torr*, equal to 1 millimeter of mercury column absolute (mmHgA). 0 torr is absolute zero, equal to 0 atmospheres, 0 PSIA, or -14.7 PSIG. Atmospheric pressure at sea level is 760 torr, equal to 1 atmosphere, 14.7 PSIA, or 0 PSIG.

If we wished to convert the car tire's pressure of 35 PSIG into torr, we would once again have to offset the initial value to get everything into absolute terms.

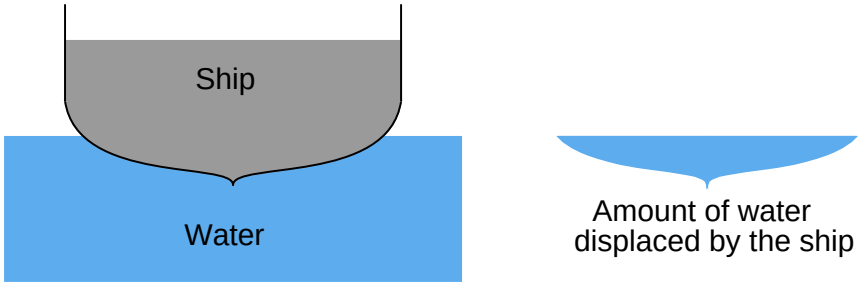
$$35\text{ PSIG} + 14.7\text{ PSI} = 49.7\text{ PSIA}$$

$$\frac{49.7\text{ PSIA}}{1} \times \frac{760\text{ torr}}{14.7\text{ PSIA}} = 2569.5\text{ torr}$$

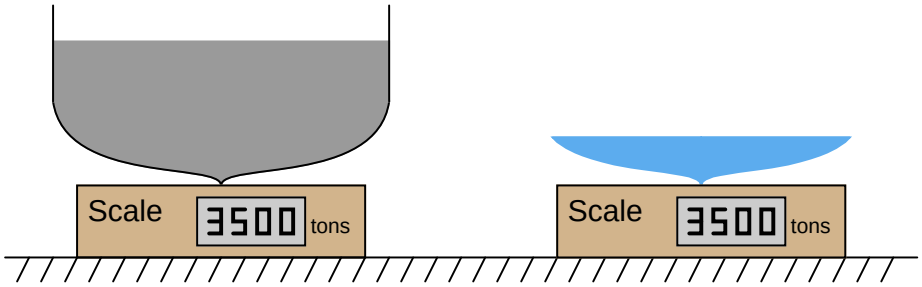
2.9.6 Buoyancy

When a solid body is immersed in a fluid, it *displaces* an equal volume of that fluid. This displacement of fluid generates an upward force on the object called the *buoyant force*. The magnitude of this force is equal to the weight of the fluid displaced by the solid body, and it is always directed exactly opposite the line of gravitational attraction. This is known as *Archimedes' Principle*.

Buoyant force is what makes ships float. A ship sinks into the water just enough so the weight of the water displaced is equal to the total weight of the ship and all it holds (cargo, crew, food, fuel, etc.):



If we could somehow measure the weight of that water displaced, we would find it exactly equals the dry weight of the ship:



Archimedes' Principle also explains why hot-air balloons and helium aircraft float. By filling a large enclosure with a gas that is less dense than the surrounding air, that enclosure experiences an upward (buoyant) force equal to the difference between the weight of the air displaced and the weight of the gas enclosed. If this buoyant force equals the weight of the craft and all it holds (cargo, crew, food, fuel, etc.), it will exhibit an apparent weight of zero, which means it will float. If the buoyant force exceeds the weight of the craft, the resultant force will cause an upward acceleration according to Newton's Second Law of motion ($F = ma$).

Submarines also make use of Archimedes' Principle, adjusting their buoyancy by adjusting the amount of water held by *ballast tanks* on the hull. Positive buoyancy is achieved by "blowing" water out of the ballast tanks with high-pressure compressed air, so the submarine weighs less (but still occupies the same hull volume and therefore displaces the same amount of water). Negative buoyancy is achieved by "flooding" the ballast tanks so the submarine weighs more. Neutral buoyancy is when the buoyant force exactly equals the weight of the submarine and the remaining water stored in

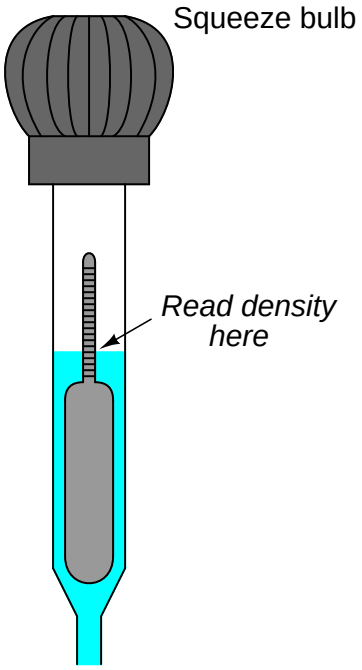
the ballast tanks, so the submarine is able to "hover" in the water with no vertical acceleration or deceleration.

An interesting application of Archimedes' Principle is the quantitative determination of an object's density by submersion in a liquid. For instance, copper is 8.96 times as dense as water, with a mass of 8.96 grams per cubic centimeter (8.96 g/cm^3) as opposed to water at 1.00 gram per cubic centimeter (1.00 g/cm^3). If we had a sample of pure, solid copper exactly 1 cubic centimeter in volume, it would have a mass of 8.96 grams. Completely submerged in pure water, this same sample of solid copper would appear to have a mass of only 7.96 grams, because it would experience a buoyant force equivalent to the mass of water it displaces (1 cubic centimeter = 1 gram of water). Thus, we see that the difference between the dry mass (mass measured in air) and the wet mass (mass measured when completely submerged in water) is the mass of the water displaced. Dividing the sample's dry mass by this mass difference (dry − wet mass) yields the ratio between the sample's mass and the mass of an equivalent volume of water, which is the very definition of specific gravity. The same calculation yields a quantity for specific gravity if *weights* instead of *masses* are used, since weight is nothing more than mass multiplied by the acceleration of gravity ($F_{weight} = mg$), and the constant g cancels out of both numerator and denominator:

$$\text{Specific Gravity } = \frac{m_{dry}}{m_{dry} - m_{wet}} = \frac{m_{dry}g}{m_{dry}g - m_{wet}g} = \frac{\text{Dry weight}}{\text{Dry weight} - \text{Wet weight}}$$

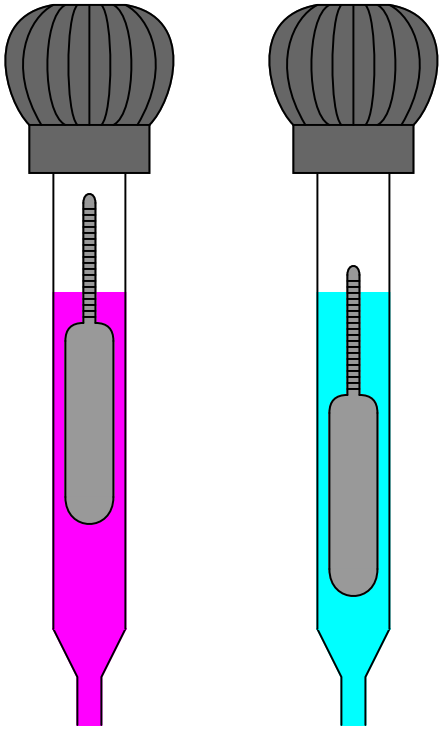
Another application of Archimedes' Principle is the use of a *hydrometer* for measuring liquid density. If a narrow cylinder of precisely known volume and weight (most of the weight concentrated at one end) is immersed in liquid, that cylinder will sink to a level dependent on the liquid's density. In other words, it will sink to a level sufficient to displace its own weight in fluid. Calibrated marks made along the cylinder's length may then serve to register liquid density in any unit desired.

A simple style of hydrometer used to measure the density of lead-acid battery electrolyte is shown in this illustration:



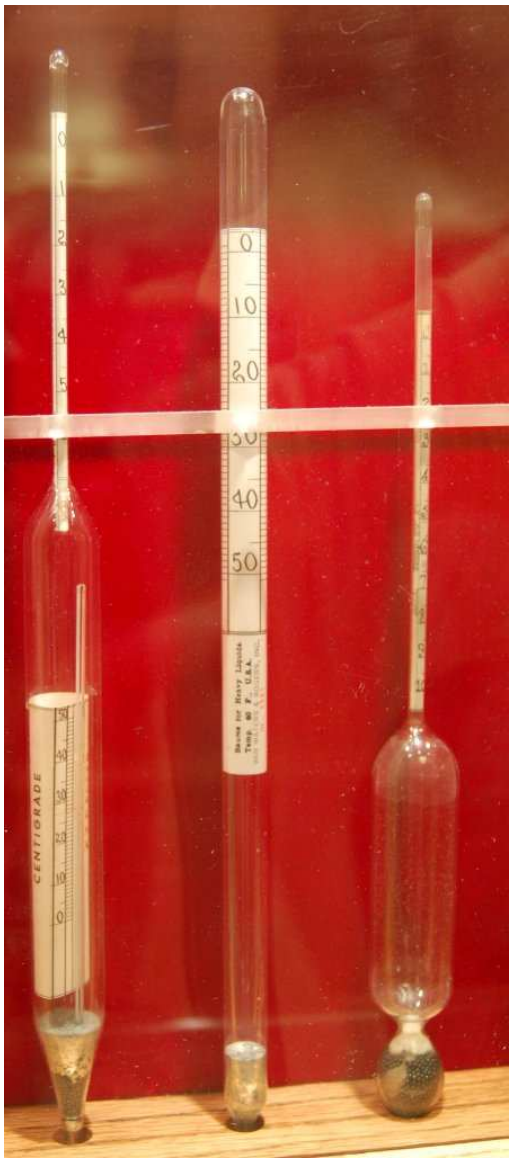
To use this hydrometer, you must squeeze the rubber bulb at the top and dip the open end of the tube into the liquid to be sampled. Relaxing the rubber bulb will draw a sample of liquid up into the tube where it immerses the float. When enough liquid has been drawn into the tube to suspend the float so that it neither rests on the bottom of the tapered glass tube or "tops out" near the bulb, the liquid's density may be read at the air/liquid interface.

A denser electrolyte liquid results in the float rising to a higher level inside the hydrometer tube:



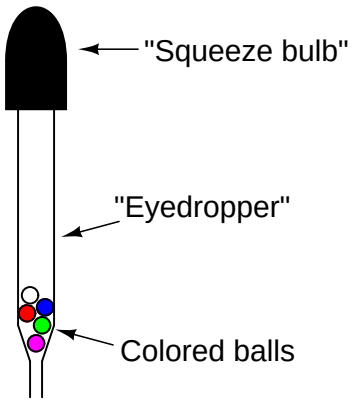
Dense liquid Light liquid

The following photograph shows a set of antique hydrometers used to measure the density of beer. The middle hydrometer bears a label showing its calibration to be in degrees Baumé (heavy):



Liquid density measurement is useful in the alcoholic beverage industry to infer alcohol content. Since alcohol is less dense than water, a sample containing a greater concentration of alcohol (a greater *proof* rating) will be less dense than a "weaker" sample, all other factors being equal.

A less sophisticated version of hydrometer uses multiple balls of differing density. A common application for such a hydrometer is in measuring the concentration of "antifreeze" coolant for automobile engines. The denser the sample liquid, the more of the balls will float (and fewer will sink):



This form of instrument yields a qualitative assessment of liquid density as opposed to the quantitative measurement given by a hydrometer with calibrated marks on a single float. When used to measure the density of engine coolant, a greater number of floating balls represents a "stronger" concentration of glycol in the coolant. "Weak" glycol concentrations represent a greater percentage of water in the coolant, with a correspondingly greater freezing temperature.

2.9.7 Gas Laws

The *Ideal Gas Law* relates pressure, volume, molecular quantity, and temperature of an ideal gas together in one neat mathematical expression:

$$PV = nRT$$

Where,
P = Absolute pressure (atmospheres)
V = Volume (liters)
n = Gas quantity (moles)
R = Universal gas constant (0.0821 L $\cdot$ atm / mol $\cdot$ K)
T = Absolute temperature (K)

An alternative form of the Ideal Gas Law uses the number of actual gas molecules (N) instead of the number of moles of molecules (n):

$$PV = NkT$$

Where,
P = Absolute pressure (atmospheres)
V = Volume (liters)
N = Gas quantity (molecules)
k = Boltzmann's constant (1.38×10^{-23} J / K)
T = Absolute temperature (K)

Although no gas in real life is ideal, the Ideal Gas Law is a close approximation for conditions of modest gas density, and no phase changes (gas turning into liquid or visa-versa).

Since the molecular quantity of an enclosed gas is constant, and the universal gas constant *must* be constant, the Ideal Gas Law may be written as a proportionality instead of an equation:

$$PV \propto T$$

Several "gas laws" are derived from this Ideal Gas Law. They are as follows:

$PV = \text{Constant}$ **Boyle's Law** (assuming constant temperature T)

$V \propto T$ **Charles's Law** (assuming constant pressure P)

$P \propto T$ **Gay-Lussac's Law** (assuming constant volume V)

You will see these laws referenced in explanations where the specified quantity is constant (or very nearly constant).

For non-ideal conditions, the "Real" Gas Law formula incorporates a corrected term for the *compressibility* of the gas:

$$PV = ZnRT$$

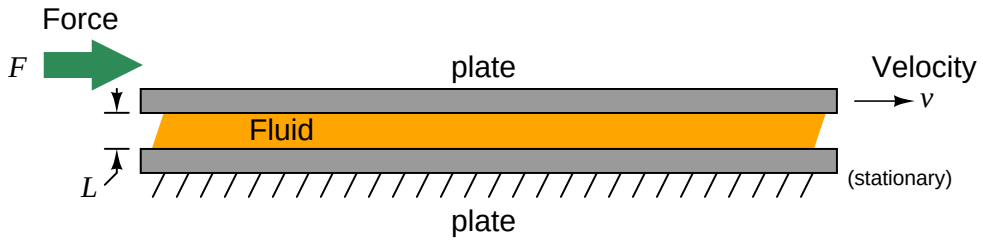
Where,
P = Absolute pressure (atmospheres)
V = Volume (liters)
Z = Gas compressibility factor (unitless)
n = Gas quantity (moles)
R = Universal gas constant (0.0821 L $\cdot$ atm / mol $\cdot$ K)
T = Absolute temperature (K)

The compressibility factor for an ideal gas is unity ($Z = 1$), making the Ideal Gas Law a limiting case of the Real Gas Law. Real gases have compressibility factors less than unity (< 1). What this means is real gases tend to compress more than the Ideal Gas Law would predict (i.e. occupies less volume for a given amount of pressure than predicted, and/or exerts less pressure for a given volume than predicted).

2.9.8 Fluid viscosity

Viscosity is a measure of a fluid's internal friction. The more "viscous" a fluid is, the "thicker" it is when stirred. Clean water is an example of a low-viscosity liquid, while honey at room temperature is an example of a high-viscosity liquid.

There are two different ways to quantify the viscosity of a fluid: *absolute viscosity* and *kinematic viscosity*. Absolute viscosity (symbolized by the Greek symbol "eta" η, or sometimes by the Greek symbol "mu" μ), also known as *dynamic viscosity*, is a direct relation between stress placed on a fluid and its rate of deformation (or shear). The textbook definition of absolute viscosity is based on a model of two flat plates moving past each other with a film of fluid separating them. The relationship between the shear stress applied to this fluid film (force divided by area) and the velocity/film thickness ratio is viscosity:



$$\eta = \frac{FL}{Av}$$

Where,

η = Absolute viscosity (pascal-seconds)
F = Force (newtons)
L = Film thickness (meters) – typically *much* less than 1 meter for any realistic demonstration!
A = Plate area (square meters)
v = Relative velocity (meters per second)

Another common unit of measurement for absolute viscosity is the *poise*, with 1 poise being equal to 0.1 pascal-seconds. Both units are too large for common use, and so absolute viscosity is often expressed in *centipoise*. Water has an absolute viscosity of very nearly 1.000 centipoise.

Kinematic viscosity (symbolized by the Greek letter "nu" ν) includes an assessment of the fluid's density in addition to all the above factors. It is calculated as the quotient of absolute viscosity and mass density:

$$\nu = \frac{\eta}{\rho}$$

Where,

ν = Kinematic viscosity (stokes)
η = Absolute viscosity (poises)
ρ = Mass density (grams per cubic centimeter)

As with the unit of poise, the unit of stokes is too large for convenient use, so kinematic viscosities are often expressed in units of *centistokes*. Water has an absolute viscosity of very nearly 1.000 centistokes.

The mechanism of viscosity in liquids is inter-molecular cohesion. Since this cohesive force is overcome with increasing temperature, most liquids tend to become "thinner" (less viscous) as they heat up. The mechanism of viscosity in gases, however, is inter-molecular collisions. Since these collisions increase in frequency and intensity with increasing temperature, gases tend to become "thicker" (more viscous) as they heat up.

As a ratio of stress to strain (applied force to yielding velocity), viscosity is often constant for a given fluid at a given temperature. Interesting exceptions exist, though. Fluids whose viscosities change with applied stress, and/or over time with all other factors constant, are referred to as *non-Newtonian fluids*. A simple example of a non-Newtonian fluid is cornstarch mixed with water, which "solidifies" under increasing stress then returns to a liquid state when the stress is removed.

2.9.9 Reynolds number

Viscous flow is when friction forces dominate the behavior of a moving fluid, typically in cases where viscosity (internal fluid friction) is great. *Inviscid flow*, by contrast, is where friction within a moving fluid is negligible. The *Reynolds number* of a fluid is a dimensionless quantity expressing the ratio between a moving fluid's momentum and its viscosity.

A couple of formulae for calculating Reynolds number of a flow are shown here:

$$\mathrm{Re} = \frac{D\overline{V}\rho}{\mu}$$

Where,

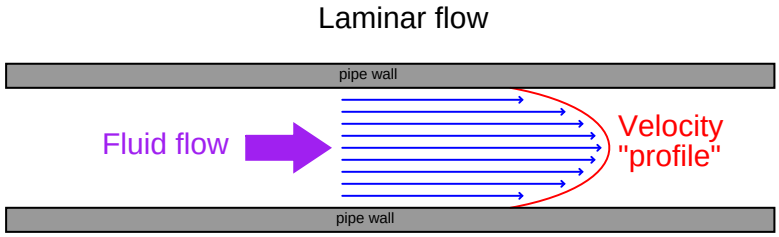
Re = Reynolds number (unitless)
D = Diameter of pipe, (meters)
$\overline{V}$ = Average velocity of fluid (meters per second)
ρ = Mass density of fluid (kilograms per cubic meter)
μ = Absolute viscosity of fluid (Pascal-seconds)

$$\mathrm{Re} = \frac{(3160)G_f Q}{D\mu}$$

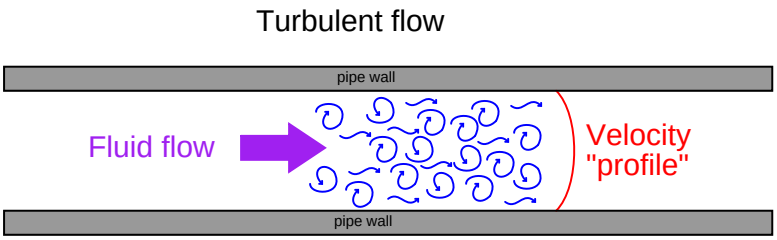
Where,

Re = Reynolds number (unitless)
G_f = Specific gravity of liquid (unitless)
Q = Flow rate (gallons per minute)
D = Diameter of pipe (inches)
μ = Absolute viscosity of fluid (centipoise)

The Reynolds number of a fluid stream may be used to qualitatively predict whether the flow regime will be *laminar* or *turbulent*. Low Reynolds number values predict laminar flow, where fluid molecules move in straight "stream-line" paths, and fluid velocity near the center of the pipe is substantially greater than near the pipe walls:



High Reynolds number values predict turbulent flow, where individual molecule motion is chaotic on a microscopic scale, and fluid velocities across the face of the flow profile are similar:



A generally accepted rule-of-thumb is that Reynolds number values less than 2,000 will probably be laminar, while values in excess of 10,000 will probably be turbulent. There is no definite threshold value for all fluids and piping configurations, though. To illustrate, I will share with you some examples of Reynolds number thresholds for laminar versus turbulent flows are given by various technical sources:

Chapter 2.8: Laminar Flowmeters of the *Instrument Engineer's Handbook, Process Measurement and Analysis, Third Edition* (pg. 105 – authors: R. Siev, J.B. Arant, B.G. Lipták) define Re $<$ 2,000 as "laminar" flow, Re $>$ 10,000 as "fully developed turbulent" flow, and any Reynolds number values between 2,000 and 10,000 as "transitional" flow.

Chapter 2: Fluid Properties – Part II of the *ISA Industrial Measurement Series – Flow* (pg. 11) define "laminar" flow as Re $<$ 2,000, "turbulent" flow as Re $>$ 4,000, and any Reynolds values in between 2,000 and 4,000 as "transitional" flow.

The Laminar Flow in a Pipe section in the *Standard Handbook of Engineering Calculations* (pg. 1-202) defines "laminar" flow as Re $<$ 2,100, and "turbulent" flow as Re $>$ 3,000. In a later section of that *same book* (Piping and Fluid Flow – page 3-384), "laminar" flow is defined as Re $<$ 1,200 and "turbulent" flow as Re $>$ 2,500.

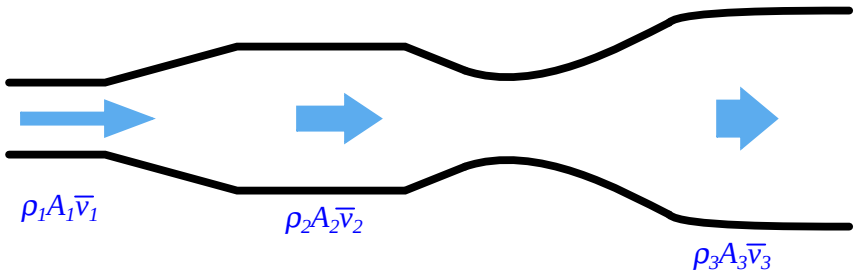
Douglas Giancoli, in his physics textbook *Physics* (third edition, pg. 11), defines "turbulent" flow as Re $<$ 2,000 and "turbulent" flow as Re $>$ 2,000.

Finally, a source on the internet (`http://flow.netfirms.com/reynolds/theory.htm`) attempts to define the threshold separating laminar from turbulent flow to an unprecedented degree of precision: Re $<$ 2,320 is supposedly the defining point of "laminar" flow, while Re $>$ 2,320 is supposedly marks the onset of "turbulent" flow.

Clearly, Reynolds number alone is insufficient for consistent prediction of laminar or turbulent flow, otherwise we would find far greater consistency in the reported Reynolds number values for each regime. Pipe roughness, swirl, and other factors influence flow regime, making Reynolds number an approximate indicator only. It should be noted that laminar flow may be sustained at Reynolds numbers significantly in excess of 10,000 under very special circumstances. For example, in certain coiled capillary tubes, laminar flow may be sustained all the way up to Re = 15,000, due to something known as the *Dean effect*!

2.9.10 Law of Continuity

Any fluid moving through a pipe obeys the Law of Continuity, which states that the product of average velocity ($\overline{v}$), pipe cross-sectional area (A), and fluid density (ρ) for a given flow stream must remain constant:



Fluid continuity is an expression of a more fundamental law of physics: the *Conservation of Mass*. If we assign appropriate units of measurement to the variables in the continuity equation, we see that the units cancel in such a way that only units of mass per unit time remain:

$$\rho A\overline{v} = \left[\frac{\text{kg}}{\text{m}^3}\right]\left[\frac{\text{m}^2}{1}\right]\left[\frac{\text{m}}{\text{s}}\right] = \left[\frac{\text{kg}}{\text{s}}\right]$$

This means we may define the product $\rho A\overline{v}$ as an expression of *mass flow rate*, or W:

$$W = \rho A\overline{v}$$

In order for the product $\rho A\overline{v}$ to differ between any two points in a pipe, mass would have to mysteriously appear and disappear. So long as the flow is continuous (not pulsing), and the pipe does not leak, it is impossible to have different rates of mass flow at different points along the flow path without violating the Law of Mass Conservation. The continuity principle for fluid through a pipe is analogous to the principle of current being the same everywhere in a series circuit, and for equivalently the same reason[39].

We refer to a flowing fluid as *incompressible* if its density does not substantially change[40]. For this limiting case, the continuity equation simplifies to the following form:

$$A_1\overline{v_1} = A_2\overline{v_2}$$

[39] In an electric circuit, the conservation law necessitating equal current at all points in a series circuit is the Law of Charge Conservation.

[40] Although not grammatically correct, this is a common use of the word in discussions of fluid dynamics. By definition, something that is "incompressible" *cannot* be compressed, but that is not how we are using the term here. We commonly use the term "incompressible" to refer to either a moving liquid (in which case the actual compressibility of the liquid is inconsequential) or a gas/vapor that does not *happen* to undergo substantial compression or expansion as it flows through a pipe. In other words, an "incompressible" flow is a moving fluid whose ρ does not substantially change, whether by actual impossibility or by circumstance.

Examining this equation in light of dimensional analysis, we see that the product $A\overline{v}$ is also an expression of flow rate:

$$A\overline{v} = \left[\frac{\text{m}^2}{1}\right]\left[\frac{\text{m}}{\text{s}}\right] = \left[\frac{\text{m}^3}{\text{s}}\right]$$

Cubic meters per second is an expression of *volumetric flow rate*, often symbolized by the variable Q:

$$Q = A\overline{v}$$

The practical implication of this principle is that fluid velocity is inversely proportional to the cross-sectional area of a pipe. That is, fluid slows down when the pipe's diameter expands, and visa-versa. We see this principle easily in nature: deep rivers run slow, while rapids are relatively shallow (and/or narrow).

For example, consider a pipe with an inside diameter of 8 inches (2/3 of a foot), passing a liquid flow of 5 cubic feet per minute. The average velocity (v) of this fluid may be calculated as follows:

$$Q = A\overline{v}$$

$$\overline{v} = \frac{Q}{A}$$

Solving for A in units of square feet:

$$A = \pi r^2$$

$$A = \pi\left(\frac{1}{3}\text{ ft}\right)^2 = \frac{\pi}{9}\text{ ft}^2$$

Now, solving for average velocity $\overline{v}$:

$$\overline{v} = \frac{\frac{5\text{ ft}^3}{\text{min}}}{\frac{\pi}{9}\text{ ft}^2}$$

$$\overline{v} = \left(\frac{5\text{ ft}^3}{\text{min}}\right)\left(\frac{9}{\pi\text{ ft}^2}\right)$$

$$\overline{v} = \frac{45}{\pi}\frac{\text{ft}}{\text{min}} = 14.32\frac{\text{ft}}{\text{min}}$$

2.9.11 Viscous flow

The pressure dropped by a slow-moving, viscous fluid through a pipe is described by the *Hagen-Poiseuille equation*. This equation applies only for conditions of low Reynolds number; i.e. when viscous forces are the dominant restraint to fluid motion through the pipe, and turbulence is nonexistent:

$$Q = k\left(\frac{\Delta P D^4}{\mu L}\right)$$

Where,

Q = Flow rate (gallons per minute)
k = Unit conversion factor = 7.86 $\times 10^5$
ΔP = Pressure drop (inches of water column)
D = Pipe diameter (inches)
μ = Liquid viscosity (centipoise) – this is a temperature-dependent variable!
L = Length of pipe section (inches)

2.9.12 Bernoulli's equation

Bernoulli's equation is an expression of the *Law of Energy Conservation* for an inviscid fluid stream, named after Daniel Bernoulli[41]. It states that the sum total energy at any point in a passive fluid stream (i.e. no pumps or other energy-imparting machines in the flow path) must be constant. Two versions of the equation are shown here:

$$z_1 \rho g + \frac{v_1^2 \rho}{2} + P_1 = z_2 \rho g + \frac{v_2^2 \rho}{2} + P_2$$

$$z_1 + \frac{v_1^2}{2g} + \frac{P_1}{\gamma} = z_2 + \frac{v_2^2}{2g} + \frac{P_2}{\gamma}$$

Where,

z = Height of fluid (from a common reference point, usually ground level)
ρ = Mass density of fluid
γ = Weight density of fluid ($\gamma = \rho g$)
g = Acceleration of gravity
v = Velocity of fluid
P = Pressure of fluid

Each of the three terms in Bernoulli's equation is an expression of a different kind of energy, commonly referred to as *head*:

$$z\rho g \qquad \text{Elevation head}$$

$$\frac{v^2 \rho}{2} \qquad \text{Velocity head}$$

$$P \qquad \text{Pressure head}$$

Elevation and Pressure heads are potential forms of energy, while Velocity head is a kinetic form of energy. Note how the elevation and velocity head terms so closely resemble the formulae for potential and kinetic energy of solid objects:

$$E_p = mgh \qquad \text{Potential energy formula}$$

$$E_k = \frac{1}{2} m v^2 \qquad \text{Kinetic energy formula}$$

The only real differences between the solid-object and fluid formulae for energies is the use of mass *density* (ρ) for fluids instead of mass (m) for solids, and the arbitrary use of the variable z for height instead of h. In essence, the elevation and velocity head terms within Bernoulli's equation come from the assumption of individual fluid molecules behaving as miniscule solid masses.

[41] According to Ven Te Chow in *Open Channel Hydraulics*, who quotes from Hunter Rouse and Simon Ince's work *History of Hydraulics*, Bernoulli's equation was first formulated by the great mathematician Leonhard Euler and made popular by Julius Weisbach, not by Daniel Bernoulli himself.

It is very important to maintain consistent units of measurement when using Bernoulli's equation! Each of the three energy terms (elevation, velocity, and pressure) *must* possess the exact same units if they are to add appropriately[42]. Here is an example of dimensional analysis applied to the first version of Bernoulli's equation (using British units):

$$z\rho g + \frac{v^2\rho}{2} + P$$

$$[\text{ft}]\left[\frac{\text{slug}}{\text{ft}^3}\right]\left[\frac{\text{ft}}{\text{s}^2}\right] + \left[\frac{\text{ft}}{\text{s}}\right]^2\left[\frac{\text{slug}}{\text{ft}^3}\right] + \left[\frac{\text{lb}}{\text{ft}^2}\right] = \left[\frac{\text{slug}}{\text{ft}\cdot\text{s}^2}\right]$$

As you can see, both the first and second terms of the equation (elevation and velocity heads) bear the same unit of slugs per foot-second squared after all the "feet" are canceled. The third term (pressure head) does not appear as though its units agree with the other two terms, until you realize that the unit definition of a "pound" is a slug of mass multiplied by the acceleration of gravity in feet per second squared, following Newton's Second Law of motion ($F = ma$):

$$[\text{lb}] = [\text{slug}]\left[\frac{\text{ft}}{\text{s}^2}\right]$$

Once we make this substitution into the pressure head term, the units are revealed to be the same as the other two terms, slugs per foot-second squared:

$$\left[\frac{\text{lb}}{\text{ft}^2}\right] = \left[\frac{\text{slug}\left[\frac{\text{ft}}{\text{s}^2}\right]}{\text{ft}^2}\right] = \left[\frac{\text{slug}}{\text{ft}\cdot\text{s}^2}\right]$$

In order for our British units to be consistent here, we must use *feet* for elevation, *slugs* per cubic *foot* for mass density, *feet* per *second* squared for acceleration, *feet* per *second* for velocity, and *pounds* per square *foot* for pressure. If one wished to use the more common pressure unit of PSI (pounds per square inch) with Bernoulli's equation instead of PSF (pounds per square foot), all the other units would have to change accordingly: elevation in *inches*, mass density in slugs per cubic *inch*, acceleration in *inches* per second squared, and velocity in *inches* per second.

Just for fun, we can try dimensional analysis on the second version of Bernoulli's equation, this time using metric units:

$$z + \frac{v^2}{2g} + \frac{P}{\gamma}$$

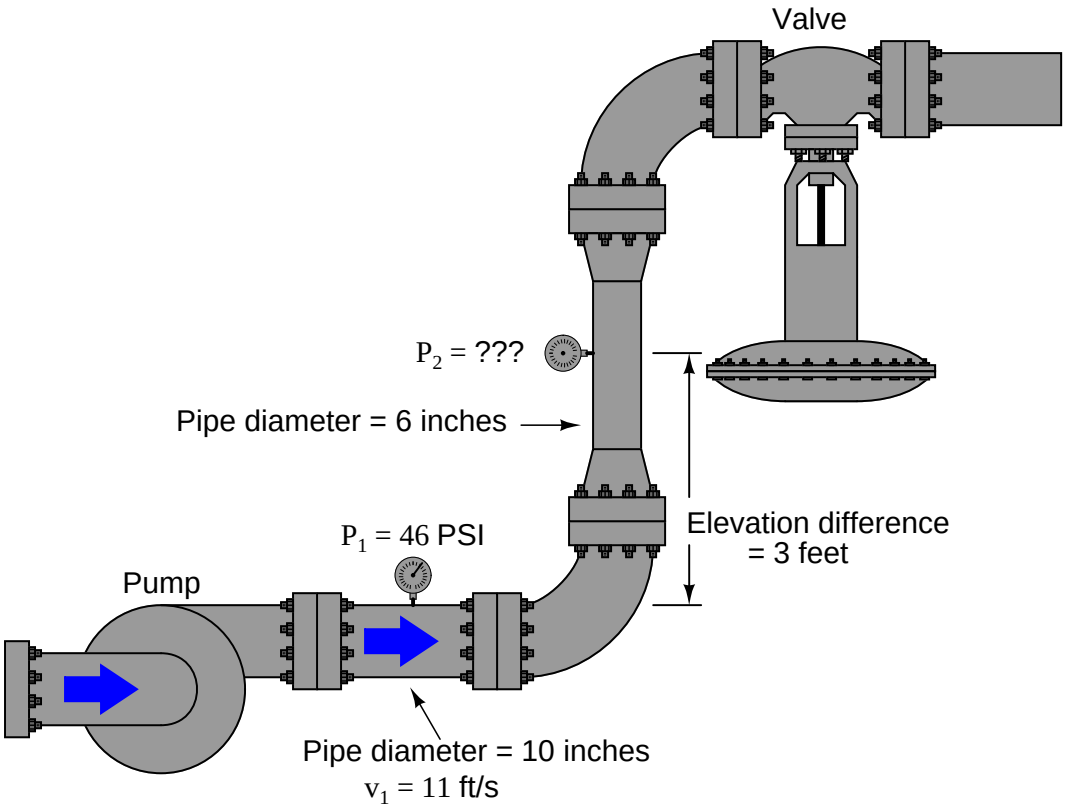
$$[\text{m}] + \left[\frac{\left[\frac{\text{m}}{\text{s}}\right]^2}{\left[\frac{\text{m}}{\text{s}^2}\right]}\right] + \left[\frac{\left[\frac{\text{N}}{\text{m}^2}\right]}{\left[\frac{\text{N}}{\text{m}^3}\right]}\right] = [\text{m}]$$

Here, we see that all three terms end up being cast in simple units of meters. That is, the fluid's elevation, velocity, and pressure heads are all expressed as simple elevations. In order for our metric units to be consistent here, we must use *meters* for elevation, *meters* per *second* for velocity, *meters*

[42] Surely you've heard the expression, "Apples and Oranges don't add up." Well, pounds per square inch and pounds per square foot don't add up either!

per *second* squared for acceleration, *pascals* (*newtons* per square *meter*) for pressure, and *newtons* per cubic *meter* for weight density.

The following example shows how we would apply Bernoulli's equation to the solution of pressure at a point in a water piping system, assuming no frictional losses at any point:



Water has a nominal density of 62.4 pounds per cubic foot, but this is *weight* density (γ) and not *mass* density (ρ). If we wish to use the form of Bernoulli's equation where all terms are in units of pressure ($z\rho g + \frac{v^2\rho}{2} + P$), we must have a value of ρ for water.

The relationship between weight density γ and mass density ρ is the exact same relationship between weight (F_W) and mass (m) in a gravitational field (g). Newton's Second Law equation relating force to mass and acceleration ($F = ma$) works well to relate weight to mass and gravitational acceleration:

$$F = ma$$

$$F_W = mg$$

Dividing both sides of this equation by volumetric units (V) (e.g. cubic feet) gives us our relationship between γ and ρ:

$$\frac{F_W}{V} = \frac{m}{V}g$$

$$\gamma = \rho g$$

Water has a weight density of 62.4 pounds per cubic foot in Earth gravity (32.2 feet per second squared), so:

$$\rho = \frac{\gamma}{g}$$

$$\rho = \frac{62.4 \text{ lb/ft}^3}{32.2 \text{ ft/s}^2} = 1.94 \text{ slugs/ft}^3$$

Now we are ready to begin our Bernoulli's equation calculations. Since we have the freedom to choose any arbitrary point in the piping system as our reference elevation ($z = 0$), we will set the location of the first pressure gauge as this reference height, so the second pressure gauge will have a positive elevation value of 3 feet. First, calculating the values of all terms (elevation, velocity, and pressure) at the first point, near the discharge of the pump:

Head	Calculation	Value
$z_1 \rho g$	(0 ft) (1.94 slugs/ft^3) (32.2 ft/s^2)	0 lb/ft^3
$v_1^2 \rho / 2$	(11 ft/s)^2 (1.94 slugs/ft^3) / 2	117.4 lb/ft^3
P	(46 lb/in^2) (144 in^2/1 ft^2)	6624 lb/ft^2
Total	0 lb/ft^2 + 6.56 lb/ft^2 + 6624 lb/ft^2	**6741.4 lb/ft^2**

Note the absolutely consistent use of units: all units of distance are *feet*. All units of time are *seconds*. Failure to maintain consistency of units will result in (often severely) incorrect results![43]

[43] It is entirely possible to perform all our calculations using inches and/or minutes as the primary units instead of feet and seconds. The only caveat is that *all* units throughout all terms of Bernoulli's equation must be consistent. This means we would also have to express mass density in units of slugs per cubic *inch*, the acceleration of gravity in *inches* per second squared (or *inches* per *minute* squared), and velocity in units of *inches* per second (or *inches* per *minute*). The only real benefit of doing this is that pressure would remain in the more customary units of pounds per square *inch*. My personal preference is to do all calculations using units of feet and seconds, then convert pressures in units of PSF to units of PSI at the very end.

Next, we will calculate the values of the elevation and velocity heads at the location of the second pressure gauge. Here, the pressure is unknown, but the elevation is given as 3 feet higher than the first gauge, and the velocity may be calculated by pipe size. We know that the pipe here is 6 inches in diameter, while it is 10 inches in diameter where the velocity is 11 feet per second. Since area varies with the *square* of the diameter, and velocity varies inversely with area, we can tell that the velocity at the second pressure gauge will be 2.78 times greater than at the first pressure gauge:

$$v_2 = v_1 \left(\frac{D_1}{D_2} \right)^2$$

$$v_2 = 11 \text{ ft/s} \left(\frac{10 \text{ in}}{6 \text{ in}} \right)^2$$

$$v_2 = (11 \text{ ft/s})(2.78) = 30.56 \text{ ft/s}$$

Tabulating our calculations and results:

Head	Calculation	Value
$z_2 \rho g$	(3 ft) (1.94 slugs/ft^3) (32.2 ft/s^2)	187.4 lb/ft^3
$v_2^2 \rho / 2$	(30.56 ft/s)^2 (1.94 slugs/ft^3) / 2	905.6 lb/ft^3
Total	187.4 lb/ft^2 + 905.6 lb/ft^2 + P_2	**1093 lb/ft^2 + P_2**

Knowing that the total head calculated at the first location was 6741.4 lb/ft^2, and the Conservation of Energy requires total heads at both locations be equal (assuming no energy lost to fluid friction along the way), P_2 must be equal to:

$$6741.4 \text{ lb/ft}^2 = 1093 \text{ lb/ft}^2 + P_2$$

$$P_2 = 6741.4 \text{ lb/ft}^2 - 1093 \text{ lb/ft}^2 = 5648.3 \text{ lb/ft}^2$$

Converting pounds per square foot into the more customary unit of pounds per square inch (PSI):

$$P_2 = (5648.3 \text{ lb/ft}^2) \left(\frac{1 \text{ ft}^2}{144 \text{ in}^2} \right)$$

$$P_2 = 39.2 \text{ lb/in}^2$$

Note how much lower the pressure is at the second gauge than at the first: 39.2 PSI compared to 46 PSI: almost a 7 PSI decrease in pressure. Note also how little vertical distance separates the two gauges: only 3 feet. Clearly, the change in elevation between those two points in insufficient to account for the large loss in pressure[44]. Given a 3 foot difference in elevation, one would expect a pressure reduction of about 1.3 PSI for a static column of water, but what we're seeing in this piping system is a pressure drop of nearly 7 PSI. The difference is due to an exchange of energy

[44] A simple approximation for pressure loss due to elevation gain is approximately 1 PSI for every 2 vertical feet of water (1 PSI for every 27.68 inches to be more exact).

from potential to kinetic form, as the fluid enters a much narrower pipe (6 inches instead of 10) and must increase velocity.

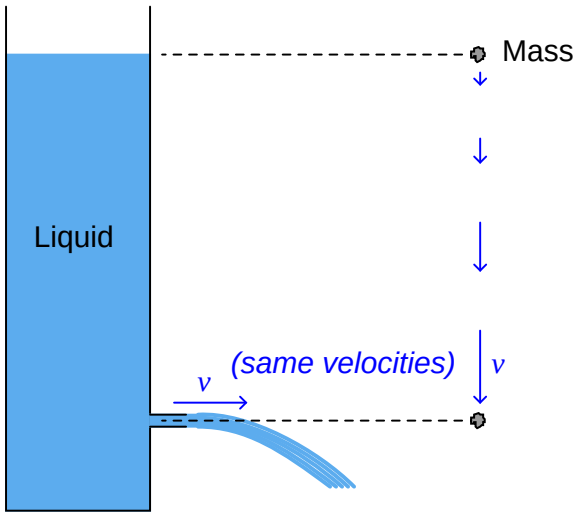
Furthermore, if we were to increase the flow rate discharged from the pump, resulting in even more velocity through the narrow pipe, pressure at P_2 might even drop lower than atmospheric. In other words, Bernoulli's equation tells us we can actually produce a *vacuum* by accelerating a fluid through a constriction. This principle is widely used in industry with devices known as *eductors*: tapered tubes through which fluid flows at high velocity to produce a vacuum at the throat.

Some eductors use high-velocity steam as the working fluid to produce significant vacuums. Other eductors use process liquid flow, such as the eductor shown in this next photograph where wastewater flow creates a vacuum to draw gaseous chlorine into the stream for biological disinfection:

Here, the eductor helps fulfill an important safety function. By creating a vacuum to draw dangerous chlorine gas from the supply tank into the water stream, the chlorine gas piping may be continuously maintained at a slightly negative pressure throughout. If ever a leak were to develop in the chlorine system, this vacuum would cause ambient air to enter the chlorine pipe rather than toxic chlorine gas to exit the pipe, making a leak far less dangerous than if the chlorine gas piping were maintained in a pressurized state.

2.9.13 Torricelli's equation

The velocity of a liquid stream exiting from a nozzle, pressured solely by a vertical column of that same liquid, is equal to the free-fall velocity of a solid mass dropped from the same height as the top of the liquid column. In both cases, potential energy (in the form of vertical height) converts to kinetic energy (motion):



This was discovered by Evangelista Torricelli almost 100 years prior to Bernoulli's more comprehensive formulation. The velocity may be determined by solving for v after setting the potential and kinetic energy formulae equal to each other (since all potential energy at the upper height must translate into kinetic energy at the bottom, assuming no frictional losses):

$$mgh = \frac{1}{2}mv^2$$

$$gh = \frac{1}{2}v^2$$

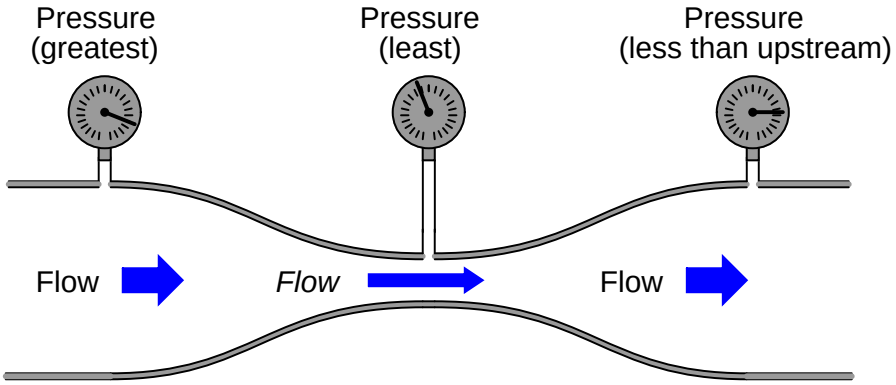
$$2gh = v^2$$

$$v = \sqrt{2gh}$$

Note how mass (m) simply disappears from the equation, neatly canceling on both sides. This means the nozzle velocity depends only on height, not the mass density of the liquid. It also means the velocity of the falling object depends only on height, not the mass of the object.

2.9.14 Flow through a venturi tube

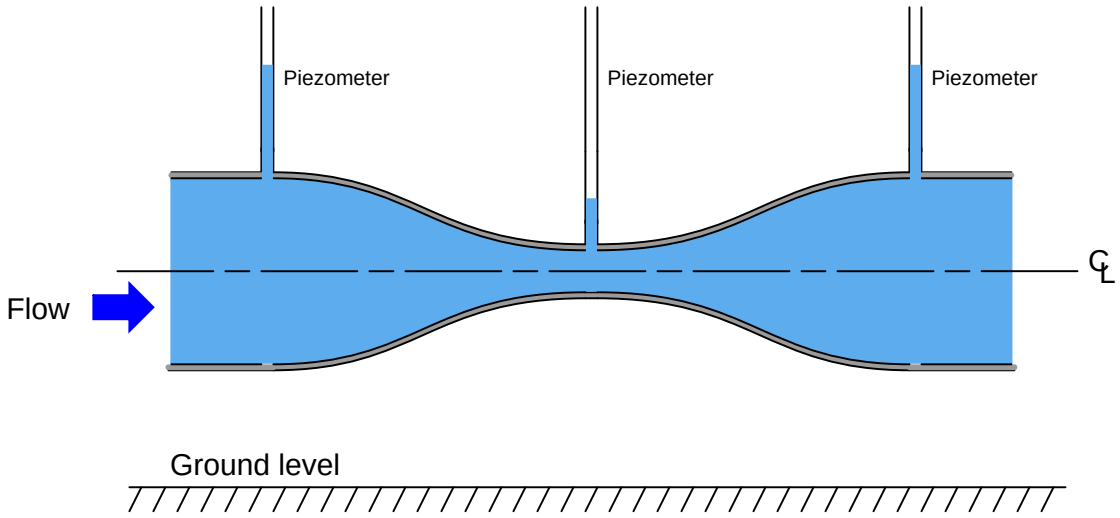
If an incompressible fluid moves through a *venturi tube* (a tube purposefully built to be narrow in the middle), the continuity principle tells us the fluid velocity must increase through the narrow portion. This increase in velocity causes kinetic energy to increase at that point. If the tube is level, there will be negligible difference in elevation (z) between different points of the tube's centerline, which means elevation head remains constant. According to the Law of Energy Conservation, some other form of energy must decrease to account for the increase in kinetic energy. This other form is the pressure head, which decreases at the throat of the venturi:



Ideally, the pressure downstream of the narrow throat should be the same as the pressure upstream, assuming equal pipe diameters upstream and down. However, in practice the downstream pressure gauge will show slightly less pressure than the upstream gauge due to some inevitable energy loss as the fluid passed through the venturi. Some of this loss is due to fluid friction against the walls of the tube, and some is due to viscous losses within the fluid driven by turbulent fluid motion at the high-velocity throat passage.

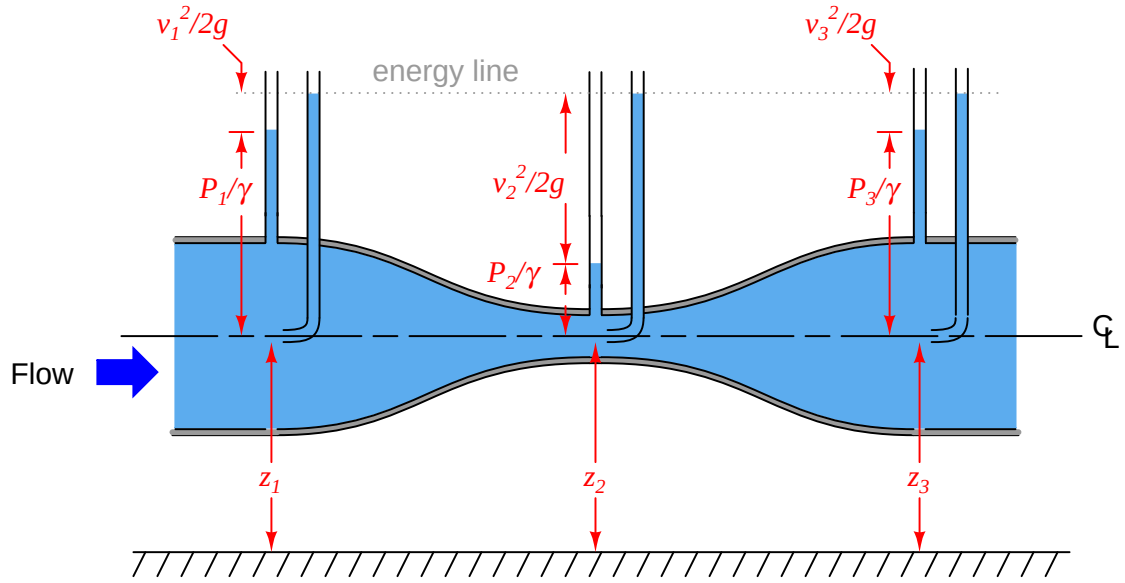
The difference between upstream and downstream pressure is called *permanent pressure loss*, while the difference in pressure between the narrow throat and downstream is called *pressure recovery*.

If we install vertical sight-tubes called *piezometers* along a horizontal venturi tube, the differences in pressure will be shown by the heights of liquid columns within the tubes. Here, we assume an ideal (inviscid) liquid with no permanent pressure loss:



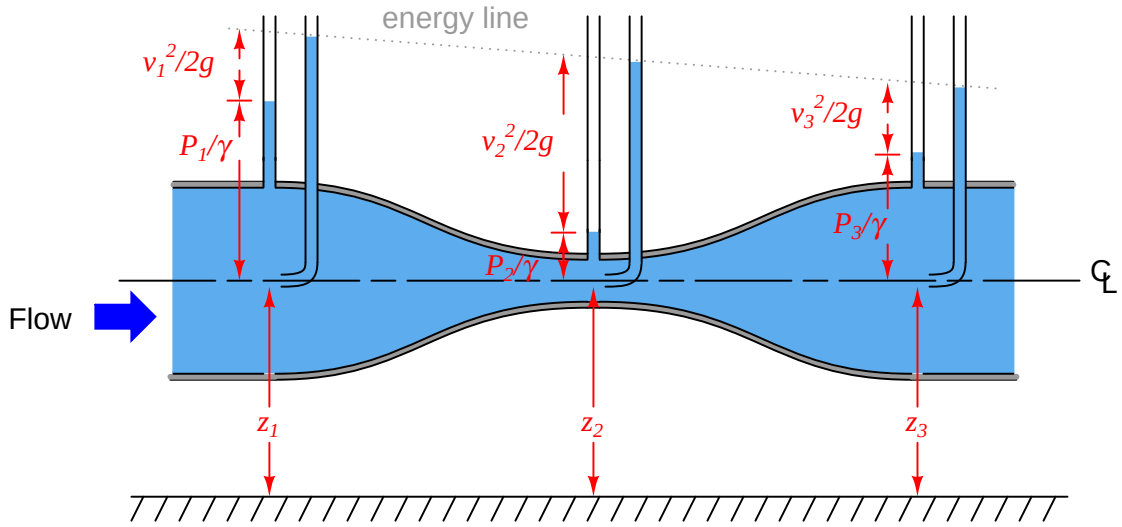
If we add three more piezometers to the venturi tube assembly, each one equipped with its own *Pitot tube* facing upstream to "catch" the velocity of the fluid, we see that total energy is indeed conserved at every point in the system. Here, each of the "heads" represented[45] in Bernoulli's equation are shown in relation to the different piezometer heights:

$$z + \frac{v^2}{2g} + \frac{P}{\gamma} = (\text{constant})$$



[45] The form of Bernoulli's equation with each term expressed in units of distance (e.g. z = [feet] ; $\frac{v^2}{2g}$ = [feet] ; $\frac{P}{\gamma}$ = [feet]) was chosen so that the piezometers' liquid heights would directly correspond.

A more realistic scenario would show the influence of energy lost in the system due to friction. Here, the total energy is seen to decrease as a result of friction:



References

Chow, Ven Te., *Open-Channel Hydraulics*, McGraw-Hill Book Company, Inc., New York, NY, 1959.

Considine, Douglas C., *Energy Technology Handbook*, McGraw-Hill Book Company, New York, NY, 1977.

Control Valve Handbook, Third Edition, Fisher Controls International, Inc., Marshalltown, IA, 1999.

Giancoli, Douglas C., *Physics for Scientists & Engineers*, Third Edition, Prentice Hall, Upper Saddle River, NJ, 2000.

Hicks, Tyler G., *Standard Handbook of Engineering Calculations*, McGraw-Hill, Inc., New York, NY, 1972.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Measurement and Analysis Volume I*, Fourth Edition, CRC Press, New York, NY, 2003.

Miller, Richard W., *Flow Measurement Engineering Handbook*, Second Edition, McGraw-Hill Publishing Company, New York, NY, 1989.

Pauling, Linus, *General Chemistry*, Dover Publications, Inc., Mineola, NY, 1988.

Rouse, Hunter, *Characteristics of Laminar and Turbulent Flow* (video), Iowa Institute of Hydraulic Research, University of Iowa.

Shapiro, Ascher H., *Pressure Fields and Fluid Acceleration* (video), Massachusetts Institute of Technology, Educational Services Incorporated, 1962.

Thompson, Ambler and Taylor, Barry N., *Guide for the Use of the International System of Units (SI)*, special publication 811 (second printing), National Institute of Standards and Technology, Gaithersburg, MD, 2008.

Vennard, John K., *Elementary Fluid Mechanics*, 3rd Edition, John Wiley & Sons, Inc., New York, NY, 1954.

Weast, Robert C.; Astel, Melvin J.; and Beyer, William H., *CRC Handbook of Chemistry and Physics*, 64th Edition, CRC Press, Inc., Boca Raton, FL, 1984.

Chapter 3

Chemistry

Chemistry is the study of how atoms join with and separate from one another. Like so many other areas of physical science, the patterns and laws we see in chemical reactions are dominated by two fundamental laws of physics: the *Conservation of Mass* and the *Conservation of Energy*. The particles of matter comprising atoms have the ability to store energy in potential form, and their tendency is to "seek" the most stable energy state[1]. The arrangement of electrons around the nucleus of an atom is largely dictated by the tendency of electrons to "prefer" stable energy states, and so is the formation of molecules (atoms bonded together): electrons seeking energy states least liable to disturbance. The rest, as they say, is all detail.

We exploit this property of energy storage in the fuels we use. Atoms bound together to form molecules are in a lower energy state than when they exist as separate atoms. Therefore, an investment of energy is required to force molecules apart (into separate atoms), and energy is returned (released) when atoms join together to form molecules. The combustion of a *fuel*, for example, is nothing more than a process of the atoms in relatively unstable (high-energy) fuel molecules joining with oxygen atoms in air to form stable (low-energy) molecules such as water (H_2O) and carbon dioxide (CO_2).

Natural gas, for example, is a relatively stable combination of hydrogen (H) and carbon (C) atoms, mostly in the form of molecules with a 4:1 hydrogen-to-carbon ratio (CH_4). However, when placed in the vicinity of free oxygen (O) atoms, and given enough energy (a spark) to cause the hydrogen and carbon atoms to separate from each other, the hydrogen atoms strongly bond with oxygen atoms to form water molecules (H_2O), while the carbon atoms also strongly bond with oxygen atoms to form carbon dioxide molecules (CO_2). These strong bonds formed between hydrogen, carbon, and oxygen in the water and carbon dioxide molecules are the result of electrons within those atoms seeking lower energy states than they possessed while forming molecules of natural gas (CH_4). In other words, the energy states of the electrons in the hydrogen and carbon atoms were higher when they were joined to form natural gas than they are when joined with oxygen to form water and carbon dioxide. As those electrons attain lower energy states, they difference of energy

[1] This generally means to seek the *lowest* available energy state, but there are important exceptions where chemical reactions actually proceed in the opposite direction (with atoms seeking *higher* energy states and absorbing energy from the surrounding environment to achieve those higher states). A more general and consistent understanding of matter and energy interactions involves a more complex concept called *entropy*.

must go somewhere (since energy cannot be created or destroyed), and so the chemical reaction releases that energy in the forms of heat and light. This is what you see and feel in the presence of a natural gas flame: the heat and light emitted by hydrogen and carbon atoms joining with oxygen atoms.

The Law of Mass Conservation plays an important role in chemistry as well. When atoms join to form molecules, their masses add. That is, the mass of a molecule is precisely equal[2] to the mass of its constituent atoms. When chemical engineers design manufacturing processes, they must pay close attention to a principle called *mass balance*, where the sum total of all masses for chemical feeds into a process precisely equals the sum total of all masses exiting the process.

Too many other practical applications of chemistry exist to summarize in these pages, but this chapter aims to give you an foundation to understand certain chemistry concepts necessary to comprehend the function of certain instruments (notably *analyzers*) and processes.

[2]This statement is not perfectly honest. When atoms join to form molecules, the subsequent release of energy is translated into an incredibly small loss of mass for the molecule, as described by Albert Einstein's famous mass-energy equation $E = mc^2$. However, this mass discrepancy is so small (typically less than one part per *billion* of the original mass!), we may safely ignore it for the purposes of understanding chemical reactions in industrial processes.

3.1 Terms and Definitions

- *Atom*: the smallest unit of matter that may be isolated by chemical means.
- *Particle*: a part of an atom, separable from the other portions only by levels of energy far in excess of chemical reactions.
- *Proton*: a type of "elementary" particle, found in the nucleus of an atom, possessing a positive electrical charge.
- *Neutron*: a type of "elementary" particle, found in the nucleus of an atom, possessing no electrical charge, and having nearly the same amount of mass as a proton.
- *Electron*: a type of "elementary" particle, found in regions surrounding the nucleus of an atom, possessing a negative electrical charge, and having just a small fraction of the mass of a proton or neutron.
- *Element*: a substance composed of atoms all sharing the same number of protons in their nuclei (e.g. hydrogen, helium, nitrogen, iron, cesium, fluorine).
- *Atomic number*: the number of protons in the nucleus of an atom – this quantity defines the chemical identify of an atom.
- *Atomic mass* or *Atomic weight*: the total number of elementary particles in the nucleus of an atom (protons + neutrons) – this quantity defines the vast majority of an atom's mass, since the only other elementary particle (electrons) are so light-weight by comparison to protons and neutrons.
- *Ion*: an atom or molecule that is not electrically balanced (equal numbers of protons and electrons).

  → *Cation*: a positively-charged ion, called a "cation" because it is attracted toward the negative electrode (cathode) immersed in a solution.

  → *Anion*: a negatively-charged ion, called an "anion" because it is attracted toward the positive electrode (anode) immersed in a solution.
- *Isotope*: a variation on the theme of an element – atoms sharing the same number of protons in their nuclei, but having different numbers of neutrons, are called "isotopes" (e.g. uranium-235 versus uranium-238).
- *Molecule*: the smallest unit of matter composed of two or more atoms joined by electron interaction in a fixed ratio (e.g. water: H_2O). The smallest unit of a *compound*.
- *Compound*: a substance composed of identical molecules (e.g. pure water).
- *Isomer*: a variation on the theme of a compound – molecules sharing the same numbers and types of atoms, but having different structural forms, are called "isomers". For example, the sugars glucose and fructose are isomers, both having the same formula $C_6H_{12}O_6$ but having disparate structures.
- *Mixture*: a substance composed of different atoms or molecules.

- *Solution*: an homogeneous mixture at the molecular level (different atoms/molecules thoroughly mixed together). A solution may be a gas, a liquid, or a solid (e.g. air, saltwater, doped silicon).

  → *Solvent*: the majority element or compound in a solution. Chemists usually consider water to be the *universal solvent*.

  → *Solute*: the minority element or compound in a solution (may be more than one).

  → *Precipitate*: (noun) solute that has "fallen out of solution" due to the solution being saturated with that element or compound; (verb) the process of solute separating from the rest of the solution. (e.g. Mixing too much salt with water results in some of that salt *precipitating* out of the water to form a solid pile at the bottom.)

  → *Supernatant*: the solution remaining above the precipitate.

- *Suspension*: an heterogeneous mixture where separation occurs due to gravity (e.g. mud).

- *Colloid* or *Colloidal suspension*: an heterogeneous mixture where separation does not occur (or occurs at a negligible pace) due to gravity (e.g. milk).

  → *Aerosol*: A colloid formed of a solid or liquid substance dispersed in a gas medium.

  → *Foam*: A colloid formed of a gas dispersed in either a liquid or a solid medium.

  → *Emulsion*: A colloid formed of a liquid dispersed in either a liquid or a solid medium.

  → *Sol*: A colloid formed of a solid dispersed in either a liquid or a solid medium.

3.2 Atomic theory and chemical symbols

The three "elementary" particles of matter comprising all atoms are *electrons*, *protons*, and *neutrons*. Combinations of these three particle types in various whole-number quantities constitute every type of atom. These fundamental particles are absolutely miniscule in comparison to the macroscopic existence of human beings. Just to illustrate, the mass of a single proton is approximately 1.67×10^{-27} kilograms: written without scientific notation, it would be 0.00000000000000000000000000167 kg. An electron is even smaller: weighing in at 9.11×10^{-31} kg (about *1800 times* less mass than a proton!). Being far smaller in size than a wavelength of visible light[3], we cannot see these particles even with the most powerful optical microscope.

Protons and neutrons are very tightly bound together in the nucleus (center) of an atom. The bind is so tight that only extraordinary forces are able to pry an atom's nucleus apart. Suffice it to say, one cannot disturb the stability of an atomic nucleus by rubbing, cutting, grinding, heating, smashing, or any other macroscopic physical process. The force binding protons and neutrons together in the nucleus is known as the *strong nuclear force*.

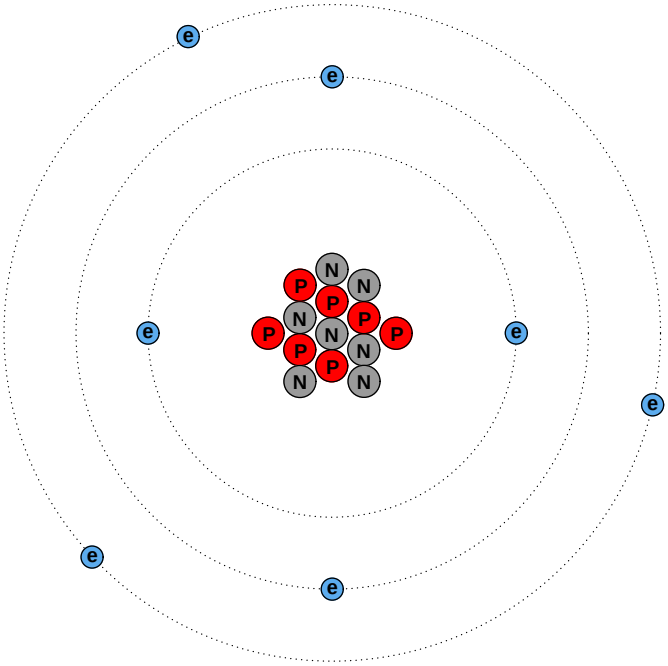
Electrons "orbit" the nucleus of atoms, and are held in proximity to those nuclei by electrostatic attraction (the so-called *electromagnetic force*), which is many orders of magnitude weaker than the strong nuclear force. Thus, electrons *can* be dislodged from or added to atoms through the agency of macroscopic forces such as rubbing, cutting, grinding, heating, etc. It is the changeable configurations of electrons that accounts for different atoms joining together to form *molecules*.

The chemical identify of any atom is a simple and direct function of how many protons that atom has in its nucleus. Each nitrogen atom, for example, has seven (7) protons in its nucleus. This quantity is called the *atomic number* of an atom. In order for an atom to have a net neutral electric charge, there must be as many electrons orbiting the nucleus as there are protons in the nucleus, since protons carry equal and opposite electric charges. Therefore, a neutral atom of nitrogen will have seven electrons orbiting around the nucleus, electrically balancing the seven protons within the nucleus.

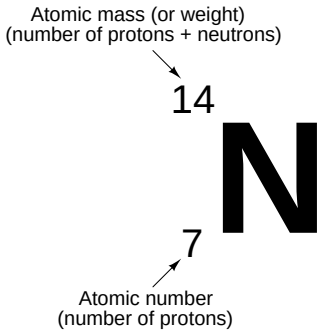
The number of neutrons within the nucleus of an atom does not affect the atom's chemical identity, but it may affect its nuclear properties (e.g. whether or not it is radioactive; to what degree it captures certain forms of particulate radiation, etc.). For example, most nitrogen atoms have seven neutrons along with seven protons in their nuclei, giving a total nuclear particle count of fourteen – the atomic *mass* of the atom, sometimes called the atomic *weight*. However, it is possible for a nitrogen atom to have eight neutrons (an atomic mass of fifteen) and still be "nitrogen," with all the same chemical properties.

[3] In order for a wave of light to be influenced at all by an object, that object must be at least the size of the wave's length. To use an analogy with water waves, it would be comparing the interaction of a water wave on a beach against a large rock (a disturbance in the wave pattern) versus the non-disturbance of that same wave as it encounters a small buoy.

A tremendously simplified model of a common nitrogen atom is shown here, with 7 protons and 7 neutrons in the nucleus, and 7 electrons in "orbit" around the nucleus:

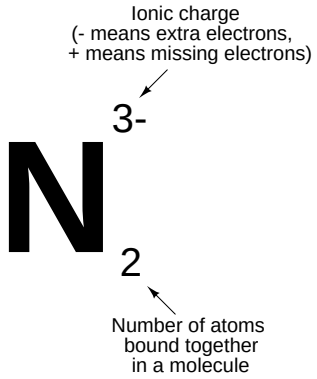


The atomic number of this atom (the number of protons in the nucleus) is seven, which is what makes it nitrogen. The *atomic mass* of this atom (the sum of protons and neutrons in the nucleus) is fourteen. The chemical symbol for this atom is shown here:

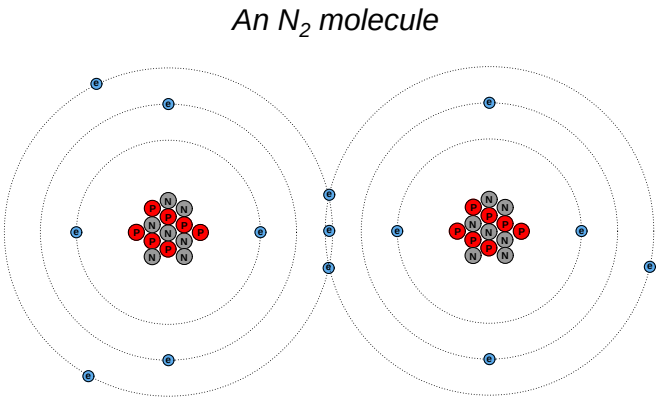


The atomic number is redundant to the letter "N" for nitrogen, since only the element nitrogen can have an atomic number of seven. The atomic mass is only relevant when we need to distinguish one *isotope* of nitrogen from another (variations of elements having the same number of protons but different numbers of neutrons), and this is seldom because the chemical properties of isotopes are identical – only their masses differ. For these reasons, you will usually find no left-hand subscripts or superscripts placed near chemical symbols of elements in chemical expressions.

By contrast, subscripts and superscripts placed to the right of a chemical symbol have very important meanings in chemistry. A right-hand subscript refers to the number of atoms bound together to form a molecule. A right-hand superscript refers to the electrical charge possessed by an atom (or by a molecule) by virtue of the number of electrons not matching the number of protons:

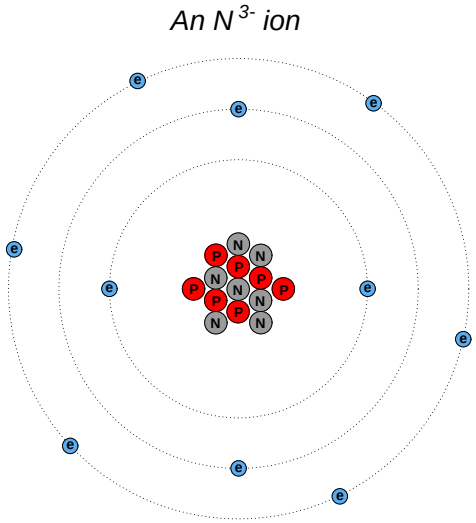


An N_2 molecule may be represented simplistically as follows, the two nitrogen atoms joined by a mutual sharing of the highest-energy (valence) electrons, shown in this illustration as those electrons residing in the largest-diameter "orbits:"



(**Two** atoms of Nitrogen bound together by the sharing of electrons)

An N^{3-} ion is an atom of nitrogen having three more electrons than it normally would when electrically balanced:



(Possesses **three** more electrons than an electrically balanced atom would)

A chemical *formula* is a written description of a molecule's composition. Ethanol (ethyl alcohol), for example, is a conglomerate of two carbon atoms, six hydrogen atoms, and one oxygen atom. One way to express this structure is to write the following formula for ethanol, the right-hand subscripts showing the relative quantities of atoms in each ethanol molecule:

$$C_2H_6O$$

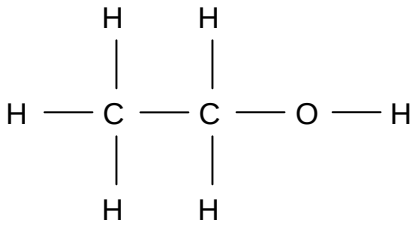
This is called a *molecular chemical formula*, because it shows the proportions of atom types comprising each molecule.

A more common way to write the formula for ethanol, though, is this:

$$C_2H_5OH$$

Here, an attempt is made to show the physical structure of the ethanol molecule, where one of the hydrogen atoms is found distant from the others. This is called a *structural* formula. If more detail is needed, a semi-graphic representation called a *displayed formula* may be used in lieu of a structural formula in lieu of a structural formula:

Displayed formula for ethanol
(C_2H_5OH)

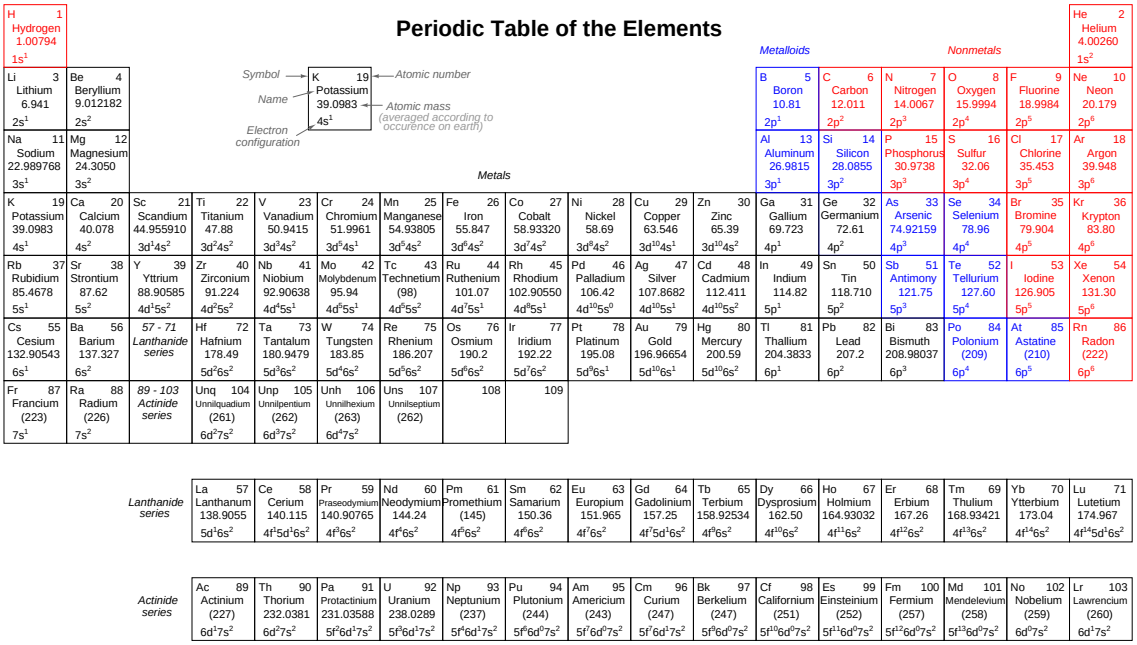


Chemical engineers often deal with processes where *mixtures* of similar compounds exist. Wastewater treatment is but one example, where an array of organic compounds must all be treated through oxidation (chemical reaction with oxygen). In such cases, it is common to write formulae expressing the *average* ratios of elements. Primary sludge clarified from municipal wastewater, for example, may be represented by the *compositional formula* $C_{22}H_{39}O_{10}N$. This does not suggest the existence of some monstrous molecule consisting of twenty-two carbon atoms, thirty-nine hydrogen atoms, ten oxygen atoms, and a lone nitrogen atom somewhere in a sample of sludge, but rather that the combined *average* carbon, hydrogen, oxygen, and nitrogen quantities in that sludge exist in a variety of molecular forms in these approximate proportions. This aggregate formula expression helps the engineer calculate the gross chemical characteristics of the sludge, such as the amount of oxygen needed to completely oxidize the sludge.

Sometimes, compositional formulae are written with non-integer subscripts. An example of this would be the compositional formula $C_{4.8}H_{8.4}O_{2.2}$, which also happens to be an average composition for wastewater sludge (ignoring nitrogen content). The same formula could just as well have been written $C_{48}H_{84}O_{22}$, or even $C_{24}H_{42}O_{11}$, because these subscript values all express the exact same proportions.

3.3 Periodic table of the elements

A Periodic Table of the Elements is a table listing the elements in order of their atomic numbers:



Multiple attributes appear for each element in the table. Two of these attributes – atomic number and atomic mass – are directly related to the number of particles in the nucleus of each atom. We will examine the table's entry for the element *potassium* (K) to explore these concepts.

Potassium has an *atomic number* (number of protons in the nucleus of each potassium atom) of 19. This number defines the element. If we were somehow to add or subtract protons from the nucleus of a potassium atom[4], it would cease being potassium and *transmutate* into a different element. Note how *every* element in the table has its own unique atomic number, and how each of these numbers is whole (no fractions or decimals).

The *atomic mass* or *atomic weight* shown for potassium is 39.0983. This quantity is the sum of protons and neutrons found in the nucleus of each potassium atom. Like the atomic number (19), we would logically expect the atomic mass to be a whole number as well, since protons and neutrons

[4]The amount of energy required to rearrange particles in the nucleus for even just a single atom is *tremendous*, lying well outside the energy ranges of chemical reactions. Such energy levels are the exclusive domain of *nuclear* reactions and high-energy radiation (subatomic particles traveling at high velocity). The extremely large energy "investment" required to alter an atom's nucleus is why atomic identities are so stable. This is precisely why alchemists of antiquity utterly failed to turn lead into gold: no materials, processes, or techniques they had at their disposal were capable of the targeted energy necessary to dislodge three protons from a nucleus of lead ($_{82}$Pb) to that it would turn into a nucleus of gold ($_{79}$Au). That, and the fact the alchemists had no clue about atomic structure to begin with, made their endeavor fruitless.

only come in whole quantities. The primary reason we see a non-whole number for potassium's atomic mass is that this table reflects the *average* atomic mass of potassium atoms as found in nature. Some potassium atoms have atomic masses greater than 39, and some have atomic masses less than 39. We know that the number of protons in every potassium atom is fixed (which is what gives potassium its elemental identity), which means the only quantity that may cause the atomic mass to vary is the number of *neutrons* in the nucleus. The most common form of potassium (^{39}K) atom possesses 19 protons and 20 neutrons in its nucleus, giving it an atomic mass of 39 (19 + 20). The next most common form of potassium found on Earth is (^{41}K), possessing 19 protons and 22 neutrons.

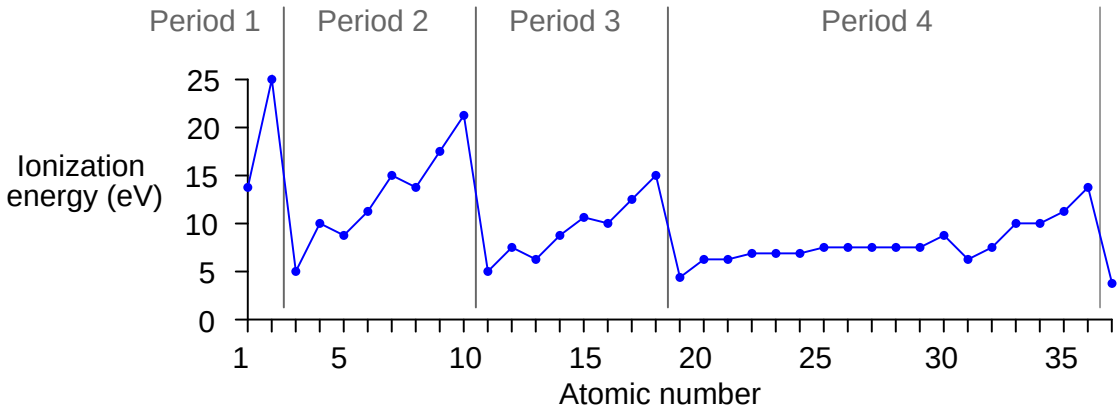
We refer to atoms of the same element with differing atomic masses as *isotopes*. From a chemical perspective, isotopes are identical. That is to say, they engage in the exact same chemical reactions in the exact same manner. To use potassium as an example, an atom of ^{39}K will join with a chlorine atom (Cl) to form the compound *potassium chloride* (KCl) just as readily as an atom of ^{41}K will join with a chlorine atom to form the same compound. The three isotopes of hydrogen (^{1}H, ^{2}H, and ^{3}H: hydrogen, deuterium, and tritium, respectively) are all chemically identical: all are highly flammable, combining with oxygen to create water (H_2O). However, deuterium (^{2}H) has twice the density of normal hydrogen (^{1}H), while tritium (^{3}H) has three times the density of normal hydrogen and is highly radioactive! Isotopes only differ in their mass and in their nuclear properties (such as *radioactivity*: the tendency for a nucleus to spontaneously decay, usually resulting in a loss or gain of protons that subsequently alters the identity of the decayed atom.).

The Periodic Table is called "periodic" because its configuration reveals a repeating pattern of chemical behaviors approximately following atomic number. Horizontal rows in the table are called *periods*, while vertical columns are called *groups*. Elements in the same group (vertical column) share similar chemical reactivities – that is, they tend to engage in the same types of chemical reactions – despite having different masses and physical properties such as melting point, boiling point, etc. This *periodicity* is a function of how electrons are arranged around the nucleus of each atom, a subject we will explore in more detail later in this chapter. As mentioned previously, chemistry is the study of how atoms bond together to form molecules, and this bonding takes place through the interaction of the electrons surrounding the atoms' nuclei. It makes perfect sense, then, that the configuration of those electrons determine the chemical (bonding) properties of atoms.

Some periodic tables show the *first ionization energy* value for each element – the amount of energy required to force the first electron of an electrically balanced atom to separate from that atom – in addition to other attributes such as atomic number and atomic mass. If we note the ionization energies of the elements, reading each element in turn from left-to-right, starting with period 1 (hydrogen and helium) and progressing to subsequent periods, we see an interesting pattern:

Element	**Period**	**First ionization energy** (measured in "electron-volts")
Hydrogen (H)	1	13.5984
Helium (He)	1	24.5874
Lithium (Li)	2	5.3917
Beryllium (Be)	2	9.3227
Boron (B)	2	8.2980
Carbon (C)	2	11.2603
Nitrogen (N)	2	14.5341
Oxygen (O)	2	13.6181
Fluorine (F)	2	17.4228
Neon (Ne)	2	21.5645
Sodium (Na)	3	5.1391
Magnesium (Mg)	3	7.6462
Aluminum (Al)	3	5.9858
Silicon (Si)	3	8.1517
Phosphorus (P)	3	10.4867
Sulfur (S)	3	10.3600
Chlorine (Cl)	3	12.9676
Argon (Ar)	3	15.7596
Potassium (K)	4	4.3407

The ionization energies increase with increasing atomic number (with an occasional down-step) until the last column of the period is reached, and then there is a comparatively enormous down-step in energy at the first column of a new period. This pattern is clearly evident when the first ionization energies are plotted against atomic number:



This periodicity suggests that as atoms grow in atomic number, the additional electrons do not simply pile on in random fashion or in a plain and simple progression from inner orbits to outer orbits. Rather, they "fill in" a structured energy pattern with major changes in structure at the start of each new period. More details of this structured pattern will be explored later in this chapter.

The low ionization energy values for all the "Group 1" elements (far left-hand column) suggest they are relatively easy to positively ionize, and indeed we find this to be the case through experimentation. Hydrogen, lithium, sodium, potassium, and the rest all readily become positively-charged ions upon interaction with other atoms, since their low ionization energy values means they may easily lose an electron.

The high ionization energy values for all the "Group 18" elements (far right-hand column) suggest they possess a very stable electron structure, which is also verified by experiment. These are the *noble* elements, possessing very little reactive potential[5].

Looking at the "Group 17" column, just to the left of the noble elements, we notice that they are all just one electron shy of the stable electron structure enjoyed by the noble atoms when in their electrically-balanced states. This suggests it might be easy to *add* one more electron to atoms of these elements, which (once again!) is a principle validated by experiment. Fluorine, chlorine, bromine, iodine, and even astatine[6] all readily ionize negatively, readily accepting an extra electron from surrounding atoms. As one might expect from this tendency, these elements readily bond through electrostatic attraction with the "Group 1" elements (hydrogen, lithium, sodium, potassium, etc.), each "Group 17" atom accepting an extra electron from each "Group 1" atom which is readily provides it. Ordinary table salt (sodium chloride, or NaCl) is an example of a compound formed by this sort of bond.

[5] It used to be believed that these elements were completely *inert*: incapable of forming molecular bonds with other atoms. However, this is not precisely true, as some compounds are now known to integrate noble elements.

[6] All isotopes of astatine (At) are radioactive with very short half-lives, making this element difficult to isolate and study.

3.4 Electronic structure

Somewhere in your education, you were probably shown a model of the atom showing a dense nucleus (comprised of protons and neutrons) surrounded by electrons whirling around like satellites around a planet. While there are some useful features of this model, it is largely in error. A more realistic view of atomic structure begins with the realization that electrons do not exist as discrete particles, but rather as wave packets. In a sense, they orbit the nucleus within certain *areas of probability*, as described by the principles of quantum mechanics. One way to envision this is to think of an electron's placement around the nucleus in the same way you might picture a city shrouded by a layer of fog. The electron does not have a discrete location (even though there *is* a discrete number of them found in every atom), but rather may be found anywhere within a certain region to varying degrees of probability.

Things get even stranger as we encounter atoms having multiple electrons. No two electrons may share the same quantum states in the same atom – a principle called the *Pauli Exclusion Principle*. This means the electrons surrounding a nucleus must exist in distinct patterns. Just a few of these patterns are shown here as *orbitals* (regions of high probability where up to two electrons may be found surrounding a nucleus):[7]

Electrons situate themselves around the nucleus of any atom according to one basic rule: the minimization of potential energy. That is, the electrons "try" to get as close to the nucleus as they can. Given the electrostatic attraction between negative electrons and the positive nucleus of an atom, there is potential energy stored in the "elevation" between an orbiting electron and the nucleus, just as there is gravitational potential energy in any object orbiting a planet. Electrons lose energy as they situate themselves closer to the nucleus, and it requires an external input of energy to move an electron further away from its parent nucleus.

In a sense, most of chemistry may be explained by this principle of minimized potential energy. Electrons "want" to "fall" as close as they can to the positively-charged nucleus. However, there is limited "seating" around the nucleus. As described by Pauli's Exclusion Principle, electrons cannot simply pile on top of each other in their quest for minimum energy, but rather must occupy certain regions of space where their quantum states will be unique.

An analogy for visualizing this is to think of it in terms of an amphitheater, having concentric rows of seats where spectators may view the event on stage. Everyone wants to be as close to the action as possible, but each person is constrained to sitting in one seat. As a result, all the inner seats are filled, with the only empty seats being further away from the stage. The concept of energy fits neatly into this analogy as well: just as electrons give up *energy* to "fall into" lower-energy

[7] These orbitals just happen to be the 1s, 2p, 3d, and 4f orbitals, as viewed from left to right. In each case, the nucleus lies at the geometric center of each shape. In a real atom, all orbitals share the same center, which means any atom having more than two electrons (that's all elements except for hydrogen and helium!) will have multiple orbitals around one nucleus. This four-set of orbital visualizations shows what some orbitals would look like if viewed in isolation.

regions around the nucleus, people must give up *money* to sit in the seats closest to the action on stage.

The energy levels available for orbiting electrons are divided into categories of *shells* and *subshells*. A "shell" (or, *principal quantum number*, n) describes the main energy level of an electron. In our amphitheater analogy, this is the equivalent of seating sections. A "subshell" (or, *subsidiary quantum number*, l) further divides the energy levels within each shell, and assigns different shapes to the electrons' probability "clouds." In the amphitheater analogy, this would be pairs of seats within each section having varying degrees of comfort, each identical seat pair being one *orbital*. Just as people want to sit as close to the action as possible (electrons occupying the lowest-value shell), they also desire to sit in the most comfortable seats they can find in each section (electrons occupying the lowest-energy subshell within each shell).

Chemists identify electron shells both by number (the value of the quantum number n) and/or by capital letters: the first shell by the letter K, the second by L, the third by M, and the fourth by N. Higher-order shells exist for atoms requiring a lot of electrons (high atomic number), and the lettering pattern is alphabetic (fifth shell is O, sixth is P, etc.). Each successive shell has a different number of subshells available, like amphitheater seating sections having different numbers of seats (the sections closest to the stage having the fewest seats, and the furthest sections having the most seats per section).

A numbering and lettering system is also used by chemists to identify subshells within each shell (the l quantum number value starting with zero, and lower-case letters beginning with "s"): the first subshell ($l = 0$) in any shell represented by the letter s, the second ($l = 1$) by p, the third ($l = 2$) by d, the fourth ($l = 3$) by f, and all others by successive lower-case letters of the alphabet. Each subshell of each shell has an even-numbered capacity for electrons, since the electrons in each subshell are organized in "orbital" regions, each orbital handling a maximum of two electrons. The number of orbitals per shell is equal to twice the l value plus one. An "s" subshell has one orbital holding up to two electrons. A "p" subshell has three orbitals holding up to six electrons total. A "d" subshell has five orbitals holding up to ten electrons total. An "f" subshell has seven orbitals holding up to 14 electrons total.

The number of subshells in any shell is the same as that shell's n value. Thus, the first (K) shell has but one subshell, "s". The second (L) shell has two subshells, an "s" and a "p". The third (M) shell has three subshells available, an "s", a "p", and a "d"; and so on.

Here is a list of the first few shells, their subshells, and electron capacity for each:

Shell	**Subshell**	**Subshell electron capacity**
$n = 1$; K	$l = 0$; s	2
$n = 2$; L	$l = 0$; s	2
	$l = 1$; p	6
$n = 3$; M	$l = 0$; s	2
	$l = 1$; p	6
	$l = 2$; d	10
$n = 4$; N	$l = 0$; s	2
	$l = 1$; p	6
	$l = 2$; d	10
	$l = 3$; f	14

The complete electron configuration for an atom may be expressed using *spectroscopic notation*, showing the shell numbers, subshell letters, and number of electrons residing within each subshell as a superscript. For example, the element Helium (with an atomic number of 2) would be expressed as $1s^2$, with just two electrons in the "s" subshell of the first shell. The following table shows the electron structures of the first nineteen elements in the periodic table, from the element hydrogen (atomic number = 1) to potassium (atomic number = 19):

Element	**Atomic number**	**Electron configuration**
Hydrogen	1	$1s^1$
Helium	2	$1s^2$
Lithium	3	$1s^2 2s^1$
Beryllium	4	$1s^2 2s^2$
Boron	5	$1s^2 2s^2 2p^1$
Carbon	6	$1s^2 2s^2 2p^2$
Nitrogen	7	$1s^2 2s^2 2p^3$
Oxygen	8	$1s^2 2s^2 2p^4$
Fluorine	9	$1s^2 2s^2 2p^5$
Neon	10	$1s^2 2s^2 2p^6$
Sodium	11	$1s^2 2s^2 2p^6 3s^1$
Magnesium	12	$1s^2 2s^2 2p^6 3s^2$
Aluminum	13	$1s^2 2s^2 2p^6 3s^2 3p^1$
Silicon	14	$1s^2 2s^2 2p^6 3s^2 3p^2$
Phosphorus	15	$1s^2 2s^2 2p^6 3s^2 3p^3$
Sulfur	16	$1s^2 2s^2 2p^6 3s^2 3p^4$
Chlorine	17	$1s^2 2s^2 2p^6 3s^2 3p^5$
Argon	18	$1s^2 2s^2 2p^6 3s^2 3p^6$
Potassium	19	$1s^2 2s^2 2p^6 3s^2 3p^6 4s^1$

In order to avoid having to write unwieldy spectroscopic descriptions of each element's electron structure, it is customary to write the notation only for subshells that are unfilled. For example, instead of writing the electron structure of the element Aluminum as $1s^2 2s^2 2p^6 3s^2 3p^1$, we might just as well write a condensed version showing only the last subshell ($3p^1$), since all the previous subshells are completely full.

Another way to abbreviate the spectroscopic notation for elements is to condense all the shells below the newest (unfilled) shell as the corresponding noble element, in brackets. To use the example of Aluminum again, we could write its spectroscopic notation as $[Ne]3s^2 3p^1$ since its shell 1 and shell 2 configurations are completely described by the electron configuration of Neon.

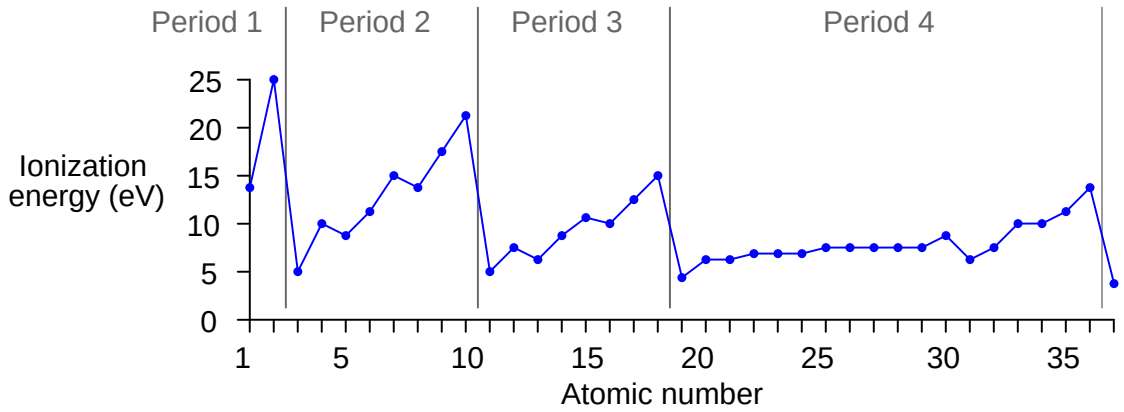
Re-writing our electron shell table for the first nineteen elements using this condensed notation:

Element	Atomic number	Electron configuration
Hydrogen	1	$1s^1$
Helium	2	$1s^2$
Lithium	3	$[He]2s^1$
Beryllium	4	$[He]2s^2$
Boron	5	$[He]2s^22p^1$
Carbon	6	$[He]2s^22p^2$
Nitrogen	7	$[He]2s^22p^3$
Oxygen	8	$[He]2s^22p^4$
Fluorine	9	$[He]2s^22p^5$
Neon	10	$[He]2s^22p^6$
Sodium	11	$[Ne]3s^1$
Magnesium	12	$[Ne]3s^2$
Aluminum	13	$[Ne]3s^23p^1$
Silicon	14	$[Ne]3s^23p^2$
Phosphorus	15	$[Ne]3s^23p^3$
Sulfur	16	$[Ne]3s^23p^4$
Chlorine	17	$[Ne]3s^23p^5$
Argon	18	$[Ne]3s^23p^6$
Potassium	19	$[Ar]4s^1$

If we progress from element to element in increasing atomic number, we see that no new shell begins to form until after we reach the noble element for that period[8] at the far right-hand column. With the beginning of each new period at the far-left end of the Table, we see the beginning of the next higher-order electron shell. The shell(s) below are represented by whichever noble element shares that same configuration, indicating a "noble core" of electrons residing in extremely stable (low-energy) regions around the nucleus.

[8] Recall the definition of a "period" in the Periodic Table being a horizontal row, with each vertical column being called a "group".

The beginning of the next higher-order shell is what accounts for the periodic cycle of ionization energies we see in elements of progressing atomic number. The first electron to take residence in a new shell is very easy to remove, unlike the electrons residing in the "noble" configuration shell(s) below:

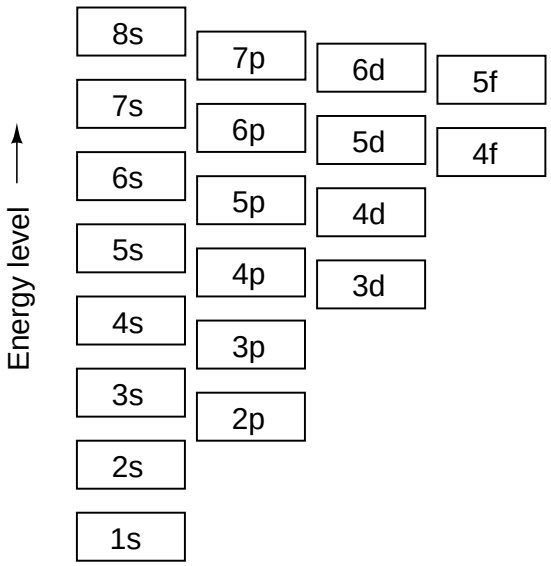


Not only is the "noble core" notation a convenience for tersely describing the electron structure of an element, but it also represents an important concept in chemistry: the idea of *valence*. Electrons residing in lower-order shells are, by definition, at lower energy states than electrons residing in higher-order shells. Therefore, the electrons in unfilled shells are more readily pulled away from the atom than those lying in completely full shells below. These "outer" electrons are called *valence electrons*, and their number determines how readily an atom will chemically interact with another atom.

If we examine the electron structures of atoms with successively greater atomic numbers (more protons in the nucleus, therefore more electrons in orbit to balance the electrical charge), we notice that the shells and subshells fill up in an interesting pattern. One might think that all the lower-order shells get completely filled before any electrons go into a higher-order shell – just as we might expect people to fill every seat in all the lower seating sections of an amphitheater before filling any of the higher seats – but this is not the case for every elements. Instead, the energy levels of subshells within shells is split, such that certain subshells within a higher shell will have a lower energy value than certain subshells within a lower shell. Referring back to our amphitheater analogy, where seating sections represented shells and seats of varying comfort represented subshells, it is as though people choose to fill the more comfortable seats in the next higher seating section before taking the less-comfortable seats in lowest available section, the desire for comfort trumping the desire for proximity to the stage.

A rule commonly taught in introductory chemistry courses called the *Madelung rule* (also referred to as *Aufbau order*, after the German verb *aufbauen* meaning "to build up") is that subshells fill with increasing atomic number in such an order that the subshell with the lowest $n + l$ value, in the lowest shell, gets filled before any others.

The following graphic illustrates this ordering:



Madelung filling order: 1s → 2s → 2p → 3s → 3p → 4s → 3d → 4p → 5s → 4d → 5p → 6s → 4f → 5d → 6p → 7s → 5f → 6d → 7p → 8s → *(etc.)*

It should be noted that exceptions exist for this rule. We see one of those exceptions with the element chromium ($_{24}$Cr). Strictly following the Madelung rule in progressing from vanadium (atomic number = 23, valence electron structure $3d^3 4s^2$) to chromium (atomic number = 24), we would expect the next electron to take residence in the "3d" subshell making chromium's valence structure be $3d^4 4s^2$, but instead we find *two more* electrons residing in chromium's 3d subshell with one less in the 4s subshell ($3d^5 4s^1$). The sequence resumes its expected progression with the next element, manganese (atomic number = 25, valence electron structure $3d^5 4s^2$). The general principle of energy minimization still holds true . . . it's just that the relative energies of succeeding subshells do not follow a simple rule structure. In other words, the Aufbau order is an over-simplified view of reality. To use the amphitheater analogy again, it's as if someone gave up one of the nice chairs in seating section 4 to be closer to a friend who just occupied one of the less comfortable chairs in seating section 3.

The practical importance of electron configurations in chemistry is the potential energy possessed by electrons as they reside in different shells and subshells. This is extremely important in the formation and breaking of chemical bonds, which occur due to the interaction of electrons between two or more atoms. A chemical bond occurs between atoms when the outer-most (valence) electrons of those atoms mutually arrange themselves in energy states that are collectively lower than they would be individually. The ability for different atoms to join in chemical bonds completely depends upon the default energy states of electrons in each atom, as well as the next available energy states in the other atoms. Atoms will form stable bonds only if the union allows electrons to assume lower energy levels. This is why different elements are very selective regarding which elements they

will chemically bond with to form compounds: not all combinations of atoms result in decreased potential energy.

The amount of energy required to break a chemical bond (i.e. separate the atoms from each other) is the same amount of energy required to restore the atoms' electrons to their previous (default) states before they joined. This is the same amount of energy released by the atoms as they come together to form the bond. Thus, we see the foundation of the general principle in chemistry that forming chemical bonds releases energy, while breaking chemical bonds requires an input of energy from an external source. We also see in this fact an expression of the Conservation of Energy: all the energy "invested" in breaking bonds between different atoms is accounted for in the energy release occurring when those atoms re-join.

In summary, the whole of chemistry is a consequence of electrons not being able to assume arbitrary positions around the nucleus of an atom. Instead, they seek the lowest possible energy levels within a framework allowing them to retain unique quantum states. Atoms with mutually agreeable electron structures readily bond together to form molecules, and they release energy in the process of joining. Molecules may be broken up into their constituent atoms, if sufficient energy is applied to overcome the bond. Atoms with incompatible electron structures do not form stable bonds with each other.

3.5 Spectroscopy

Much of our knowledge about atomic structure comes from experimental data relating the interaction between *light* and atoms of the different elements. Light may be modeled as an electromagnetic wave, consisting of an oscillating electric field and an oscillating magnetic field. Like any wave, the relationship between propagation velocity, wavelength, and frequency is described by the following equation:

$$v = \lambda f$$

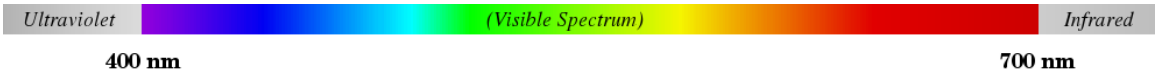
Where,

v = Velocity of propagation (e.g. meters per second)
λ = Wavelength (e.g. meters)
f = Frequency of wave (e.g. Hz, or 1/seconds)

When applied to light waves, this equation is typically written as $c = \lambda f$, where c is the velocity of light in a vacuum: one of the fundamental constants of physics.

Light that is visible to the human eye has wavelengths approximately between 400 nm (400 *nanometers*) at the violet end of the spectrum and 700 nm at the red end of the spectrum. Given the velocity of light (approximately 3×10^8 m/s), this equates to a frequency range for visible light between 7.5×10^{14} Hz and 4.286×10^{14} Hz.

A computer-generated image of the visible light spectrum (plus the ultraviolet and infrared regions outside of the visible range, shown in grey) appears here. A real spectrum may be generated by taking "white" light and passing it through either a prism or a diffraction grating so that the different wavelengths separate from each other:



Just like buoyant objects are moved up and down by waves of water, electrically-charged objects may be moved about by waves of electrical fields such as light. In the case of electrons, their positions around the nucleus of an atom may be altered if struck by light of the correct wavelength.

One of the major breakthrough discoveries of modern physics was the realization that a ray of light could be modeled as a *particle* (called a *photon*) possessing a definite amount of energy, in addition to being modeled as a wave with a definite frequency. The combined work of physicists Max Planck in 1900 and Albert Einstein in 1905 resulted in the following equation relating a photon's energy to its frequency:

$$E = hf$$

Where,

E = Energy carried by a single "photon" of light (joules)
h = Planck's constant (6.626×10^{-34} joule-seconds)
f = Frequency of light wave (Hz, or 1/seconds)

We may re-write this equation to express a photon's energy in terms of its wavelength (λ) rather than its frequency (f), knowing the equation relating those two variables for waves of light ($c = \lambda f$):

$$E = \frac{hc}{\lambda}$$

If the amount of energy carried by a photon happens to match the energy required to make an atomic electron "jump" from one energy level to another within the atom, the photon will be consumed in the work of that task when it strikes the atom. Conversely, when that "excited" electron returns to its original (lower) energy level in the atom, it releases a photon of the same frequency as the original photon that excited the electron.

Since the energy levels available for an electron to "jump" within an atom are limited to certain fixed values, this means only certain specific frequencies or wavelengths of light will be able to make an electron of a particular atom move to new shells and/or subshells. The sentence you just read is actually backward from an historical perspective: what came first was the discovery that only certain wavelengths of light were associated with atomic energy changes, and from that came the extrapolation that the energy levels of atomic electrons must be *quantized* (limited to definite, fixed values and not continuously variable as previously thought). This was a tremendous discovery, and it put physics on a whole new path toward a *quantum* model of matter and energy.

This is why the notation used in the previous section to describe electron configurations (e.g. $1s^2 2s^2 2p^1$) is called *spectroscopic* notation: the discovery of shells, subshells, and orbitals owes itself to the analysis of light wavelengths associated with different types of atoms, studied with a device called a *spectroscope* constructed to analyze the wavelengths of light across the visible spectrum.

3.5.1 Emission spectroscopy

If we take a sample of atoms, all of the same element and at a low density[9] (e.g. a gas or vapor), and "excite" them with a source of energy such as an electric arc, we will notice those atoms emit colors of light that are characteristically unique to that element. The unique electron configurations of each element creates a unique set of energy values between which atomic electrons of that element may "jump." Since no two elements have the exact same electron configurations, no two elements will have the same exact set of available energy levels for their electrons. When excited electrons fall back into their normal ("ground state") energy levels, the photons they emit will have distinct wavelengths. The result is an *emission spectrum* of light wavelengths, much like a "fingerprint" unique to that element. Indeed, just as fingerprints may be used to identify a person, the spectrum of light emitted by an "excited" sample of an element may be used to identify that element!

For example, we see here the emission spectrum for *hydrogen*, shown immediately below the continuous spectrum of visible light for convenient reference[10]:

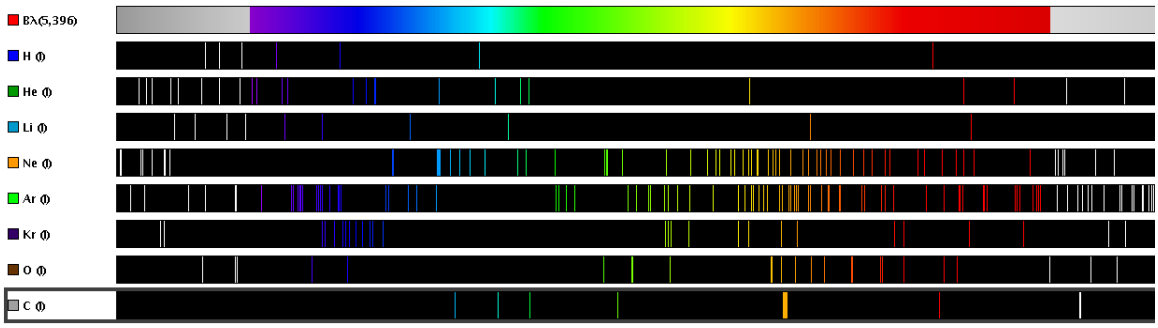
Each of the colored "lines" in the emission spectrum for hydrogen represents the photon wavelength emitted when the excited electron loses energy and falls back into a lower-level position. The larger the energy difference between energy levels (i.e. the bigger the jump), the more energy the photon carries away, and consequently the shorter the wavelength (higher the frequency) of the photon. The violet color line, therefore, represents one of the larger "jumps" while the red color line represents one of the smaller. Hydrogen happens to emit four different wavelengths within the visible range (656 nm, 486 nm, 434 nm, and 410 nm)[11], and many others outside the visible range.

[9] Solids and liquids tend to emit a broad spectrum of wavelengths when heated, in stark contrast to the distinct "lines" of color emitted by isolated atoms.

[10] To create these spectra, I used a computer program called *Spectrum Explorer*, or `SPEX`.

[11] These wavelengths are part of the *Balmer series* of spectral lines, corresponding to electrons falling from the 3rd, 4th, 5th, and 6th shells down to the 2nd shell. Photons emitted by excited electrons returning to the "ground level" shell lie outside the visible range, since the transition from the second shell (L shell, $n = 2$) to the first shell (K shell, $n = 1$) is a much larger energy gap, producing a photon of much greater energy and shorter wavelength (approximately 122 nm) than those found within the visible light spectrum.

This next graphic shows the emission spectra of several elements contrasted against a continuous spectrum covering both visible light and portions of the ultraviolet and infrared ranges:



Note how complex the emission spectra are for some of the elements, and how spectral lines for most elements (including hydrogen) extend past the visible light range.

Not only may the wavelengths of photons emitted from "excited" electrons returning to lower-energy conditions be used to positively identify different elements, but we may also use those wavelengths as universal standards, since the fundamental properties of elements are not liable to change. For example, the SI (Système International) definition for the base unit of the *meter* is standardized as 1,650,763.73 wavelengths of light emitted by a krypton-86 (^{86}Kr) atom as its electrons transition between the $2p^{10}$ and $5d^5$ subshells[12].

3.5.2 Absorption spectroscopy

If we take a sample of atoms, all of the same element and at a low density (e.g. a gas or vapor), and pass a continuous ("white") spectrum of light wavelengths through that sample, we will notice certain colors of light *missing* from the light exiting the sample. Not only are these missing wavelengths characteristically unique to that element, but they are *the exact same wavelengths of light found in the emission spectrum for that element!* The same photon wavelengths produced by an atom when "excited" by an external energy source will be readily *absorbed* by that atom if exposed to it. Thus, the spectrum of light missing characteristic wavelengths after passing through a gas sample is called an *absorption spectrum*, and may be used to identify elements just as easily[13] as an emission spectrum.

[12] The wavelength of this light happens to lie within the visible range, at approximately 606 nm.

[13] In fact, it is often easier to obtain an absorption spectrum of a sample than to create an emission spectrum, due to the relative simplicity of the absorption spectrometer test fixture. We don't have to energize a sample to incandescence to obtain an absorption spectrum – all we must do is pass white light through enough of it to absorb the characteristic colors.

The absorption spectrum of hydrogen gas is shown at the bottom of this three-spectrum graphic image, contrasted against the continuous spectrum of visible light (top) and the emission spectrum for hydrogen (middle):

Note how the four colored lines in the emission spectrum characteristic of hydrogen appear as *missing* colors (black lines) in the absorption spectrum. It is almost as though one spectrum were a photographic "negative" of the other: each of the colors present in the emission spectrum is distinctly missing in the absorption spectrum. Although the color patterns may be inverted, the positions of the lines within the spectrum are the same, and are *uniquely* representative of hydrogen.

Individual atoms are not the only forms of matter possessing uniquely identifying spectra – many *molecules* have spectral "signatures" of their own as well. The absorption spectra for molecular substances are substantially more complex than the absorption spectra of pure elements, owing to the many more different ways in which light energy may be absorbed by a molecule. In addition to electron shell and subshell "jumps" capable of absorbing a photon's energy, the atoms within a molecule are also able to vibrate, rotate, and twist about each other like mechanical oscillators. Photons of light possessing just the right frequencies are able to "excite" certain molecules in a manner not unlike AC electrical waveforms resonating with tuned LC (inductor-capacitor) circuits. Just as tuned LC circuits absorb and store energy at certain frequencies, molecular oscillators absorb and store energy from photons.

The multiplicity of energy-absorbing modes for certain molecules gives them wide *bands* of absorption in the light spectrum, not just thin "lines" as is the case with individual atoms. These bands are still unique to each molecule type, but they typically cover a far broader swath of wavelengths than is typical for atomic absorption spectra.

The absorption of ultraviolet light by ozone gas (O_3) high in Earth's atmosphere is an example of absorption spectroscopy on a grand scale. These molecules serve as a protective "blanket" against ultraviolet light rays from the sun which have detrimental effects on life (e.g. sunburn, skin cancer). The ozone does not absorb light in the visible spectrum, and so its protective effects are not visually apparent, but the attenuation of ultraviolet light is definitely measurable. This attenuation also covers far more than just one or two specific wavelengths of ultraviolet light, which is good for life on Earth because otherwise ozone wouldn't offer much protection.

Many chemical substances of interest in process industries have well-known *absorption signatures* for ultraviolet and infrared light. This makes spectroscopy a powerful tool for the identification (and quantitative measurement) of chemical composition in process fluids, exhaust gases, and sometimes even in solid materials. For more detail on the practical application of spectroscopy to analytical measurement, refer to section 22.4 beginning on page 1170.

3.6 Formulae for common chemical compounds

Most of these formulae appear in *molecular chemical* form rather than structural form. For example, ethanol appears here as C_2H_6O rather than C_2H_5OH. Also, the entries for fructose and glucose are identical ($C_6H_{12}O_6$) despite the two compounds having different structures. This means most of the formulae shown in this section merely represent the ratios of each element in a compound, making little or no attempt to convey the *structure* of the molecule.

It should be noted that this list is definitely *not* exhaustive, but merely attempts to show formulae for some common compounds.

- Acetone: C_3H_6O
- Acetylene: C_2H_2
- Alcohol, methyl (methanol): CH_4O
- Alcohol, butyl (butanol): $C_4H_{10}O$
- Alcohol, ethyl (ethanol): C_2H_6O
- Alcohol, phenol: C_6H_6O
- Aluminum oxide (alumina): Al_2O_3
- Ammonia: NH_3
- Ammonium carbonate: $(NH_4)_2CO_3$
- Ammonium chloride (sal ammoniac): NH_4Cl
- Ammonium nitrate: $N_2H_4O_3$
- Aromatic hydrocarbons:
  - Acetylene: C_2H_2
  - Ethylene: C_2H_4
  - Propylene: C_3H_6
  - Butylene: C_4H_8
  - Benzene: C_6H_6
  - Toluene: C_7H_8
  - Styrene: C_8H_8
  - Napthalene: $C_{10}H_8$
- Calcium carbonate (limestone, marble): $CaCO_3$
- Calcium chloride: $CaCl_2$
- Calcium hydroxide: $Ca(OH)_2$
- Calcium oxide (lime or quicklime): CaO

- Calcium sulfate (gypsum): $CaSO_4$
- Carbon monoxide: CO
- Carbon dioxide: CO_2
- Carbon tetrachloride: CCl_4
- Carbonic acid: H_2CO_3
- Cellulose: $(C_6H_{10}O_5)_n$
- Clay (or shale): $H_4Al_2Si_2O_9$
- Copper oxide (cuprite): Cu_2O
- Copper oxide (tenorite): CuO
- Cyanic acid: HOCN
- Dextrose (synonym for biological glucose): $C_6H_{12}O_6$
- Ethyl mercaptan: C_2H_6S
- Ethylene glycol: $C_2H_6O_2$
- Ethylene oxide: C_2H_4O
- Formaldehyde: CH_2O
- Folic acid: $C_{19}H_{19}N_7O_6$
- Formaldehyde: CH_2O
- Formic acid: CH_2O_2
- Fructose (same molecular formula as glucose): $C_6H_{12}O_6$
- Glycerol: $C_3H_8O_3$
- Hydrazine: N_2H_4N
- Hydrocyanic acid: HCN
- Hydrofluoric acid: HF
- Hydrochloric acid: HCl
- Hydrogen peroxide: H_2O_2
- Hydrogen sulfide: H_2S
- Iron oxide: Fe_2O_3
- Magnesium hydroxide (milk of magnesia): $Mg(OH)_2$

- Nitric acid: HNO_3
- Nitric oxide: NO
- Nitrogen dioxide: NO_2
- Nitrogen trioxide: NO_2
- Nitroglycerine: $C_3H_5N_3O_9$
- Nitromethane: CH_3NO_2
- Nitrous oxide: N_2O
- Dinitrogen trioxide: N_2O_3
- Ozone: O_3
- Paraffinic hydrocarbons:
  - Methane: CH_4
  - Ethane: C_2H_6
  - Propane: C_3H_8
  - Butane: C_4H_{10}
  - Pentane: C_5H_{12}
  - Hexane: C_6H_{14}
  - Heptane: C_7H_{16}
  - Octane: C_8H_{18}
  - Nonane: C_9H_{20}
  - Decane: $C_{10}H_{22}$
- Phosgene: $COCl_2$
- Phosphoric acid: H_3PO_4
- Potassium chloride: KCl
- Potassium cyanide: KCN
- Potassium hydroxide: KOH
- Potassium sulfate: K_2SO_4
- Silane: SiH_4
- Silica: SiO_2
- Silicon carbide: SiC
- Sodium chloride (table salt): NaCl

- Sodium hydroxide: $NaOH$
- Sodium fluoride: NaF
- Strychnine: $C_{21}H_{22}N_2O_2$
- Sucrose: $C_{12}H_{22}O_{11}$
- Sulfuric acid: H_2SO_4
- Sulfur dioxide: SO_2
- Sulfur hexafluoride: SF_6
- Testosterone: $C_{19}H_{28}O_2$
- Turpentine: $C_{10}H_{16}$ (approx.)
- Zinc sulfate: $ZnSO_4$

3.7 Molecular quantities

Sample sizes of chemical substances are often measured in *moles*. One mole of a substance is defined as a sample having 6.022×10^{23} (*Avogadro's number*) molecules[14]. An elemental sample's mass is equal to its molecular quantity in moles multiplied by the element's atomic mass in *amu* (atomic mass units, otherwise known as *Daltons*). For example, 2.00 moles of naturally-occurring potassium will have a mass of 78.2 grams. The value of Avogadro's number is not arbitrary – it was chosen to be a direct proportion between an element's atomic mass value and the mass of a pure monatomic sample of that element, in order to simplify calculations of sample masses based on known composition.

Molar quantities make it convenient to relate macroscopic samples of elements and compounds with each other. We know, for instance, that one mole of naturally occurring iron (Fe) atoms will have a mass of 55.8 grams, and that one mole of naturally occurring oxygen (O) atoms will have a mass of 16.0 grams, because the average atomic mass of naturally occurring iron is 55.8 amu, and the average atomic mass of naturally occurring oxygen is 16.0 amu. One mole of naturally occurring oxygen *molecules* (O_2) will have a mass of 32.0 grams, since each molecule is a *pair* of oxygen atoms at 16 amu each, and "moles" counts the number of discrete entities which in the case of molecular oxygen is the number of O_2 *molecules* rather than the number of O *atoms*. Applying the same reasoning, one mole of ozone (O_3) molecules will have a mass of 48.0 grams.

The same mathematical proportions apply to compounds as they do to elements, since compounds are nothing more than different elements bound together in whole-number ratios, and the Conservation of Mass tells us a molecule cannot have a mass greater or less than the sum total of the constituent elements' masses. To illustrate this principle, we may calculate the mass of one mole of iron oxide (Fe_2O_3), the principal component of *rust*: $55.8 \times 2 + 16.0 \times 3 = 159.6$ grams. Likewise, we may calculate the mass of five moles of pure glucose ($C_6H_{12}O_6$): $5 \times (12.01 \times 6 + 1.01 \times 12 + 16.0 \times 6) = 900.0$ grams. A convenient sum of the atomic weights of a molecule's constituent atoms is called the *molecular weight*. In the case of iron oxide, the molecular weight is 159.6 (typically rounded up to 160). In the case of molecular oxygen (O_2), the molecular weight is 32, since each O_2 molecule contains two atoms at 16 amu each.

When referring to liquid solutions, the concentration of a solute is often expressed as a *molarity*, defined as the number of moles of solute per liter of solution. Molarity is usually symbolized by an italicized capital letter *M*. It is important to bear in mind that the volume used to calculate molarity is that of the total solution (solute plus solvent) and not the solvent alone.

Suppose we had a solution of salt-water, comprised of 33.1 grams of table salt thoroughly mixed with pure water to make a total volume of 1.39 liters. In order to calculate the molarity of this solution, we first need to determine the equivalence between moles of salt and grams of salt. Since table salt is sodium chloride (NaCl), and we know the atomic masses of both sodium (23.0 amu) and chlorine (35.5 amu), we may easily calculate the mass of one mole of salt:

$$1 \text{ mole of NaCl} = 23.0 \text{ g} + 35.5 \text{ g} = 58.5 \text{ g}$$

Another way to state this is to say that sodium chloride (NaCl) has a *formula weight* of 58.5 amu (58.5 grams of mass per mole).

[14] Truth be told, a "mole" is 6.022×10^{23} of literally *any* discrete entities. There is nothing wrong with measuring the amount of eggs in the world using the unit of the mole. Think of "mole" as a *really* big dozen!

We may use this equivalence as a unity fraction to help us convert the number of grams of salt per unit volume of solution into a molarity (moles of salt molecules per liter):

$$\left(\frac{33.1 \text{ g}}{1.39 \text{ l}}\right)\left(\frac{1 \text{ mol}}{58.5 \text{ g}}\right) = 0.407\frac{\text{mol}}{\text{l}} = 0.407\ M$$

3.8 Stoichiometry

Stoichiometry is the accounting of atoms in a chemical equation. It is an expression of the *Law of Mass Conservation*, in that elements are neither created nor destroyed in a chemical reaction, and that mass is an intrinsic property of every element. Thus, the numbers, types of atoms, and total mass in a reaction product sample must be the same as the numbers, types of atoms, and total mass in the reactants which reacted to produce it. For example, in the combustion of natural gas in an oxygen-rich environment, the fuel (CH_4) and oxidizer (O_2) are the reactants, while water vapor (H_2O) and carbon dioxide gas (CO_2) are the reaction products:

$$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O$$

Reactants	**Reaction products**	**Mass** (per mole of CH_4)
Carbon = 1 × 1	Carbon = 1 × 1	12 grams
Hydrogen = 1 × 4	Hydrogen = 2 × 2	4 grams
Oxygen = 2 × 2	Oxygen = (1 × 2) + (2 × 1)	64 grams

As you can see in this example, every single atom (and its mass) entering the reaction is accounted for in the reaction products. The only exception to this rule is in *nuclear reactions* where elements transmutate into different elements, with gains or losses in nuclear particles. No such transmutation occurs in any mere *chemical* reaction, and so we may safely assume equal numbers and types of atoms before and after any chemical reaction. Chemical reactions strictly involve re-organization of molecular bonds, with electrons as the constituent particles comprising those bonds. Nuclear reactions involve the re-organization of atomic nuclei (protons, neutrons, etc.), with far greater energy levels associated.

Often in chemistry, we know both the reactant and reaction product molecules, but we need to determine their relative numbers before and after a reaction. The task of writing a general chemical equation and then assigning multiplier values for each of the molecules is called *balancing the equation.*

3.8.1 Balancing chemical equations by trial-and-error

Balancing a chemical equation is a task that may be done by trial-and-error. For example, let us consider the case of complete combustion for the hydrocarbon fuel *ethane* (C_2H_6) with oxygen (O_2). If combustion is complete, the reaction products will be water vapor (H_2O) and carbon dioxide (CO_2). The unbalanced equation representing all reactants and products for this reaction is shown here, along with a table showing the numbers of atoms on each side of the equation:

$$C_2H_6 + O_2 \rightarrow H_2O + CO_2$$

Reactants	**Reaction products**
Carbon = 2	Carbon = 1
Hydrogen = 6	Hydrogen = 2
Oxygen = 2	Oxygen = 3

Clearly, this is not a balanced equation, since the numbers of atoms for each element are unequal between the two sides of the equation.

A good place to start in balancing this equation is to look for an element represented by only one molecule on each side of the equation. Carbon is an example (present in the ethane but not in the oxygen molecule on the left-hand side, and in the carbon dioxide but not the water on the right-hand side) and hydrogen is another.

Beginning with carbon, we see that each ethane molecule contains two carbon atoms while each carbon dioxide molecule contains just one carbon atom. Therefore, we may conclude that the ratio of carbon dioxide to ethane must be 2-to-1, no matter what the other ratios might be. So, we double the number of carbon dioxide molecules on the right-hand side and re-check our atomic quantities:

$$C_2H_6 + O_2 \rightarrow H_2O + 2CO_2$$

Reactants	**Reaction products**
Carbon = 2	Carbon = 2
Hydrogen = 6	Hydrogen = 2
Oxygen = 2	Oxygen = 5

Next, we will balance the hydrogen atom numbers, since we know hydrogen is an element found in only one molecule on each side of the equation. Our hydrogen ratio is now 6:2 (left:right), so we know we need three times as many hydrogen-containing molecules on the right-hand side. Tripling the number of water molecules gives us:

$$C_2H_6 + O_2 \rightarrow 3H_2O + 2CO_2$$

Reactants	**Reaction products**
Carbon = 2	Carbon = 2
Hydrogen = 6	Hydrogen = 6
Oxygen = 2	Oxygen = 7

Unfortunately, the numbers of oxygen atoms on each side of the equation are unequal, and it is not immediately obvious how to make them equal. We need five more atoms of oxygen on the left-hand side, but we cannot add exactly five more because oxygen atoms only come to us in pairs (O_2), limiting us to even-number increments.

However, if we *double* all the other molecular quantities, it will make the disparity of oxygen atoms an even number instead of an odd number:

$$2C_2H_6 + O_2 \rightarrow 6H_2O + 4CO_2$$

Reactants	**Reaction products**
Carbon = 4	Carbon = 4
Hydrogen = 12	Hydrogen = 12
Oxygen = 2	Oxygen = 14

Now it is a simple matter to balance the number of oxygen atoms, by adding six more oxygen molecules to the left-hand side of the equation:

$$2C_2H_6 + 7O_2 \rightarrow 6H_2O + 4CO_2$$

Reactants	**Reaction products**
Carbon = 4	Carbon = 4
Hydrogen = 12	Hydrogen = 12
Oxygen = 14	Oxygen = 14

Now the equation is balanced: the quantities of each type of atom on both sides of the equation are equal.

3.8.2 Balancing chemical equations using algebra

A more mathematically sophisticated approach to stoichiometry involves the use of *simultaneous systems of linear equations*. The fundamental problem chemists must solve when balancing reaction equations is to determine the ratios of reactant and product molecules. If we assign a variable to each molecular quantity, we may then write a mathematical equation for each element represented by the reaction, and use algebra to solve for the variable values.

To illustrate, let us balance the equation describing the attack of aluminum metal's protective "passivation" layer of oxide by acid rain. When aluminum metal is exposed to oxygen, the outer surface of the metal quickly forms a layer of aluminum oxide (Al_2O_3) which acts to impede further oxidation of the metal. This protective layer, however, may be attacked by the presence of sulfuric acid (H_2SO_4). This acid finds its way into rainwater by way of sulfur compounds emitted during the combustion of fuels containing sulfur. The products of this reaction between sulfuric acid and aluminum oxide are a sulfate molecule ($Al(SO_4)_3$) and water (H_2O), as illustrated in this *unbalanced* chemical equation:

$$H_2SO_4 + Al_2O_3 \rightarrow Al_2(SO_4)_3 + H_2O$$

This equation contains four different molecules (acid, aluminum oxide, sulfate, and water), which means we ultimately must solve for four different quantities. It also contains four different elements (H, S, O, and Al). Since the mathematical requirement for solving a system of linear equations is to have at least one equation per variable, it would first appear as though we could set up a 4×4 matrix (four equations of four variables). However, this will not work. If we tried to solve for four unknown quantities, we would ultimately be foiled by an infinite number of solutions. This makes sense upon further inspection, since any stoichiometric solution to this chemical reaction will have an infinite number of correct *proportions* to satisfy it[15]. What we need to do is arbitrarily set one of these molecular quantities to a constant value (such as 1), then solve for the quantities of the other three. The result will be ratios or proportions of all the other molecules to the fixed number we assigned to the one molecule type.

[15] Take the combustion of hydrogen and oxygen to form water, for example. We know we will need two H_2 molecules for every one O_2 molecule to produce two H_2O molecules. However, *four* hydrogen molecules combined with *two* oxygen molecules will make *four* water molecules just as well! So long as we consider all three molecular quantities to be unknown, we will never be able to solve for just *one* correct answer, because there is no one correct set of absolute quantities, only one correct set of *ratios* or *proportions*.

As an example, I will choose to set the number of acid molecules to 1, and use the variables x, y, and z to solve for the numbers of the other molecules (oxide, sulfate, and water, respectively):

1	x	=	y	z
H_2SO_4	Al_2O_3	$\rightarrow$	$Al_2(SO_4)_3$	H_2O

Now, I will write four different equations, each one representing the stoichiometric balance of one element in the chemical equation. The following table shows each of the four elements with their respective balance equations:

Element	**Balance equation**
Hydrogen	$2 + 0x = 0y + 2z$
Sulfur	$1 + 0x = 3y + 0z$
Oxygen	$4 + 3x = 12y + 1z$
Aluminum	$0 + 2x = 2y + 0z$

Simplifying each equation by eliminating all zero values and "1" coefficients:

Element	**Balance equation**
Hydrogen	$2 = 2z$
Sulfur	$1 = 3y$
Oxygen	$4 + 3x = 12y + z$
Aluminum	$2x = 2y$

We can see by examination of the first, second, and fourth equations that z must be equal to 1, y must be equal to $\frac{1}{3}$, and that x and y are equal to each other (therefore, x must be equal to $\frac{1}{3}$ as well). Plugging these values into the variables of the third equation confirms this ($4 + 1 = 4 + 1$). Thus, our solution to this multi-variable system of equations is:

$$x = \frac{1}{3} \qquad y = \frac{1}{3} \qquad z = 1$$

It makes little sense to speak of *fractions* of a molecule, which is what the values of x and y seem to suggest, but we must recall these values represent *proportions* only. In other words, we need but one-third as many oxide and sulfate molecules as acid and water molecules to balance this equation. If we multiply all these values by three (as well as the initial constant we chose for the number of acid molecules), the quantities will be whole numbers and the chemical reaction will still be balanced:

$$x = 1 \qquad y = 1 \qquad z = 3$$

Thus, our final (balanced) equation showing the attack of aluminum metal's passivation layer by acid rain is as follows:

$$3H_2SO_4 + Al_2O_3 \rightarrow Al_2(SO_4)_3 + 3H_2O$$

Another example to illustrate this method of balancing chemical equations is the oxidation of wastewater (sewage) sludge. Here, the reactant is not a single type of molecule, but rather a complex mixture of carbohydrates, proteins, fats, and other organic compounds. A practical way of dealing with this problem is to represent the average quantities of carbon, hydrogen, and oxygen in the form of a *compositional formula* determined from a gross analysis of the wastewater sludge:

$$C_{4.8}H_{8.4}O_{2.2}$$

We know that the reaction products will be carbon dioxide and water, but the question is how much oxygen we will need to permit complete oxidation. The following (unbalanced) chemical equation shows the reactants and reaction products:

$$C_{4.8}H_{8.4}O_{2.2} + O_2 \rightarrow CO_2 + H_2O$$

The non-integer subscripts complicate trial-and-error stoichiometry, but they pose absolutely no obstacle at all to simultaneous equations. Assigning variables x, y, and z to the unknown molecular quantities:

1	x	=	y	z
$C_{4.8}H_{8.4}O_{2.2}$	O_2	$\rightarrow$	CO_2	H_2O

Now, we may write three different equations, each one representing the stoichiometric balance of one element in the chemical equation. The following table shows each of the three elements with their respective balance equations:

Element	**Balance equation**
Carbon	$4.8 + 0x = 1y + 0z$
Hydrogen	$8.4 + 0x = 0y + 2z$
Oxygen	$2.2 + 2x = 2y + 1z$

Simplifying each equation by eliminating all zero values and "1" coefficients:

Element	**Balance equation**
Carbon	$4.8 = y$
Hydrogen	$8.4 = 2z$
Oxygen	$2.2 + 2x = 2y + z$

We may tell from the first and second equations that $y = 4.8$ and $z = 4.2$, which then leads to a solution of $x = 5.8$ once the values for y and z have been inserted into the third equation. The final result is this balanced compositional equation for the oxidation of wastewater sludge:

$$C_{4.8}H_{8.4}O_{2.2} + 5.8O_2 \rightarrow 4.8CO_2 + 4.2H_2O$$

My own personal experience with the use of simultaneous linear equations as a tool for stoichiometry is that it is much faster (especially when balancing complex reaction equations) than trial-and-error, and relatively easy to set up once the general principles are understood.

3.8.3 Stoichiometric ratios

Regardless of the technique used to balance the equation for a chemical reaction, the most practical *purpose* of balancing the equation is to be able to relate the reactant and reaction product quantities to each other. For instance, we may wish to know how much oxygen will be required to completely combust with a given quantity of fuel, so that we will be able to engineer a burner system capable of handling the necessary flow rates of fuel and oxygen. Balancing the chemical reaction is just the first part of the solution. Once we have a balanced equation, we need to consider the ratios of the substances to each other.

For example, let us consider the balanced (stoichiometric) chemical equation for the combustion of ethane fuel with pure oxygen:

$$2C_2H_6 + 7O_2 \rightarrow 6H_2O + 4CO_2$$

From the balanced chemical equation we can see that for every 2 molecules of ethane, we will need 7 molecules of oxygen gas to completely combust, producing 6 molecules of water vapor and 4 molecules of carbon dioxide gas. The numerical multipliers preceding each molecule in the balanced equation tell us the *molar ratios* of those substances to each other. For oxygen to ethane the ratio is 7:2, for water to ethane the ratio is 6:2 (or 3:1), for carbon dioxide to water the ratio is 4:6 (2:3), etc. If for some reason we needed to calculate the number of moles of CO_2 produced after burning 80 moles of ethane, we could easily calculate that by multiplying the 80 moles of ethane by the 2:4 (1:2) ethane-to-carbon dioxide ratio to arrive at a figure of 160 moles of CO_2. If we wished, we could even solve this using the same method of *unity fractions* we commonly apply in unit-conversion problems, writing the carbon dioxide-to-ethane ratio as a fraction of two equivalent quantities:

$$\left(\frac{80 \text{ mol ethane}}{1}\right)\left(\frac{4 \text{ molecules carbon dioxide}}{2 \text{ molecules ethane}}\right) = 160 \text{ mol carbon dioxide}$$

If any substances involved in the reaction happen to be gases at nearly the same pressure and temperature[16], the molar ratios defined by the balanced equation will similarly define the *volumetric ratios* for those substances. For example, knowing our ideal oxygen-to-ethane molar ratio is 7:2 tells us that the volumetric flow rate of oxygen to ethane should also be (approximately) 7:2, assuming both the oxygen and ethane are gases flowing through their respective pipes at the same pressure and at the same temperature. Recall that the Ideal Gas Law ($PV = nRT$) is approximately true for *any* gas far from its critical phase-change values. So long as pressure (P) and temperature (T) are the same for both gases, each gas's volume (V) will be directly proportional to its molar quantity (n), since R is a constant. This means any molar ratio ($\frac{n_1}{n_2}$) for two gases under identical pressure and temperature conditions will be equal to the corresponding volumetric ratio ($\frac{V_1}{V_2}$) for those gases.

[16] These assumptions are critically important to equating volumetric ratios with molar ratios. First, the compared substances must both be *gases*: the volume of one mole of steam is huge compared to the volume of one mole of liquid water. Next, we cannot assume temperatures and pressures will be the same after a reaction as before. This is especially true for our example here, where ethane and oxygen are *burning* to produce water vapor and carbon dioxide: clearly, the reaction products will be at a greater temperature than the reactants!

It is important to understand that these molar ratios are not the same as the *mass* ratios for the reactants and products, simply because the different substances do not all have the same mass per mole.

If we regard each of the multipliers in the balanced equation as a precise molar quantity (i.e. exactly 2 moles of ethane, 7 moles of oxygen, etc.) and calculate the mass of the reactants, we will find this value precisely equal to the total mass of the reaction products because the Law of Mass Conservation holds true for this (and all other) chemical reactions:

$$2C_2H_6 = 2[(12)(2) + (1)(6)] = 60 \text{ grams}$$

$$7O_2 = 7[(16)(2)] = 224 \text{ grams}$$

$$6H_2O = 6[(1)(2) + 16] = 108 \text{ grams}$$

$$4CO_2 = 4[12 + (16)(2)] = 176 \text{ grams}$$

Calculating mass based on 2 moles of ethane, we have a total reactant mass of 284 grams (60 grams ethane plus 224 grams oxygen), and a total product mass of 284 grams as well (108 grams water plus 176 grams carbon dioxide gas). We may write the mass ratios for this chemical reaction as such:

$$\text{(ethane)} : \text{(oxygen)} : \text{(water)} : \text{(carbon dioxide)}$$

$$60 : 224 : 108 : 176$$

If for some reason we needed to calculate the mass of one of these substances in relation to the other for this reaction, we could easily do so using the appropriate mass ratios. For example, assume we were installing a pair of mass flowmeters to measure the mass flow rates of ethane and pure oxygen gas flowing into the combustion chamber of some industrial process. Supposing the ethane flowmeter had a calibrated range of 0 to 20 kg/min, what range should the oxygen's mass flowmeter be calibrated to in order to match in perfect stoichiometric ratio (so that when one flowmeter is at the top of its range, the other flowmeter should be also)?

The answer to this question is easy to calculate, knowing the required mass ratio of oxygen to ethane for this chemical reaction:

$$\left(\frac{20 \text{ kg ethane}}{1}\right)\left(\frac{224 \text{ g oxygen}}{60 \text{ g ethane}}\right) = 74.67 \text{ kg oxygen}$$

Therefore, the oxygen mass flowmeter should have a calibrated range of 0 to 74.67 kg/min. Note how the unit of mass used in the initial quantity (20 *kilograms* ethane) does not have to match the mass units used in our unity fraction (grams). We could have just as easily calculated the number of *pounds* per minute of oxygen given pounds per minute of ethane, since the mass ratio (like all ratios) is a unitless quantity[17].

[17] Looking at the unity-fraction problem, we see that "grams" (g) will cancel from top and bottom of the unity fraction, and "ethane" will cancel from the given quantity and from the bottom of the unity fraction. This leaves "kilograms" (kg) from the given quantity and "oxygen" from the top of the unity fraction as the only units remaining after cancellation, giving us the proper units for our answer: *kilograms of oxygen*.

3.9 Energy in chemical reactions

A chemical reaction resulting in a net release of energy is called *exothermic*. Conversely, a chemical reaction requiring a net input of energy to occur is called *endothermic*. The relationship between chemical reactions and energy exchange is correlated with the breaking or making of chemical bonds. Atoms bonded together represent a lower state of total energy than those same atoms existing separately, all other factors being equal. Thus, when separate atoms join together to form a molecule, they go from a high state of energy to a low state of energy, releasing the difference in energy in some form (heat, light, etc.). Conversely, an input of energy is required to break that chemical bond and force the atoms to separate.

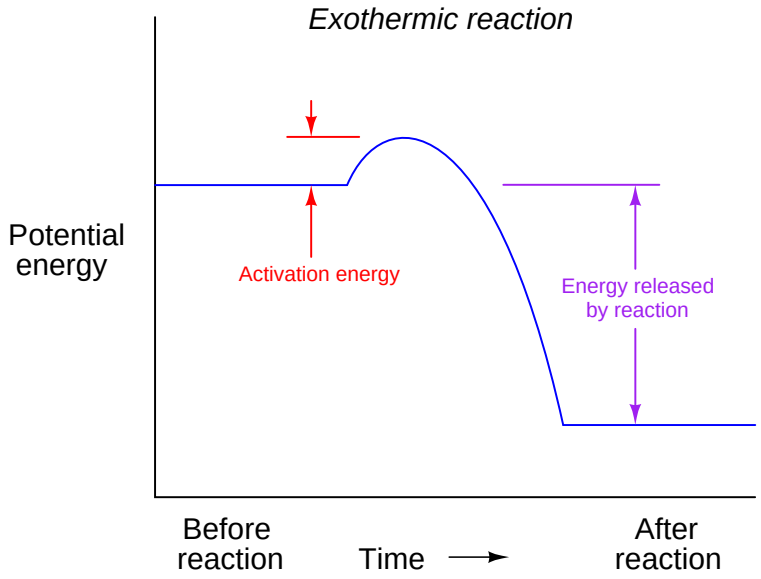
An example of this is the strong bond between two atoms of hydrogen (H) and one atom of oxygen (O), to form water (H_2O). When hydrogen and oxygen atoms bond together to form water, they release energy. This, by definition, is an exothermic reaction, but we know it better as *combustion*: hydrogen is flammable in the presence of oxygen.

A reversal of this reaction occurs when water is subjected to an electrical current, breaking water molecules up into hydrogen and oxygen gas molecules. This process of forced separation requires a substantial input of energy to accomplish, which by definition makes it an *endothermic* reaction. Specifically, the use of electricity to cause a chemical reaction is called *electrolysis*.

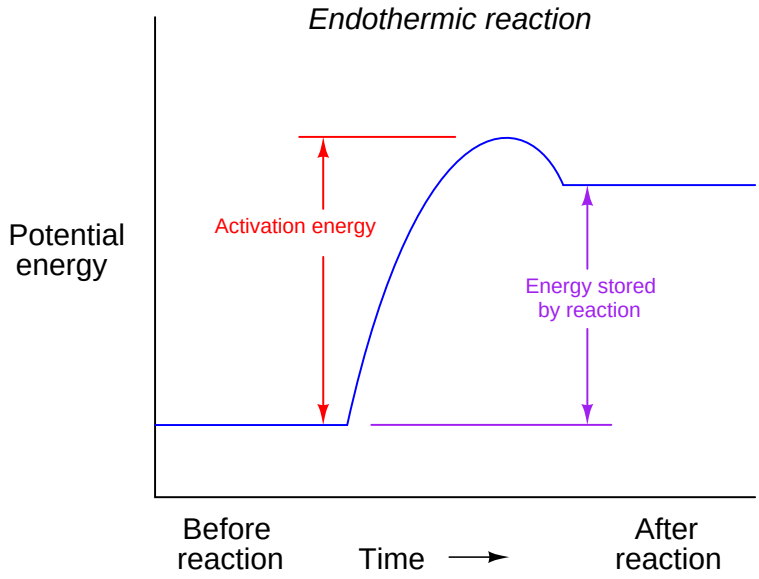
Energy storage and release is the purpose of the so-called "hydrogen economy" where hydrogen is a medium of energy distribution. The reasoning behind a hydrogen economy is that different sources of energy will be used to separate hydrogen from oxygen in water, then that hydrogen will be transported to points of use and consumed as a fuel, releasing energy. All the energy released by the hydrogen at the point of use comes from the energy sources tapped to separate the hydrogen from oxygen in water. Thus, the purpose of hydrogen in a hydrogen economy is to function as an energy storage and transport medium. The fundamental principle at work here is the energy stored in chemical bonds: invested in the separation of hydrogen from oxygen, and later returned in the re-combination of hydrogen and oxygen back into water.

The fact that hydrogen and oxygen as separate gases possess potential energy does not mean they are guaranteed to spontaneously combust when brought together. By analogy, just because rocks sitting on a hillside possess potential energy (by virtue of being elevated above the hill's base) does not means all rocks in the world spontaneously roll downhill. Some rocks need a push to get started because they are caught on a ledge or resting in a hole. Likewise, many exothermic reactions require an initial investment of energy before they can proceed. In the case of hydrogen and oxygen, what is generally needed is a spark to initiate the reaction. This initial requirement of input energy is called the *activation energy* of the reaction.

Activation energy may be shown in graphical form. For an exothermic reaction, it appears as a "hill" that must be climbed before the total energy can fall to a lower (than original) level:

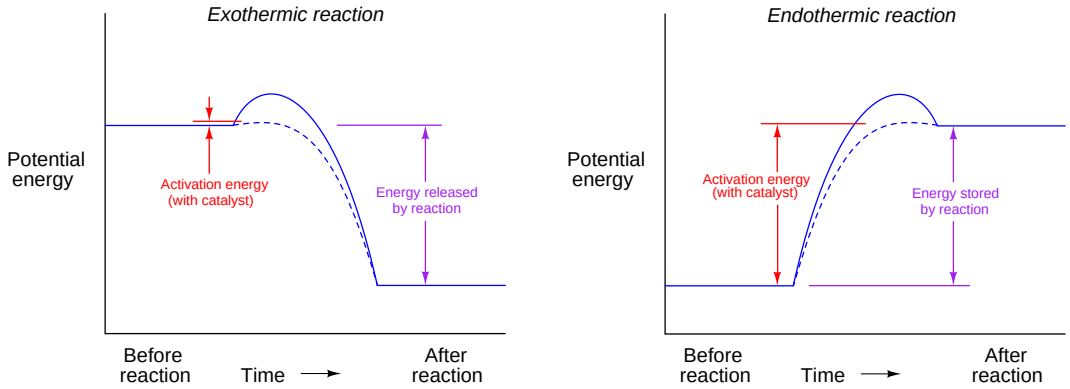


For an endothermic reaction, activation energy is much greater, a part of which never returns but is stored in the reaction products as potential energy:



A *catalyst* is a substance that works to minimize activation energy in a chemical reaction without being altered by the reaction itself. Catalysts are popularly used in industry to accelerate both exothermic and endothermic reactions, reducing the gross amount of energy that must be initially input to a process to make a reaction occur. A common example of a catalyst is the *catalytic converter* installed in the exhaust pipe of an automobile engine, helping to reduce oxidize unburnt fuel molecules and certain combustion products such as carbon monoxide (CO) to compounds which are not as polluting. Without a catalytic converter, the exhaust gas temperature is not hot enough to overcome the activation energy of these reactions, and so they will not occur (at least not at the rate necessary to make a significant difference). The presence of the catalyst allows the reactions to take place at standard exhaust temperatures.

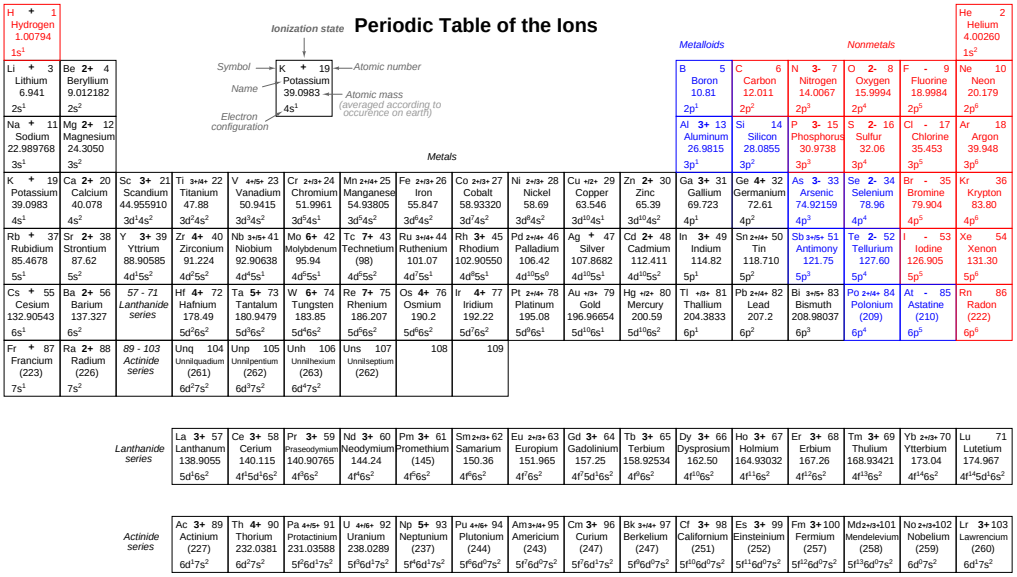
The effect of a catalyst on activation energy may be shown by the following graphs, the dashed-line curve showing the energy progression with a catalyst and the solid-line curve showing the reaction progressing without the benefit of a catalyst:



It should be noted that the presence of a catalyst has absolutely no effect on the *net* energy loss or gain resulting from a chemical reaction. With or without a catalyst, the difference in potential energy before and after a reaction will be the same. The only difference a catalyst makes to a chemical reaction is how much energy must be *initially invested* to spark the reaction.

To use the example of hydrogen and oxygen gas once again, the presence of a catalyst does not cause the combustion of hydrogen and oxygen to release more energy. All the catalyst does is make it easier for the combustion to begin.

3.10 Periodic table of the ions



3.11 Ions in liquid solutions

Many liquid substances undergo a process whereby their constituent molecules split into positively and negatively charged ion pairs, the positively-charge ion called a *cation* and the negatively-charged ion called an *anion*[18]. Liquid *ionic* compounds split into ions completely or nearly completely, while only a small percentage of the molecules in a liquid *covalent* compound split into ions. The process of neutral molecules separating into ion pairs is called *dissociation* when it happens to ionic compounds, and *ionization* when it happens to covalent compounds.

Molten salt (NaCl) is an example of the former, while pure water (H_2O) is an example of the latter. The large presence of ions in molten salt explains why it is a good conductor of electricity, while the comparative lack of ions in pure water explains why it is often considered an insulator. In fact, the electrical conductivity of a liquid substance is the definitive test of whether it is an ionic or a covalent ("molecular") substance.

Pure water ionizes into positive hydrogen ions[19] (H^+) and negative hydroxyl ions (OH^-). At room temperature, the concentration of hydrogen and hydroxyl ions in a sample of pure water is quite small: a molarity of 10^{-7} M (moles per liter) each.

Given the fact that pure water has a mass of 1 kilogram (1000 grams) per liter, and one mole of pure water has a mass of 18 grams, we must conclude that there are approximately 55.56 moles of water molecules in one liter (55.56 M). If only 10^{-7} moles of those molecules ionize at room temperature, that represents an extremely small percentage of the total:

$$\frac{10^{-7}\ M}{55.56\ M} = 0.0000000018 = 0.00000018\% = 0.0018 \text{ ppm (parts per million)}$$

It is not difficult to see why pure water is such a poor conductor of electricity. With so few ions available to act as charge carriers, pure water is practically an insulator. The vast majority of water molecules remain un-ionized and therefore cannot transport electric charges from one point to another.

The molarity of both hydrogen and hydroxyl ions in a pure water sample increases with increasing temperature. For example, at 60^o C, the molarity of hydrogen and hydroxyl ions increases to 3.1×10^{-7} M, which is still only 0.0056 parts per million, but definitely larger than the concentration at room temperature (25^o C).

The electrical conductivity of water may be greatly enhanced by dissolving an ionic compound in it, such as salt. When dissolved, the salt molecules (NaCl) immediately dissociate into sodium cations (Na^+) and chlorine anions (Cl^-), becoming available as charge carriers for an electric current. In industry, we may exploit this relationship between electrical conductivity and ionic dissociation to detect the presence of ionic compounds in otherwise pure water.

[18] These names have their origin in the terms used to classify positive and negative electrodes immersed in a liquid solution. The positive electrode is called the "anode" while the negative electrode is called the "cathode." An *anion* is an ion attracted to the anode. A *cation* is an ion attracted to the cathode. Since opposite electrical charges tend to attract, this means "anions" are negatively charged and "cations" are positively charged.

[19] Actually, the more common form of positive ion in water is *hydronium*: H_3O^+, but we often simply refer to the positive half of an ionized water molecule as hydrogen (H^+).

3.12 pH

Hydrogen ion activity in aqueous (water-solvent) solutions is a very important parameter for a wide variety of industrial processes. A number of reactions important to chemical processing are inhibited or significantly slowed if the hydrogen ion activity of a solution does not fall within a narrow range. Some additives used in water treatment processes (e.g. flocculants) will fail to function efficiently if the hydrogen ion activity in the water is not kept within a certain range. Alcohol and other fermentation processes strongly depend on tight control of hydrogen ion activity, as an incorrect level of ion activity will not only slow production, but may also spoil the product. Hydrogen ions are always measured on a logarithmic scale, and referred to as *pH*.

Free hydrogen ions (H^+) are rare in a liquid solution, and are more often found attached to whole water molecules to form a positive ion called *hydronium* (H_3O^+). However, process control professionals usually refer to these positive ions simply as "hydrogen" even though the truth is a bit more complicated.

pH is mathematically defined as the negative common logarithm of hydrogen ion activity in a solution. Hydrogen ion activity is expressed as a molarity (number of moles of active ions per liter of solution), with "pH" being the unit of measurement for the logarithmic result:

$$\text{pH} = -\log[H^+]$$

For example, an aqueous solution with an active hydrogen concentration of 0.00044 M has a pH value of 3.36 pH.

Water is a covalent compound, and so there is little separation of water molecules in liquid form. Most of the molecules in a sample of pure water remain as whole molecules (H_2O) while a very small percentage ionize into positive hydrogen ions (H^+) and negative hydroxyl ions (OH^-). The mathematical product of hydrogen and hydroxyl ion molarity in water is known as the *ionization constant* (K_w), and its value varies with temperature:

$$K_w = [H^+] \times [OH^-]$$

At 25 degrees Celsius (room temperature), the value of K_w is very nearly equal to 1.0×10^{-14}. Since each one of the water molecules that does ionize in this absolutely pure water sample separates into exactly one hydrogen ion (H^+) and one hydroxyl ion (OH^-), the molarities of hydrogen and hydroxyl ions must be equal to each other. The equality between hydrogen and hydroxyl ions in a pure water sample means that pure water is *neutral*, and that the molarity of hydrogen ions is equal to the square root of K_w:

$$[H^+] = \sqrt{K_w} = \sqrt{1.0 \times 10^{-14}} = 1.0 \times 10^{-7}\ M$$

Since we know pH is defined as the negative logarithm of hydrogen ion activity, and we can be assured all hydrogen ions present in the solution will be "active" since there are no other positive ions to interfere with them, the pH value for water at 25 degrees Celsius is:

$$\text{pH of pure water at } 25^o\text{C} = -\log(1.0 \times 10^{-7}\ M) = 7.0 \text{ pH}$$

As the temperature of a pure water sample changes, the ionization constant changes as well. Increasing temperature causes more of the water molecules to ionize into H^+ and OH^- ions, resulting in a larger K_w value. The following table shows K_w values for pure water at different temperatures:

Temperature	K_W
0^o C	1.139×10^{-15}
5^o C	1.846×10^{-15}
10^o C	2.920×10^{-15}
15^o C	4.505×10^{-15}
20^o C	6.809×10^{-15}
25^o C	1.008×10^{-14}
30^o C	1.469×10^{-14}
35^o C	2.089×10^{-14}
40^o C	2.919×10^{-14}
45^o C	4.018×10^{-14}
50^o C	5.474×10^{-14}
55^o C	7.296×10^{-14}
60^o C	9.614×10^{-14}

This means that while any pure water sample is *neutral* (an equal number of positive hydrogen ions and negative hydroxyl ions) at any temperature, the pH value of pure water actually changes with temperature, and is only equal to 7.0 pH at one particular ("standard") temperature: 25^o C. Based on the K_w values shown in the table, pure water will be 6.51 pH at 60^o C and 7.47 pH at freezing.

If we add an electrolyte to a sample of pure water, (at least some of) the molecules of that electrolyte will separate into positive and negative ions. If the positive ion of the electrolyte happens to be a hydrogen ion (H^+), we call that electrolyte an *acid*. If the negative ion of the electrolyte happens to be a hydroxyl ion (OH^-), we call that electrolyte a *caustic*, or *alkaline*, or *base*. Some common acidic and alkaline substances are listed here, showing their respective positive and negative ions in solution:

Sulfuric acid is an *acid* (produces H^+ in solution)
$H_2SO_4 \rightarrow 2H^+ + {SO_4}^{2-}$

Nitric acid is an *acid* (produces H^+ in solution)
$HNO_3 \rightarrow H^+ + {NO_3}^-$

Hydrocyanic acid is an *acid* (produces H^+ in solution)
$HCN \rightarrow H^+ + CN^-$

Hydrofluoric acid is an *acid* (produces H^+ in solution)
$HF \rightarrow H^+ + F^-$

Lithium hydroxide is a *caustic* (produces OH^- in solution)
$LiOH \rightarrow Li^+ + OH^-$

Potassium hydroxide is a *caustic* (produces OH^- in solution)
$KOH \rightarrow K^+ + OH^-$

Sodium hydroxide is a *caustic* (produces OH^- in solution)
$NaOH \rightarrow Na^+ + OH^-$

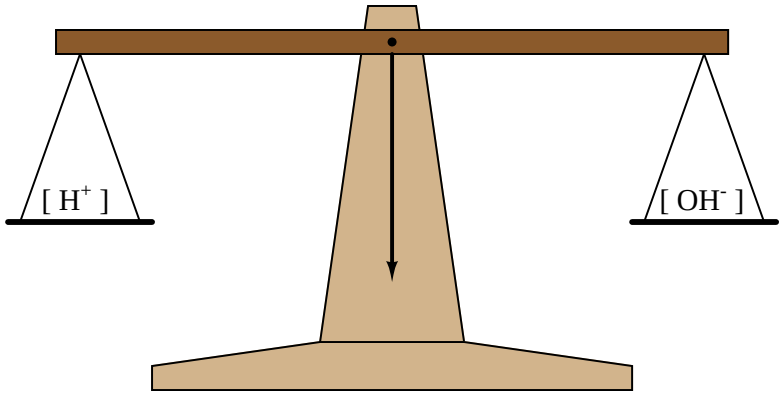
Calcium hydroxide is a *caustic* (produces OH^- in solution)
$Ca(OH)_2 \rightarrow Ca^{2+} + 2OH^-$

When an acid substance is added to water, some of the acid molecules dissociate into positive hydrogen ions (H^+) and negative ions (the type of negative ions depending on what type of acid it is). This increases the molarity of hydrogen ions (the number of moles of H^+ ions per liter of solution). The addition of hydrogen ions to the solution also decreases the molarity of hydroxyl ions (the number of moles of OH^- ions per liter of solution) because some of the water's OH^- ions combine with the acid's H^+ ions to form deionized water molecules (H_2O).

If an alkaline substance (otherwise known as a *caustic*, or a *base*) is added to water, some of the alkaline molecules dissociate into negative hydroxyl ions (OH^-) and positive ions (the type of positive ions depending on what type of alkaline it is). This increases the molarity of OH^- ions in the solution, as well as decreases the molarity of hydrogen ions (again, because some of the caustic's OH^- ions combine with the water's H^+ ions to form deionized water molecules, H_2O).

The result of this complementary effect (increasing one type of water ion, decreasing the other) keeps the overall ionization constant relatively constant, at least for dilute solutions. In other words, the addition of an acid or a caustic may change $[H^+]$, but it has little effect on K_w.

A simple way to envision this effect is to think of a laboratory balance scale, balancing the number of hydrogen ions in a solution against the number of hydroxyl ions in the same solution:



When the solution is pure water, this imaginary scale is balanced (neutral), with $[H^+] = [OH^-]$. Adding an acid to the solution tips the scale one way, while adding a caustic to the solution tips it the other way[20].

If an electrolyte has no effect on the hydrogen and hydroxyl ion activity of an aqueous solution, we call it a *salt*. The following is a list of some common salts, showing their respective ions in solution:

Potassium chloride is a *salt* (produces neither H^+ nor OH^- nor O^{2-} in solution)
$KCl \rightarrow K^+ + Cl^-$

Sodium chloride is a *salt* (produces neither H^+ nor OH^- nor O^{2-} in solution)
$NaCl \rightarrow Na^+ + Cl^-$

Zinc sulfate is a *salt* (produces neither H^+ nor OH^- nor O^{2-} in solution)
$ZnSO_4 \rightarrow Zn^+ + SO_4{}^-$

The addition of a salt to an aqueous solution should have no effect on pH, because the ions created neither add to nor take away from the hydrogen ion activity[21].

Acids and caustics tend to neutralize one another, the hydrogen ions liberated by the acid combining (and canceling) with the hydroxyl ions liberated by the caustic. This process is called *pH neutralization*, and it is used extensively to adjust the pH value of solutions. If a solution is too

[20] It should be noted that the solution never becomes *electrically* imbalanced with the addition of an acid or caustic. It is merely the balance of hydrogen to hydroxyl ions we are referring to here. The net electrical charge for the solution should still be zero after the addition of an acid or caustic, because while the balance of hydrogen to hydroxyl ions does change, that electrical charge imbalance is made up by the other ions resulting from the addition of the electrolyte (anions for acids, cations for caustics). The end result is still one negative ion for every positive ion (equal and opposite charge numbers) in the solution no matter what substance(s) we dissolve into it.

[21] Exceptions do exist for strong concentrations, where hydrogen ions may be present in solution yet unable to react because of being "crowded out" by other ions in the solution.

acidic, just add caustic to raise its pH value. If a solution is too alkaline, just add acid to lower its pH value.

The result of a perfectly balanced mix of acid and caustic is deionized water (H_2O) and a salt formed by the combining of the acid's and caustic's *other* ions. For instance, when hydrochloric acid (HCl) and potassium hydroxide (KOH) neutralize one another, the result is water (H_2O) and potassium chloride (KCl), a salt. This production of salt is a necessary side-effect of pH neutralization, which may require addressing in later stages of solution processing. Such neutralizations are exothermic, owing to the decreased energy states of the hydrogen and hydroxyl ions after combination. Mixing of pure acids and caustics together without the presence of substantial quantities of water (as a solvent) is often violently exothermic, presenting a significant safety hazard to anyone near the reaction.

References

"Fundamental Physical Constants – Extensive Listing", from `http://physics.nist.gov/constants`, National Institute of Standards and Technology (NIST), 2006.

Giancoli, Douglas C., *Physics for Scientists & Engineers*, Third Edition, Prentice Hall, Upper Saddle River, NJ, 2000.

Haug, Roger Tim, *The Practical Handbook of Compost Engineering*, CRC Press, LLC, Boca Raton, FL, 1993.

Mills, Ian; Cvitaš, Tomislav; Homann, Klaus; Kallay, Nikola; Kuchitsu, Kozo, *Quantities, Units and Symbols in Physical Chemistry* (the "Green Book"), Second Edition, International Union of Pure and Applied Chemistry (IUPAC), Blackwell Science Ltd., Oxford, England, 1993.

"NIOSH Pocket Guide to Chemical Hazards", DHHS (NIOSH) publication # 2005-149, Department of Health and Human Services (DHHS), Centers for Disease Control and Prevention (CDC), National Institute for Occupational Safety and Health (NIOSH), Cincinnati, OH, September 2005.

Pauling, Linus, *General Chemistry*, Dover Publications, Inc., Mineola, NY, 1988.

Rosman, K.J.R. and Taylor, P.D.P, *Isotopic Compositions of the Elements 1997*, International Union of Pure and Applied Chemistry (IUPAC), 1997.

Scerri, Eric R., *How Good Is the Quantum Mechanical Explanation of the Periodic System?*, Journal of Chemical Education, Volume 75, Number 11, pages 1384-1385, 1998.

Weast, Robert C.; Astel, Melvin J.; and Beyer, William H., *CRC Handbook of Chemistry and Physics*, 64th Edition, CRC Press, Inc., Boca Raton, FL, 1984.

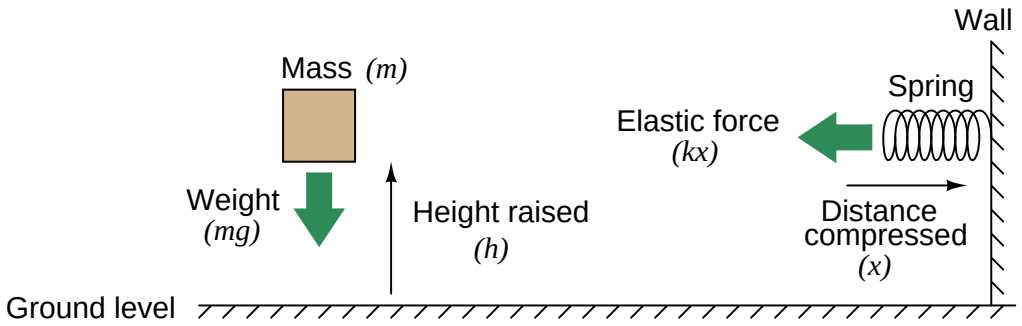
Whitten, Kenneth W.; Gailey, Kenneth D.; and Davis, Raymond E., *General Chemistry*, Third Edition, Saunders College Publishing, Philadelphia, PA, 1988.

Chapter 4

DC electricity

4.1 Electrical voltage

Voltage is the amount of *specific potential energy* available between two points in an electric circuit. Potential energy is energy that is potentially available to do work. Looking at this from a classical physics perspective, potential energy is what we accumulate when we lift a weight above ground level, or when we compress a spring:



In either case, potential energy is calculated by the work done in exerting a force over a parallel distance. In the case of the weight, potential energy (E_p) is the simple product of weight (gravity g acting on the mass m) and height (h):

$$E_p = mgh$$

For the spring, things are a bit more complex. The force exerted by the spring against the compressing motion increases with compression ($F = kx$, where k is the elastic constant of the spring). It does not remain steady as the force of weight does for the lifted mass. Therefore, the potential energy equation is nonlinear:

$$E_p = \frac{1}{2}kx^2$$

Releasing the potential energy stored in these mechanical systems is as simple as dropping the mass, or letting go of the spring. The potential energy will return to the original condition (zero) when the objects are at rest in their original positions. If either the mass or the spring were attached to a machine to harness the return-motion, that stored potential energy could be used to do useful tasks.

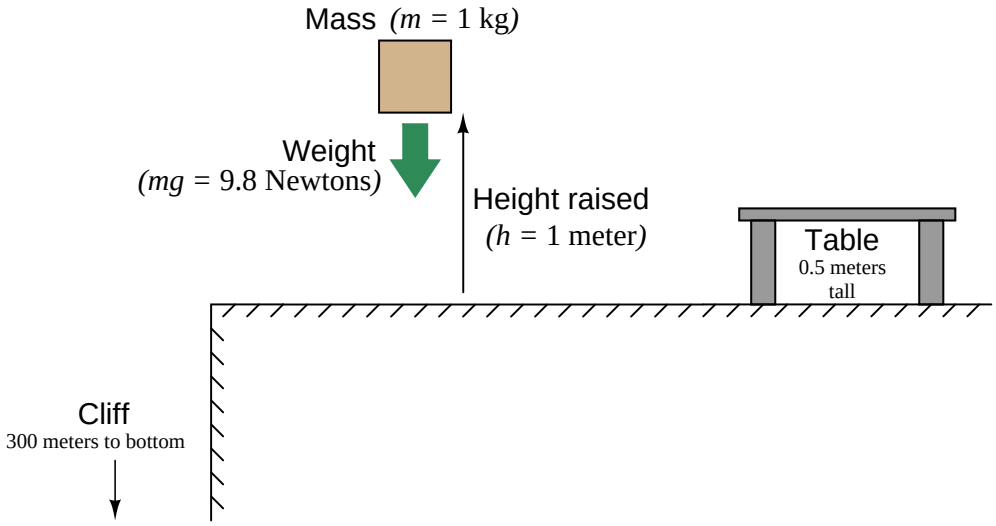
Potential energy may be similarly defined and quantified for *any* situation where we exert a force over a parallel distance, regardless of where that force or the motivating distance comes from. For instance, the static cling you experience when you pull a cotton sock out of a dryer is an example of a force. By pulling that sock away from another article of clothing, you are doing *work*, and storing *potential energy* in the tension between that sock and the rest of the clothing. In a similar manner, that stored energy could be released to do useful tasks if we placed the sock in some kind of machine that harnessed the return motion as the sock went back to its original place on the pile of laundry inside the dryer.

If we make use of non-mechanical means to move electric charge from one location to another, the result is no different. Moving attracting charges apart from one another means doing *work* (a force exerted over a parallel distance) and storing potential energy in that physical tension. When we use chemical reactions to move electrons from one metal plate to another in a solution, or when we spin a generator and electro-magnetically motivate electrons to seek other locations, we impart potential energy to those electrons. We could express this potential energy in the same unit as we do for mechanical systems (the *Joule*). However, it is actually more useful to express the potential energy in an electric system in terms of how many joules are available per a specific quantity of electric charge (a certain number of electrons). This measure of *specific* potential energy is simply called *electric potential* or *voltage*, and we measure it in units of *Volts*, in honor of the Italian physicist Alessandro Volta, inventor of the first electrochemical battery.

$$1 \text{ Volt} = \frac{1 \text{ Joule of potential energy}}{1 \text{ Coulomb of electric charge}}$$

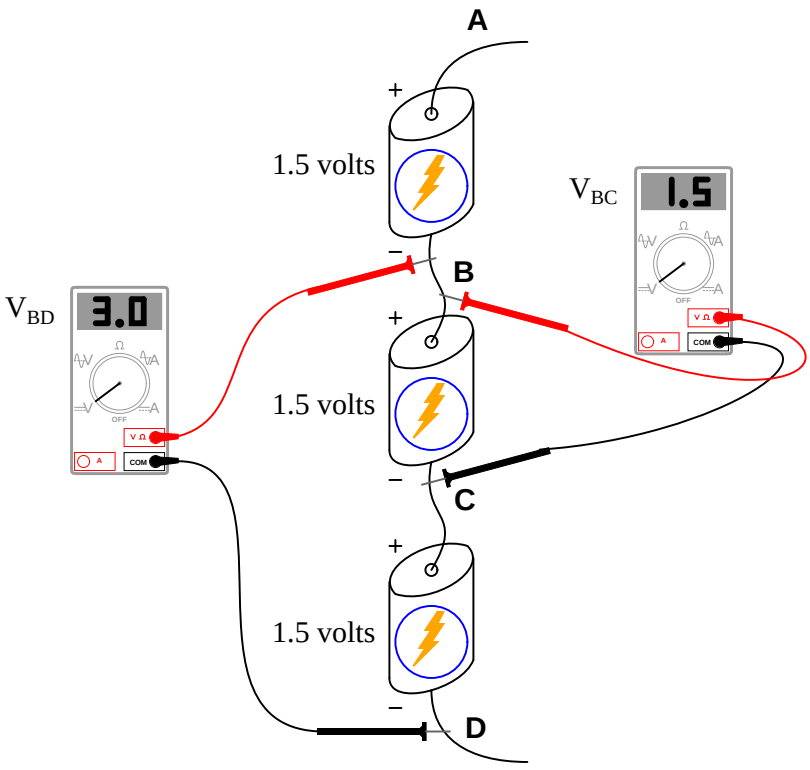
In other words, if we forced 1 Coulomb's worth of electrons (6.24×10^{18} of them, to be exact) away from a positively-charged place, and did one Joule's worth of work in the process, we would have generated one Volt of electric potential.

Electric potential (voltage) and potential energy share a common, yet confusing property: both quantities are fundamentally *relative* between two physical locations. There is really no such thing as specifying a quantity of potential energy at a single location. The amount of potential energy in any system is always relative between two different points. If I lift a mass off the ground, I can specify its potential energy, *but only in relation to its former position on the ground*. The amount of energy that mass is potentially capable of releasing by free-fall depends on how far it could possibly fall. To illustrate, imagine lifting a 1 kilogram mass 1 meter off the ground. That 1-kilo mass weighs 9.8 Newtons on Earth, and the distance lifted was 1 meter, so the potential energy stored in the mass is 9.8 joules, right? Consider the following scenario:



If we drop the mass over the spot we first lifted it from, it will release all the potential energy we invested in it: 9.8 joules. But what if we carry it over to the table and release it there? Since now it can only fall half a meter, it will only release 4.9 joules in the process. How much potential energy did the mass have while suspended above that table? What if we carry it over to the edge of the cliff and release it there? Falling 301 meters, it will release 2.95 kilojoules (kJ) of energy. How much potential energy did the mass have while suspended over the cliff?

As you can see, potential energy is a relative quantity. We must know the mass's position relative to its falling point before we can quantify its potential energy. Likewise, we must know an electric charge's position relative to its return point before we can quantify the voltage it has. Consider a series of batteries connected as shown:

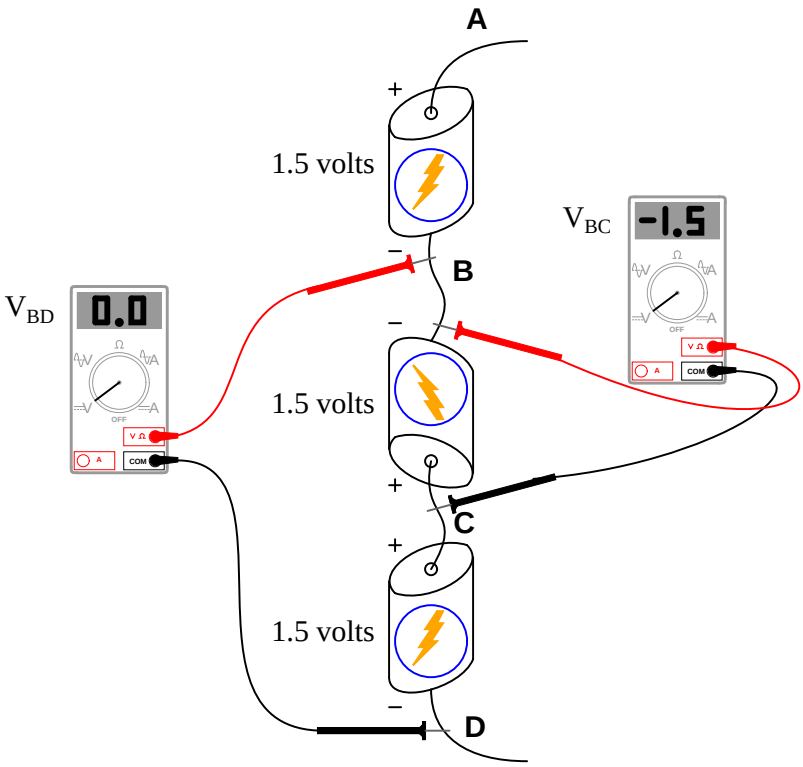


The voltage as measured between any two points directly across a single battery will be 1.5 volts:
$V_{AB} = 1.5$ volts
$V_{BC} = 1.5$ volts
$V_{CD} = 1.5$ volts

If, however, we span more than one battery with our voltmeter connections, our voltmeter will register more than 1.5 volts:
$V_{AC} = 3.0$ volts
$V_{BD} = 3.0$ volts
$V_{AD} = 4.5$ volts

There is no such thing as "voltage" at a single point in a circuit. The concept of voltage has meaning only *between* pairs of points in a circuit, just as the concept of potential energy for a mass has meaning only *between* two physical locations: where the mass is, and where it could potentially fall to.

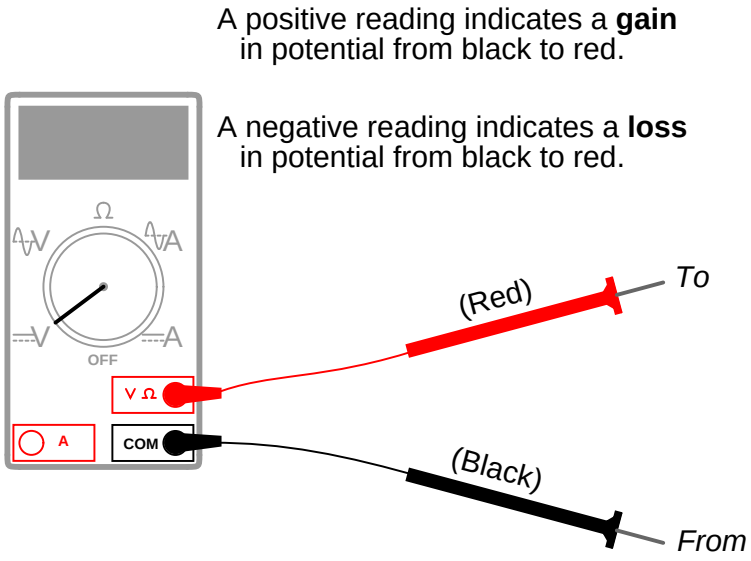
Things get interesting when we connect voltage sources in different configurations. Consider the following example, identical to the previous illustration except the middle battery has been reversed:



Note the "+" and "-" signs next to the ends of the batteries. These signs show the *polarity* of each battery's voltage. Also note how the two voltmeter readings are different from before. Here we see an example of *negative potential* with the middle battery connected in opposition to the other two batteries. While the top and bottom batteries are both "lifting" electric charges to greater potential (going from point **D** to point **A**), the middle battery is decreasing potential from point **C** to point **B**. It's like taking a step forward, then a step back, then another step forward. Or, perhaps more appropriately, like lifting a mass 1.5 meters up, then setting it down 1.5 meters, then lifting it 1.5 meters up again. The first and last steps accumulate potential energy, while the middle step releases potential energy.

This explains why it is important to install multiple batteries the same way into battery-powered devices such as radios and flashlights. The batteries' voltages are supposed to add to a make a larger total required by the device. If one or more batteries are placed backwards, potential will be lost instead of gained, and the device will not receive enough voltage.

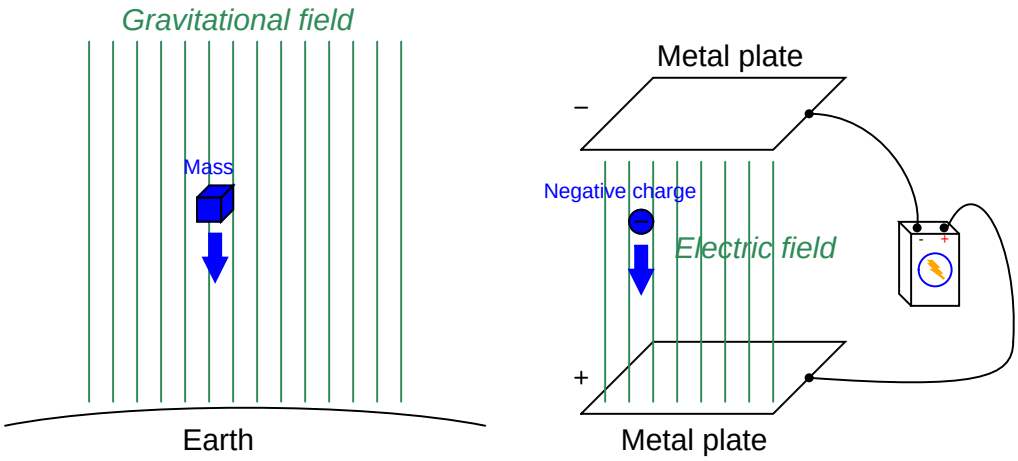
Here we must pay special attention to how we use our voltmeter, since polarity matters. All voltmeters are standardized with two colors for the test leads: red and black. To make sense of the voltmeter's indication, especially the positive or negative *sign* of the indication, we must understand what the red and black test lead colors mean:



Connecting these test leads to different points in a circuit will tell you whether there is potential gain or potential loss from one point (black) to the other point (red).

4.2 Electrical current

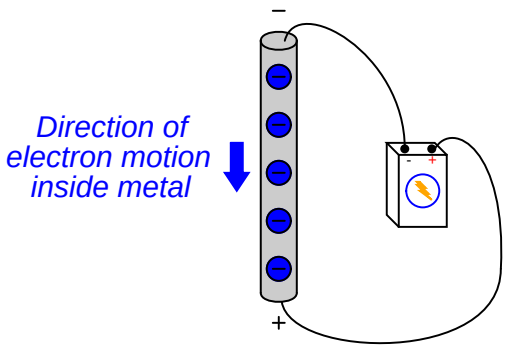
Current is the name we give to the motion of electric charges from a point of high potential to a point of low potential. All we need to form an electric current is a source of potential (voltage) and some electric charges that are free to move between the poles of that potential. For instance, if we connected a battery to two metal plates, we would create an electric field between those plates, analogous to a gravitational field except it only acts on electrically charged objects, while gravity acts on anything with mass. A free charge placed between those plates would "fall" toward one of the plates just as a mass would fall toward a larger mass:



An electric charge will "fall" in an electric field just as a mass will fall in a gravitational field.

Some substances, most notably metals, have very mobile electrons. That is, the outer (valence) electrons are very easily dislodged from the parent atoms to drift to and fro throughout the material. In fact, the electrons of metals are so free that physicists sometimes refer to the structure of a metal as atoms floating in a "sea of electrons". The electrons are almost fluid in their mobility throughout a solid metal object, and this property of metals may be exploited to form definite pathways for electric currents.

If the poles of a voltage source are joined by a continuous path of metal, the free electrons within that metal will drift toward the positive pole (electrons having a negative charge, opposite charges attracting one another):

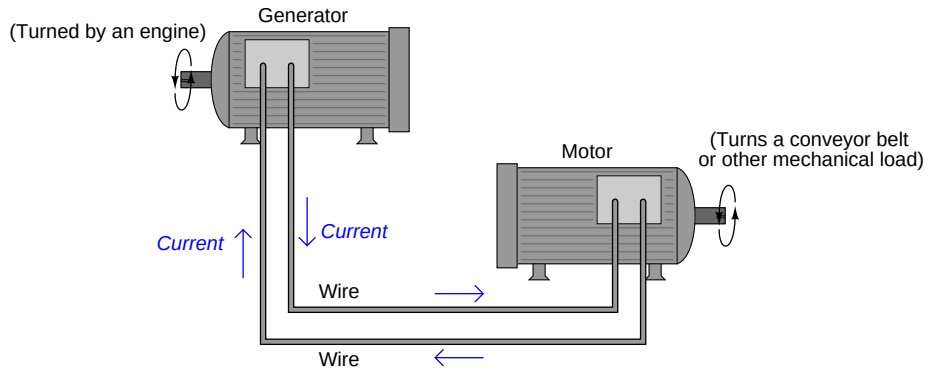


If the source of this voltage is continually replenished by chemical energy, mechanical energy, or some other form of energy, the free electrons will continually loop around this circular path. We call this unbroken path an *electric circuit*.

We typically measure the amount of current in a circuit by the unit of *amperes*, or *amps* for short (named in honor of the French physicist André Ampère. One ampere of current is equal to one coulomb of electric charge (6.24×10^{18} electrons) moving past a point in a circuit for every second of time.

Like masses falling toward a source of gravity, these electrons continually "fall" toward the positive pole of a voltage source. After arriving at that source, the energy imparted by that source "lifts" the electrons to a higher potential state where they once again "fall down" to the positive pole through the circuit.

Like rising and falling masses in a gravitational field, these electrons act as carriers of energy within the electric field of the circuit. This is very useful, as we can use them to convey energy from one place to another, using metal wires as conduits for this energy. This is the basic idea behind electric power systems: a source of power (a *generator*) is turned by some mechanical engine (windmill, water turbine, steam engine, etc.), creating an electric potential. This potential is then used to motivate free electrons inside the metal wires to drift in a common direction. The electron drift is conveyed in a circuit through long wires, where they can do useful work at a *load* device such as an electric motor, light bulb, or heater.



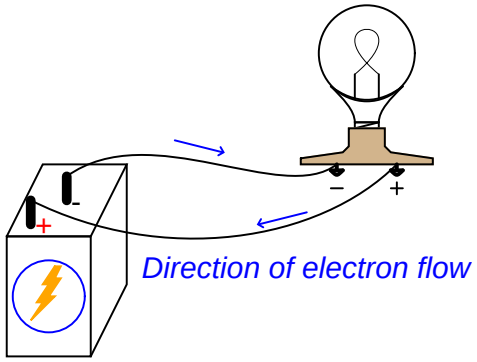
Given the proper metal alloys, the friction that electrons experience within the metal wires may be made very small, allowing nearly all the energy to be expended at the load (motor), with very little wasted along the path (wires). This makes electricity the most efficient means of energy transport known.

The electric currents common in electric power lines may range from hundreds to thousands of amperes. The currents conveyed through power receptacles in your home typically are no more than 15 or 20 amperes. The currents in the small battery-powered circuits you will build are even less: fractions of an ampere. For this reason, we commonly use the metric prefix *milli* (one one-thousandth) to express these small currents. For instance, 10 milliamperes is 0.010 amperes, and 500 milliamperes is one-half of an ampere.

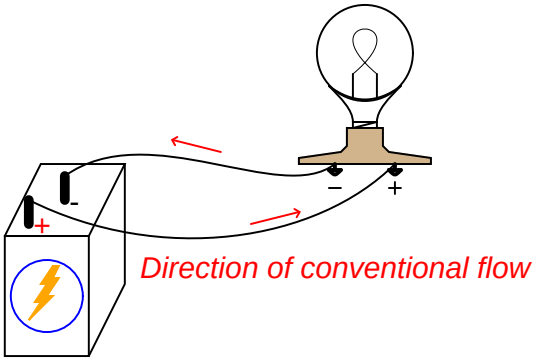
4.2.1 Electron versus conventional flow

When Benjamin Franklin advanced his single-fluid theory of electricity, he defined "positive" and "negative" as the surplus and deficiency of electric charge, respectively. These labels were largely arbitrary, as Mr. Franklin had no means of identifying the actual nature of electric charge carriers with the primitive test equipment and laboratory techniques of his day. As luck would have it, his hypothesis was precisely opposite of the truth for metallic conductors, where electrons are the dominant charge carrier.

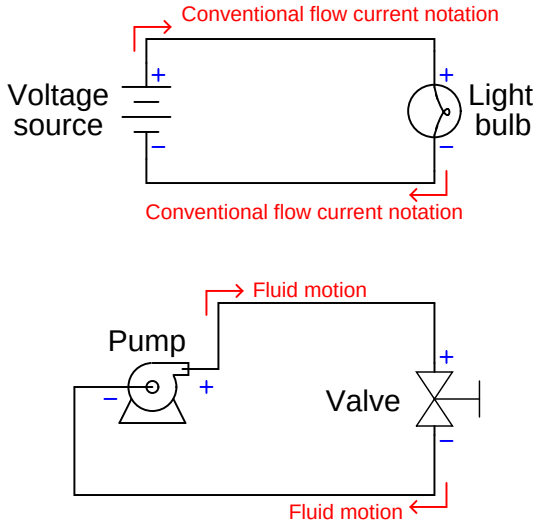
This means that in an electric circuit consisting of a battery and a light bulb, electrons slowly move from the negative side of the battery, through the metal wires, through the light bulb, and on to the positive side of the battery as such:



Unfortunately, scientists and engineers had grown accustomed to Franklin's false hypothesis long before the true nature of electric current in metallic conductors was discovered. Their preferred notation was to show electric current flowing from the positive pole of a source, through the load, returning to the negative pole of the source:

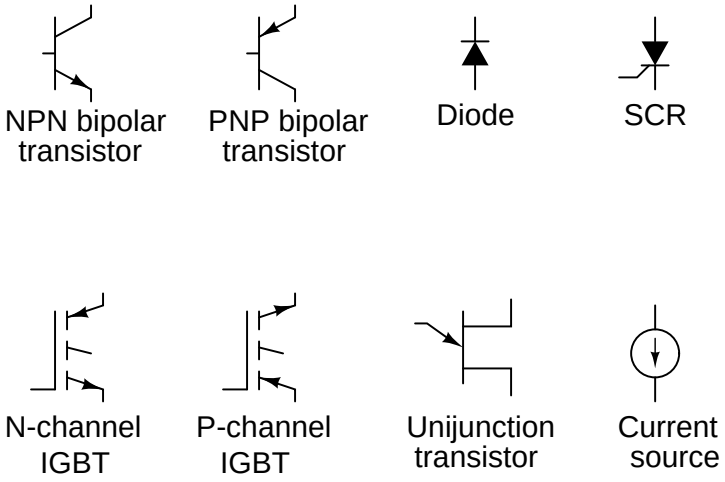


This relationship between voltage polarity marks and conventional flow current makes more intuitive sense than electron flow notation, because it is reminiscent of fluid pressure and flow direction:



If we take the "+" sign to represent *more* pressure and the "-" sign to represent *less* pressure, it makes perfect sense that fluid should move from the high-pressure (discharge) port of the pump through the hydraulic "circuit" and back to the low-pressure (suction) port of the pump. It also makes perfect sense that the upstream side of the valve (a fluid restriction) will have a greater pressure than the downstream side of the valve. In other words, conventional flow notation best honors Mr. Franklin's original intent of modeling current as though it were a fluid, even though he was later proven to be mistaken in the case of metallic conductors where electrons are the dominant charge carrier.

This convention was so well-established in the electrical engineering realm that it held sway despite the discovery of electrons. Engineers, who create the symbols used to represent the electronic devices they invent, consistently chose to draw arrows in the direction of conventional flow rather than electron flow. In each of the following symbols, the arrow heads point in the direction that *positive* charge carriers would move (opposite the direction that electrons actually move):



This stands in contrast to electronics technicians, who historically have been taught using electron flow notation. I remember sitting in a technical school classroom being told by my teacher to always imagine the electrons moving *against the arrows* of the devices, and wondering why it mattered.

It is truly a sad situation when the members of two branches within the same field do not agree on something as fundamental as the convention used to denote flow in diagrams. It is even worse when people within the field argue over which convention is best. So long as one is consistent with their convention and with their thinking, *it does not matter!* Many fine technologists may be found on either side of this "fence," and some are adept enough to switch between both without getting confused.

For what it's worth, I personally prefer conventional flow notation. The only objective arguments I have in favor of this preference are as follows:

- Conventional flow notation makes more intuitive sense to someone familiar with fluid systems (as all instrument technicians need to be!).

- Conventional flow notation matches all device arrows; no need to "go against the arrow" when tracing current in a schematic diagram.

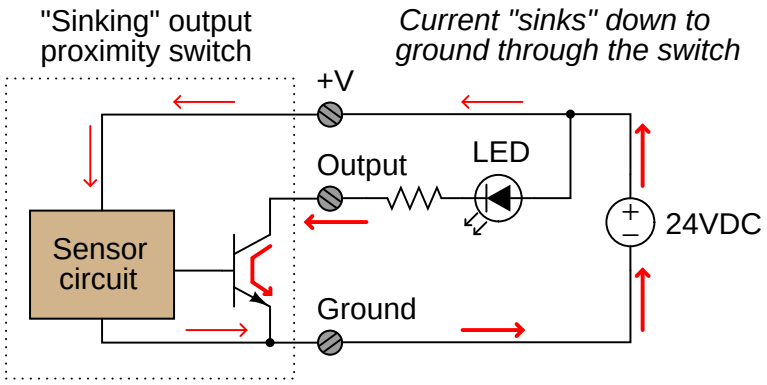
- Conventional flow notation is consistent with the "right-hand rule" for vector cross products (which are essential for understanding electromagnetics at advanced academic levels). The so-called "left-hand rule" taught to students learning electron flow notation is mathematically wrong, and must be un-learned if the student ever progresses to the engineering level in his or her studies.

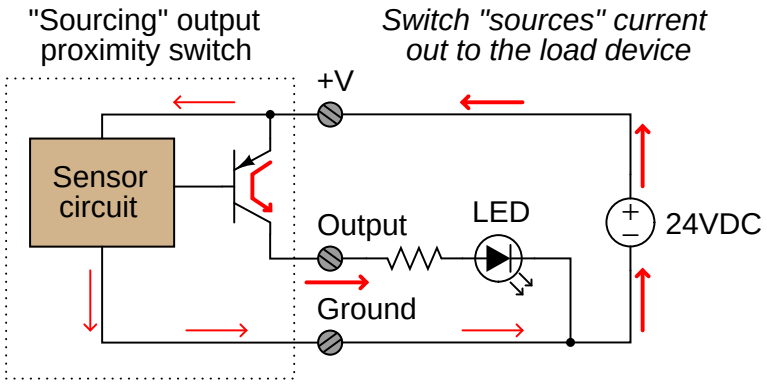
- Conventional flow notation is the *standard* for modern manufacturers' documentation (reference manuals, troubleshooting guides, datasheets, etc.)[1].
- Conventional flow notation makes sense of the descriptive terms *sourcing* and *sinking*.

This last point merits further investigation. The terms "sourcing" and "sinking" are often used in the study of digital electronics to describe the direction of current in a switching circuit. A circuit that "sources" current to a load is one where the direction of conventional flow points outward from the sourcing circuit to the load device.

[1] I have yet to read a document of any kind written by an equipment manufacturer that uses electron flow notation, and this is after scrutinizing literally hundreds of documents looking for this exact detail! For the record, though, most technical documents do not bother to draw a direction for current at all, leaving it to the imagination of the reader instead. It is only when a direction must be drawn that one sees a strong preference in industry for conventional flow notation.

For example, here are two schematic diagrams showing two different kinds of electronic proximity switch. The first switch *sinks* current in from the LED through its output terminal, through its transistor, and down to ground. The second switch *sources* current from the positive supply terminal through its transistor and out to the LED through its output terminal (note the direction of the thick arrow near the output screw terminal in each circuit):





These terms simply make no sense when viewed from the perspective of electron flow notation. If you were to actually trace the directions of the electrons, you would find that a device "sourcing" current has electrons flowing *into* its connection terminal, while a device "sinking" current sends electrons *out* to another device where they travel (up) to a point of more positive potential.

In fact, the association between conventional flow notation and sourcing/sinking descriptions is so firm that I have yet to see a professionally published textbook on digital circuits that uses electron flow[2]. This is true even for textbooks written for technicians and not engineers!

[2]If by chance I have missed anyone's digital electronics textbook that does use electron flow, please accept my apologies. I can only speak of what I have seen myself.

Once again, though, it should be understood that either convention of current notation is adequate for circuit analysis. I dearly wish this horrible state of affairs would come to an end, but the plain fact is it will not. Electron flow notation may have the advantage of greater correspondence to the actual state of affairs (in the vast majority of circuits), but conventional flow has the weight of over a hundred years of precedent, cultural inertia, and convenience. No matter which way you choose to think, at some point you will be faced with the opposing view.

Pick the notation you like best, and may you live long and prosper.

4.3 Electrical resistance and Ohm's Law

To review, *voltage* is the measure of potential energy available to electric charges. *Current* is the uniform drifting of electric charges in response to a voltage. We can have a voltage without having a current, but we cannot have a current without first having a voltage to motivate it[3]. Current without voltage would be equivalent to motion without a motivating force.

When electric charges move through a material such as metal, they will naturally encounter some friction, just as fluid moving through a pipe will inevitably encounter friction[4]. We have a name for this friction to electrical charge motion: *resistance*. Like voltage and current, resistance has its own special unit of measurement: the *ohm*, named in honor of the German physicist Georg Simon Ohm.

At this point it would be good to summarize and compare the symbols and units we use for voltage, current, and resistance:

Quantity	**Algebraic symbol**	**Unit**	**Unit abbreviation**
Voltage	V (or E)	Volt	V
Current	I	Ampere (or Amp)	A
Resistance	R	Ohm	Ω

Ohm defined resistance as the mathematical ratio between applied voltage and resulting current:

$$R = \frac{V}{I}$$

Verbally expressed, resistance is how much voltage it takes to force a certain rate of current through a conductive material. Many materials have relatively stable resistances, while others do not. Devices called *resistors* are sold which are manufactured to possess a very precise amount of resistance, for the purpose of limiting current in circuits (among other things).

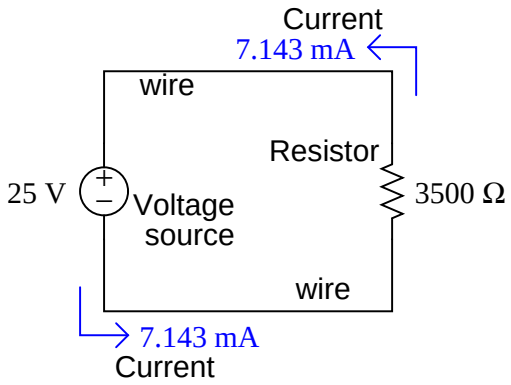
Here is an example of Ohm's Law in action: calculate the amount of current in a circuit with a voltage source of 25 V and a total resistance of 3500 Ω. Taking 25 volts and dividing by 3500 ohms, you should arrive at a result of 0.007143 amperes, or 7.143 milliamperes (7.143 mA).

One of the most challenging aspect of Ohm's Law is remembering to *keep all variables in context*. This is a common problem for many students when studying physics as well: none of the equations learned in a physics class will yield the correct results unless all the variables relate to the same object or situation. For instance, it would make no sense to try to calculate the kinetic energy of a moving object ($E = \frac{1}{2}mv^2$) by taking the mass of one object (m) and multiplying it by the square of the velocity of some *other* object (v^2). Likewise, with Ohm's Law, we must make sure the voltage, current, and resistance values we are using all relate to the same portion of the same circuit.

[3] Except in the noteworthy case of *superconductivity*, a phenomenon occurring at extremely low temperatures.

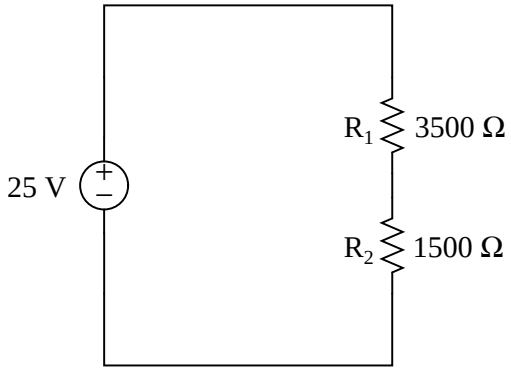
[4] Except in the noteworthy case of *superfluidity*, another phenomenon occurring at extremely low temperatures.

If the circuit in question has only one source of voltage, one resistance, and one path for current, there cannot be any mix-ups. Expressing the previous example in a schematic diagram:

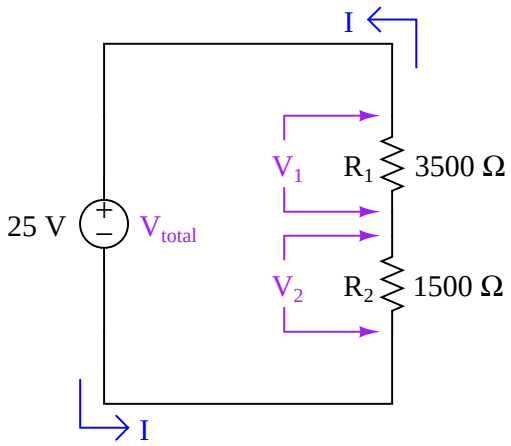


Note: arrows point in the direction of electron motion

However, if we look at a more complex circuit, we encounter the potential for mix-ups:



Which resistance do we use to calculate current in this circuit? Do we divide our 25 volts by 3500 ohms like we did last time, or do we divide it by 1500 ohms, or something entirely different? The answer to this question lies in the identification of voltages and currents. We know that the 25 volt potential will be impressed across the *total* of the two resistances R_1 and R_2, and since there is only one path for current they must share the same current. Thus, we actually have *three* voltages (V_1, V_2, and V_{total}), *three* resistances (R_1, R_2, and R_{total}), and only *one* current (I):



Note: arrows point in the direction of electron motion

Manipulating the Ohm's Law equation originally given ($R = \frac{V}{I}$) to solve for V, we end up with three equations for this circuit:

$$V_{total} = IR_{total} = I(R_1 + R_2)$$

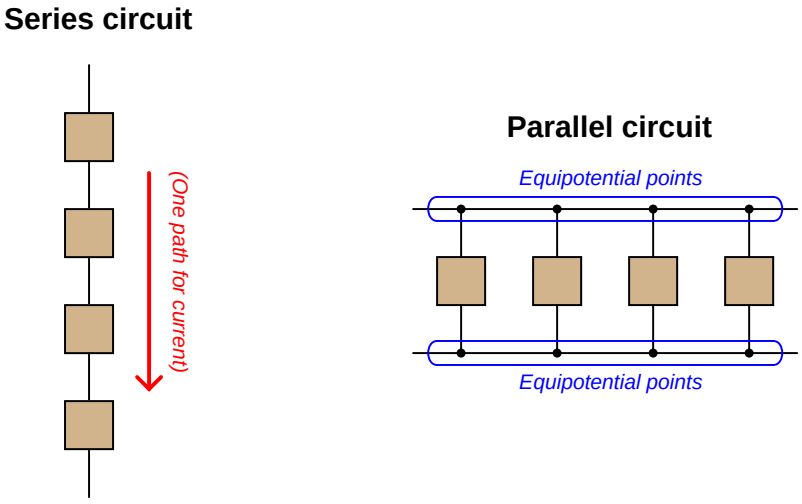
$$V_1 = IR_1$$

$$V_2 = IR_2$$

Thus, the current in this circuit is 5 milliamps (5 mA), the voltage across resistor R_1 is 17.5 volts, and the voltage across resistor R_2 is 7.5 volts.

4.4 Series versus parallel circuits

In addition to Ohm's Law, we have a whole set of rules describing how voltages, currents, and resistances relate in circuits comprised of multiple resistors. These rules fall evenly into two categories: *series* circuits and *parallel* circuits. The two circuit types are shown here, with squares representing any type of two-terminal electrical component:

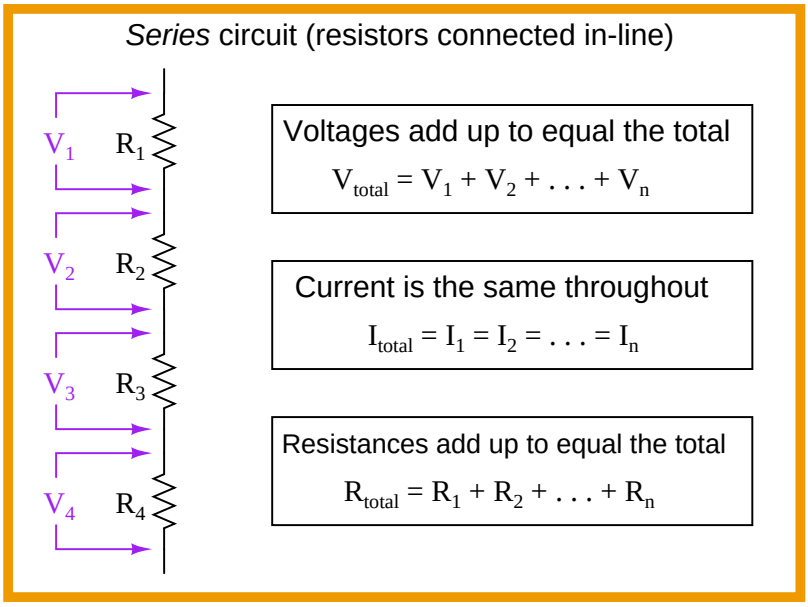


The defining characteristic of a series electrical circuit is it provides just one path for current. This means there can be only one value for current anywhere in the circuit, the exact same current for all components at any given time[5]. The principle of current being the same everywhere in a series circuit is actually an expression of a more fundamental law of physics: the *Conservation of Charge*, which states that electric charge cannot be created or destroyed. In order for current to have different values at different points in a series circuit indefinitely, electric charge would have to somehow appear and disappear to account for greater rates of charge flow in some areas than in others. It would be the equivalent of having different rates of water flow at different locations along one length of pipe[6].

[5] Interesting exceptions do exist to this rule, but only on very short time scales, such as in cases where we examine the a transient (pulse) signal nanosecond by nanosecond, and/or when very high-frequency AC signals exist over comparatively long conductor lengths.

[6] Those exceptional cases mentioned earlier in the footnote are possible only because electric charge may be temporarily stored and released by a property called *capacitance*. Even then, the law of charge conservation is not violated because the stored charges re-emerge as current at later times. This is analogous to pouring water into a bucket: just because water is poured into a bucket but no water leaves the bucket does not mean that water is magically disappearing! It is merely being stored, and can re-emerge at a later time.

Series circuits are defined by having only one path for current, and this means the steady-state current in a series circuit must be the same at all points of that circuit. It also means that the sum of all voltages dropped by load devices must equal the sum total of all source voltages, and that the total resistance of the circuit will be the sum of all individual resistances:



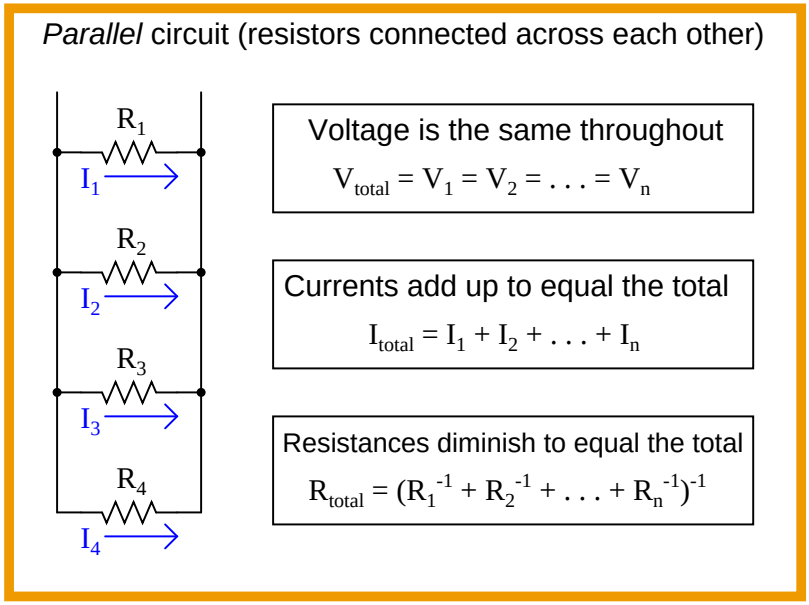
The defining characteristic of a parallel circuit, by contrast, is that all components share the same two equipotential points. "Equipotential" simply means "at the same potential" which points along an uninterrupted conductor must be[7]. This means there can be only one value of voltage anywhere in the circuit, the exact same voltage for all components at any given time[8]. The principle of voltage being the same across all parallel-connected components is (also) an expression of a more fundamental law of physics: the *Conservation of Energy*, in this case the conservation of specific potential energy which is the definition of voltage. In order for voltage to differ between parallel-connected components, the potential energy of charge carriers would have to somehow appear and disappear to account for lesser and greater voltages. It would be the equivalent of having a "high spots" and "low spots" of water mysteriously appear on the quiet surface of a lake, which we know cannot happen because water has the freedom to move, meaning any high spots would rush to fill any low spots[9].

[7] An ideal conductor has no resistance, and so there is no reason for a difference of potential to exist along a pathway where nothing stands in the way of charge motion. If ever a potential difference developed, charge carriers within the conductor would simply move to new locations and neutralize the potential.

[8] Again, interesting exceptions do exist to this rule on very short time scales, such as in cases where we examine the a transient (pulse) signal nanosecond by nanosecond, and/or when very high-frequency AC signals exist over comparatively long conductor lengths.

[9] The exceptional cases mentioned in the previous footnote exist only because the electrical property of *inductance* allows potential energy to be stored in a magnetic field, manifesting as a voltage different along the length of a conductor. Even then, the law of energy conservation is not violated because the stored energy re-emerges at a later time.

The sum of all component currents must equal the total current in a parallel circuit, and total resistance will be *less* than the smallest individual resistance value:



The rule for calculating total resistance in a parallel circuit perplexes many students with its weird compound reciprocal notation. There is a more intuitive way to understand this rule, and it involves a different quantity called *conductance*, symbolized by the letter G.

Conductance is defined as the reciprocal of resistance; that is, a measure of how *easily* electrical charge carriers may move through a substance. If the electrical resistance of an object doubles, then it now has *half* the conductance it did before:

$$G = \frac{1}{R}$$

It should be intuitively apparent that conductances add in parallel circuits. That is, the total amount of conductance for a parallel circuit must be the sum total of all individual conductances, because the addition of more conductive pathways must make it easier overall for charge carriers to move through the circuit. Thus,

$$G_{total} = G_1 + G_2 + \cdots + G_n$$

The formula shown here should be familiar to you. It has the same form as the total resistance formula for series circuits. Just as resistances add in series (more series resistance makes the overall resistance to current increase), conductances add in parallel (more conductive branches makes the overall conductance increase).

Knowing that resistance is the reciprocal of conductance, we may substitute $\frac{1}{R}$ for G wherever we see it in the conductance equation:

$$\frac{1}{R_{total}} = \frac{1}{R_1} + \frac{1}{R_2} + \cdots + \frac{1}{R_n}$$

Now, to solve for R_{total}, we need to reciprocate both sides:

$$R_{total} = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2} + \cdots + \frac{1}{R_n}}$$

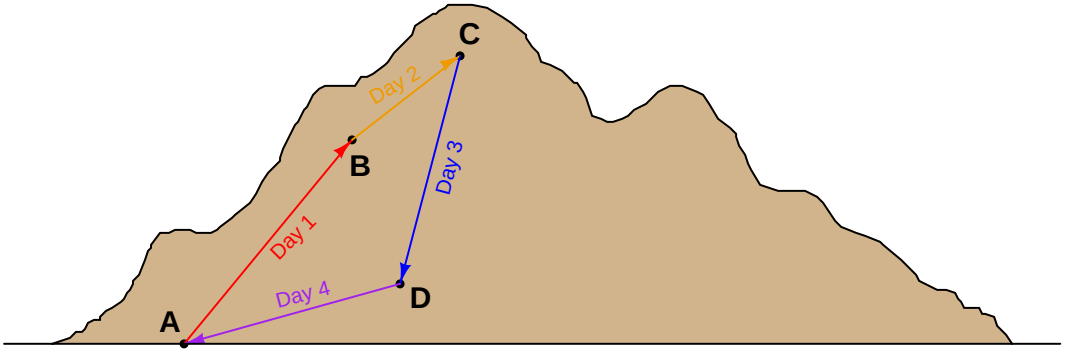
For both series and parallel circuits, total power dissipated by all load devices is equal to the total power delivered by all source devices. The configuration of a circuit is irrelevant to the balance between power supplied and power lost, because this balance is an expression of the Law of Energy Conservation.

4.5 Kirchhoff's Laws

Two extremely important principles in electric circuits were codified by Gustav Robert Kirchhoff in the year 1847, known as *Kirchhoff's Laws*. His two laws refer to voltages and currents in electric circuits, respectively.

Kirchhoff's Voltage Law states that the algebraic sum of all voltages in a closed loop is equal to zero. Another way to state this law is to say that for every rise in potential there must be an equal fall, if we begin at any point in a circuit and travel in a loop back to that same starting point.

An analogy for visualizing Kirchhoff's Voltage Law is hiking up a mountain. Suppose we start at the base of a mountain and hike to an altitude of 5,000 feet to set up camp for an overnight stay. Then, the next day we set off from camp and hike further up another 3,500 feet. Deciding we've climbed high enough for two days, we set up camp again and stay the night. The next day we hike down 6,200 feet to a third location and camp once gain. On the fourth day we hike back to our original starting point at the base of the mountain. We can summarize our hiking adventure as a series of rises and falls like this:

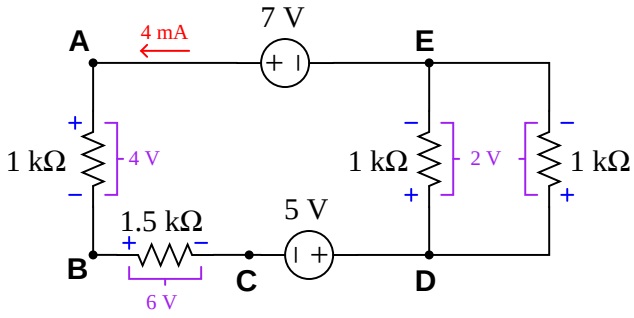


Day	Path	Altitude gain/loss
Day 1	AB	+5,000 feet
Day 2	BC	+3,500 feet
Day 3	CD	-6,200 feet
Day 4	DA	-2,300 feet
(Total)	ABCDA	0 feet

Of course, no one would brag to their friends that they spent four days hiking a total altitude of 0 feet, so people generally speak in terms of the *highest* point reached: in this case 8,500 feet. However, if we track each day's gain or loss in algebraic terms (maintaining the mathematical sign, either positive or negative), we see that the end sum is zero (and indeed *must always be zero*) if we finish at our starting point.

If we view this scenario from the perspective of potential energy as we lift a constant mass from point to point, we would conclude that we were doing work on that mass (i.e. investing energy in it by lifting it higher) on days 1 and 2, but letting the mass do work on us (i.e. releasing energy by lowering it) on days 3 and 4. After the four-day hike, the net potential energy imparted to the mass is zero, because it ends up at the exact same altitude it started at.

Let's apply this principle to a real circuit, where total current and all voltage drops have already been calculated for us:



If we trace a path ABCDEA, we see that the algebraic voltage sum in this loop is zero:

Path	Voltage gain/loss
AB	- 4 volts
BC	- 6 volts
CD	+ 5 volts
DE	- 2 volts
EA	+ 7 volts
ABCDEA	0 volts

We can even trace a path that does not follow the circuit conductors or include all components, such as EDCBE, and we will see that the algebraic sum of all voltages is still zero:

Path	Voltage gain/loss
ED	+ 2 volts
DC	- 5 volts
CB	+ 6 volts
BE	- 3 volts
EDCBE	0 volts

Kirchhoff's Voltage Law is often a difficult subject for students, precisely because voltage itself is a difficult concept to grasp. Remember that there is no such thing as voltage at a single point; rather, voltage exists only as a *differential* quantity. To intelligently speak of voltage, we must refer to either a *loss* or *gain* of potential between **two points**.

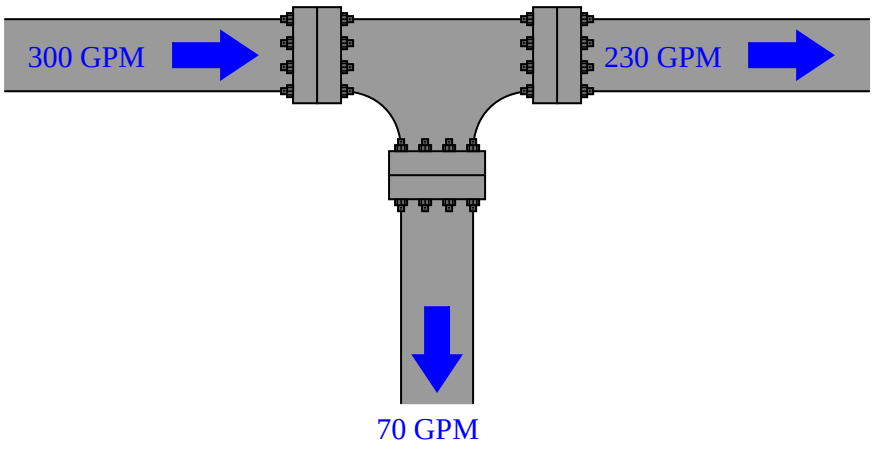
Our analogy of altitude on a mountain is particularly apt. We cannot intelligently speak of some point on the mountain as having a specific altitude unless we assume a point of reference to measure from. If we say the mountain summit is 9,200 feet high, we usually mean 9,200 feet *higher than sea level*, with the level of the sea being our common reference point. However, our hiking adventure where we climbed 8,500 feet in two days did not imply that we climbed to an absolute altitude of 8,500 feet above sea level. Since I never specified the sea-level altitude at the base of the mountain,

it is impossible to calculate our absolute altitude at the end of day 2. All you can tell from the data given is that we climbed 8,500 feet *above* the mountain base, wherever that happens to be with reference to sea level.

So it is with electrical voltage as well: most circuits have a point labeled as *ground* where all other voltages are referenced. In DC-powered circuits, this ground point is often the negative pole of the DC power source[10]. Voltage is fundamentally a quantity relative between two points: a measure of how much potential has *increased* or *decreased* moving from one point to another.

Kirchhoff's Current Law is a much easier concept to grasp. This law states that the algebraic sum of all currents at a junction point (called a *node*) is equal to zero. Another way to state this law is to say that for every electron entering a node, one must exit somewhere.

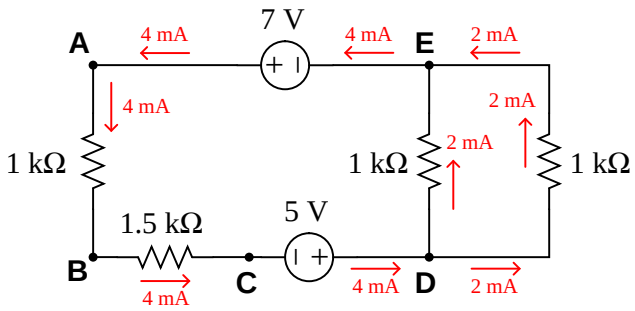
An analogy for visualizing Kirchhoff's Current Law is water flowing into and out of a "tee" fitting:



So long as there are no leaks in this piping system, every drop of water entering the tee must be balanced by a drop exiting the tee. For there to be a continuous mis-match between flow rates would imply a violation of the Law of Mass Conservation.

[10] But not always! There do exist positive-ground systems, particularly in telephone circuits and in some early automobile electrical systems.

Let's apply this principle to a real circuit, where all currents have been calculated for us:



At nodes where just two wires connect (such as points A, B, and C), the amount of current going in to the node exactly equals the amount of current going out (4 mA, in each case). At nodes where three wires join (such as points D and E), we see one large current and two smaller currents (one 4 mA current versus two 2 mA currents), with the directions such that the sum of the two smaller currents form the larger current.

4.6 Electrical sources and loads

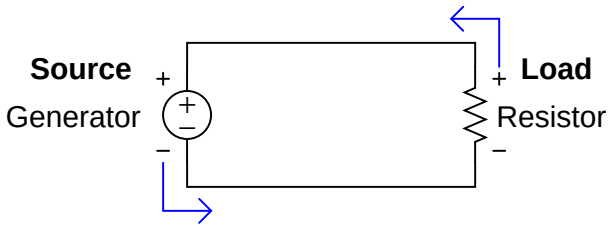
By definition, and *source* is a device that inputs energy into a system, while a *load* is a device that extracts energy from a system. Examples of typical electrical sources include generators, photovoltaic cells, thermopiles, and primary-cell batteries. Examples of typical electrical loads include resistors, lamps, and electric motors.

In a working circuit, electrical sources and loads may be easily distinguished by comparison of their current directions and voltage drop polarities. An electrical source always manifests a voltage polarity in a direction that *assists* the direction of charge flow. An electrical source always manifests a voltage polarity in a direction that *opposes* the direction of charge flow.

The convention used to designate direction of current (charge flow) becomes very important here. Since there are two commonly accepted notations – electron flow and "conventional" flow, exactly opposite of each other – it is easy to become confused.

First we see a diagram showing a source and a load, using electron flow notation. Electrons, being negatively charged particles, are repelled by the negative (-) poles of both source and load, and attracted to the positive (+) poles of both source and load. The difference between source and load is that the source device *motivates* the flow of electrons while the load device *resists* the flow of electrons:

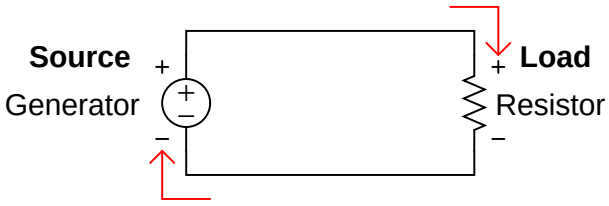
Shown using electron flow notation



Electrons are repelled by the (-) poles
and attracted to the (+) poles

Next we see a diagram showing the same source and load, this time using "conventional" flow notation to designate the direction of current. Here we must imagine positively-charged carriers moving through the wires instead of electrons. These positive charge carriers are repelled by any positive (+) pole and attracted to any negative (-) pole. Viewed in this light, we see the exact same principle at work: the source device is seen to *motivate* the flow of these positive charge carriers while the load device *resists* the flow:

Shown using conventional flow notation



Positive charge carriers are repelled by the
(+) poles and attracted to the (-) poles

In later sections, we encounter devices with the ability to act as sources and loads at different times. Both capacitors (see section 4.10 starting on page 226) and inductors (see section 4.11 starting on page 228) have the ability to temporarily contribute to and extract energy from electrical circuits, both having the ability to act as energy *storage* devices.

4.7 Resistors

Resistance is dissipative opposition to the flow of charge carriers. All conductors (except superconductors) possess some electrical resistance. The relationship between voltage, current, and resistance is known as *Ohm's Law*:

$$V = IR$$

Conductance (G) is the reciprocal of resistance:

$$G = \frac{1}{R}$$

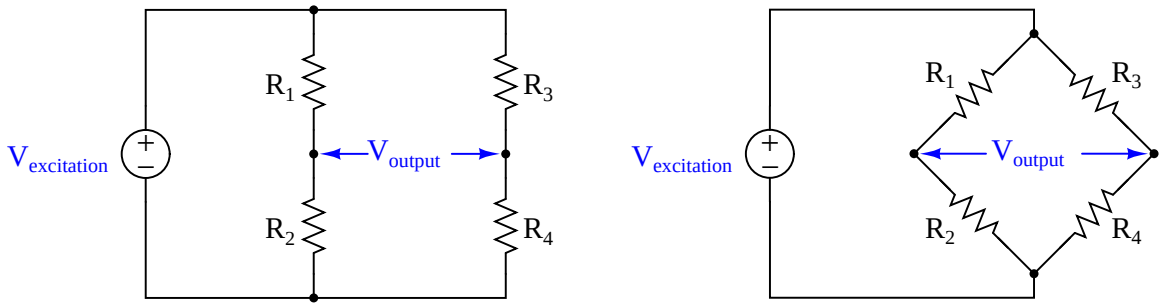
Resistors are devices expressly designed and manufactured to possess electrical resistance. They are constructed of a partially conductive material such as carbon or metal alloy. Resistors have power dissipation ratings as well as resistance ratings. Here are some schematic symbols for resistors:

The amount of power dissipated by a resistance may be calculated as a function of either voltage or current, and is known as *Joule's Law*:

$$P = IV \qquad P = \frac{V^2}{R} \qquad P = I^2R$$

4.8 Bridge circuits

A *bridge* circuit is basically a pair of voltage dividers where the circuit output is taken as the difference in potential between the two dividers. Bridge circuits may be drawn in schematic form in an H-shape or in a diamond shape, although the diamond configuration is more common:

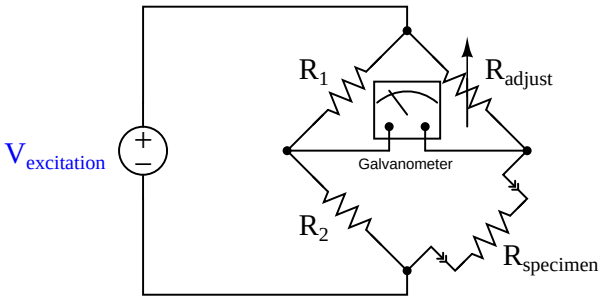


The voltage source powering the bridge circuit is called the *excitation* source. This source may be DC or AC depending on the application of the bridge circuit. The components comprising the bridge need not be resistors, either: capacitors, inductors, lengths of wire, sensing elements, and other component forms are possible, depending on the application.

Two major applications exist for bridge circuits, which will be explained in the following subsections.

4.8.1 Component measurement

Bridge circuits may be used to test components. In this capacity, one of the "arms" of the bridge circuit is comprised of the component under test, while at least one of the other "arms" is made adjustable. The common *Wheatstone bridge* circuit for resistance measurement is shown here:

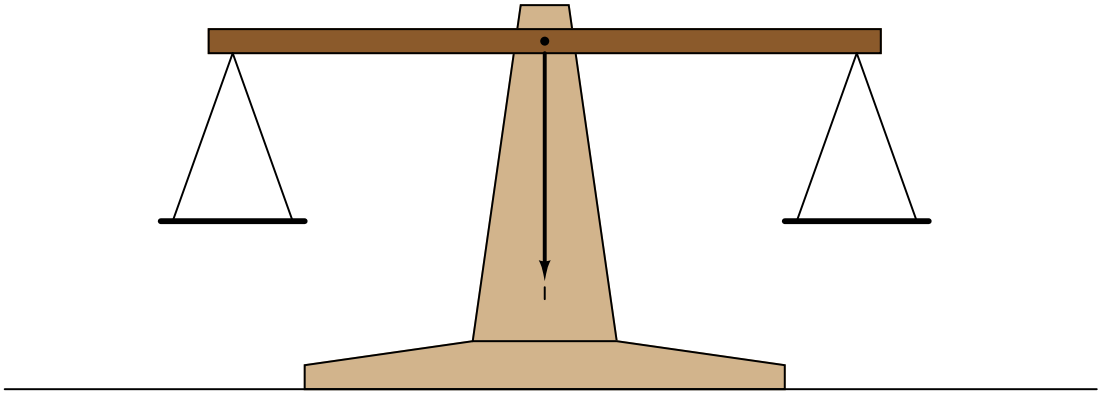


Fixed resistors R_1 and R_2 are of precisely known value and high precision. Variable resistor R_{adjust} has a labeled knob allowing for a person to adjust and read its value to a high degree of precision. When the ratio of the variable resistance to the specimen resistance equals the ratio of the two fixed resistors, the sensitive galvanometer will register exactly zero volts regardless of the excitation source's value. This is called a *balanced* condition for the bridge circuit:

$$\frac{R_1}{R_2} = \frac{R_{adjust}}{R_{specimen}}$$

When the two resistance ratios are equal, the voltage drops across the respective resistances will also be equal. Kirchhoff's Voltage Law declares that the voltage differential between two equal and opposite voltage drops must be zero, accounting for the meter's indication of balance.

It would not be inappropriate to relate this to the operation of a laboratory balance-beam scale, comparing a specimen of unknown mass against a set of known masses. In either case, the instrument is merely comparing an unknown quantity against an (adjustable) known quantity, indicating a condition of equality between the two:

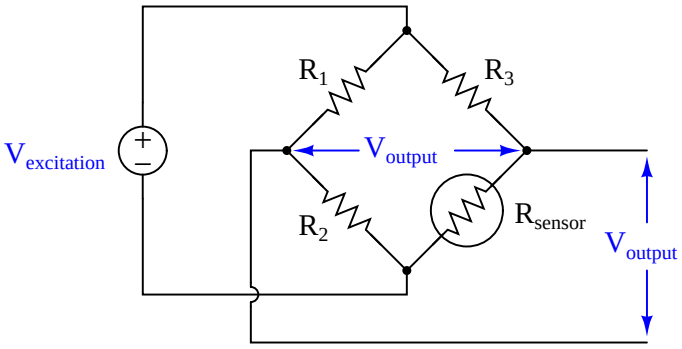


Many legacy instruments were designed around the concept of a *self-balancing* bridge circuit, where an electric servo motor drove a potentiometer to achieve a balanced condition against the voltage produced by some process sensor. Analog electronic paper chart recorders often used this principle. Almost all pneumatic process instruments use this principle to translate the force of a sensing element into a variable air pressure.

Modern bridge circuits are mostly used in laboratories for extremely precise component measurements. Very rarely will you encounter a Wheatstone bridge circuit used in the process industries.

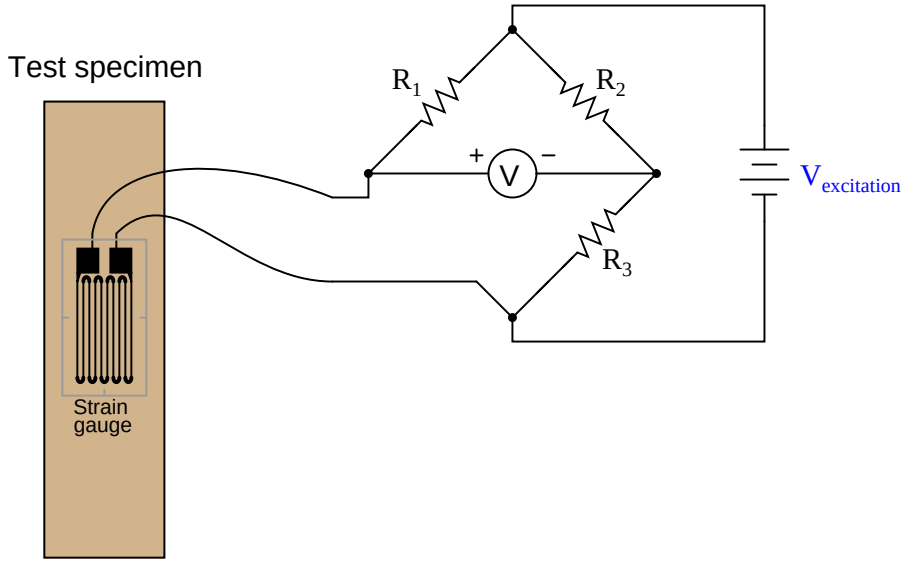
4.8.2 Sensor signal conditioning

A different application for bridge circuits is to convert the output of an electrical sensor into a voltage signal representing some physical measurement. This is by far the most popular use of bridge measurement circuits in industry, and here we see the same circuit used in an entirely different manner from that of the balanced Wheatstone bridge circuit.



Here, the bridge will be balanced only when R_{sensor} is at one particular resistance value. Unlike the Wheatstone bridge, which serves to measure a component's value when the circuit is balanced, this bridge circuit will probably spend most of its life in an unbalanced condition. The output voltage changes as a function of sensor resistance, which makes that voltage a reflection of the sensor's physical condition. In the above circuit, we see that the output voltage increases (positive on the top wire, negative on the bottom wire) as the resistance of R_{sensor} increases.

One of the most common applications for this kind of bridge circuit is in strain measurement, where the mechanical strain of an object is converted into an electrical signal. The sensor used here is a device known as a *strain gauge*: a folded wire designed to stretch and compress with the object under test, altering its electrical resistance accordingly.



When the specimen is stretched along its long axis, the metal wires in the strain gauge stretch with it, increasing their length and decreasing their cross-sectional area, both of which work to increase the wire's electrical resistance. This stretching is microscopic in scale, but the resistance change is measurable and repeatable within the specimen's elastic limit. In the above circuit example, stretching the specimen will cause the voltmeter to read upscale (as defined by the polarity marks). Compressing the specimen along its long axis has the opposite effect, decreasing the strain gauge resistance and driving the meter downscale.

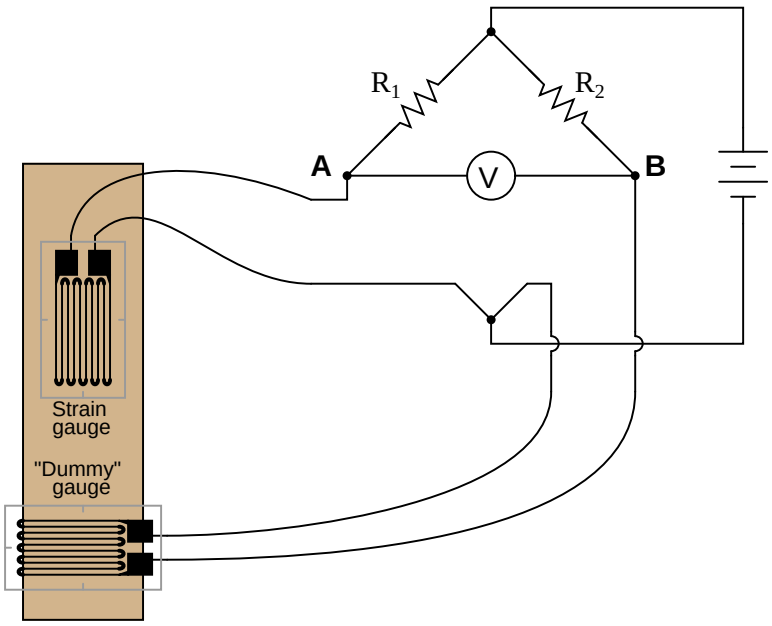
Strain gauges are used to precisely measure the strain (stretching or compressing motion) of mechanical elements. One application for strain gauges is the measurement of strain on machinery components, such as the frame components of an automobile or airplane undergoing design development testing. Another application is in the measurement of force in a device called a *load cell*. A "load cell" is comprised of one or more strain gauges bonded to the surface of a metal structure having precisely known elastic properties. This metal structure will stretch and compress very precisely with applied force, as though it were an extremely stiff spring. The strain gauges bonded to this structure measure the strain, translating applied force into electrical resistance changes.

You can see what a load cell looks like in the following photograph:

Strain gauges are not the only dynamic element applicable to bridge circuits. In fact, any resistance-based sensor may be used in a bridge circuit to translate a physical measurement into an electrical (voltage) signal. Thermistors (changing resistance with temperature) and photocells (changing resistance with light exposure) are just two alternatives to strain gauges.

It should be noted that the amount of voltage output by this bridge circuit depends both on the amount of resistance change of the sensor *and* the value of the excitation source. This dependency on source voltage value is a major difference between a sensing bridge circuit and a Wheatstone (balanced) bridge circuit. In a perfectly balanced bridge, the excitation voltage is irrelevant: the output voltage is zero no matter what source voltage value you use. In an unbalanced bridge circuit, however, source voltage value matters! For this reason, these bridge circuits are often rated in terms of how many millivolts of output they produce *per volt of excitation* per unit of physical measurement (microns of strain, newtons of stress, etc.).

An interesting feature of a sensing bridge circuit is its ability to cancel out unwanted variables. In the case of a strain gauge, for example, mechanical strain is not the only variable affecting gauge resistance. Temperature also affects gauge resistance. Since we do not wish our strain gauge to also act as a thermometer (which would make measurements very uncertain – how would we differentiate the effects of changing temperature from the effects of changing strain?), we must find some way to nullify resistance changes due solely to temperature, such that our bridge circuit will respond only to changes in strain. The solution is to creatively use a "dummy" strain gauge as another arm of the bridge:

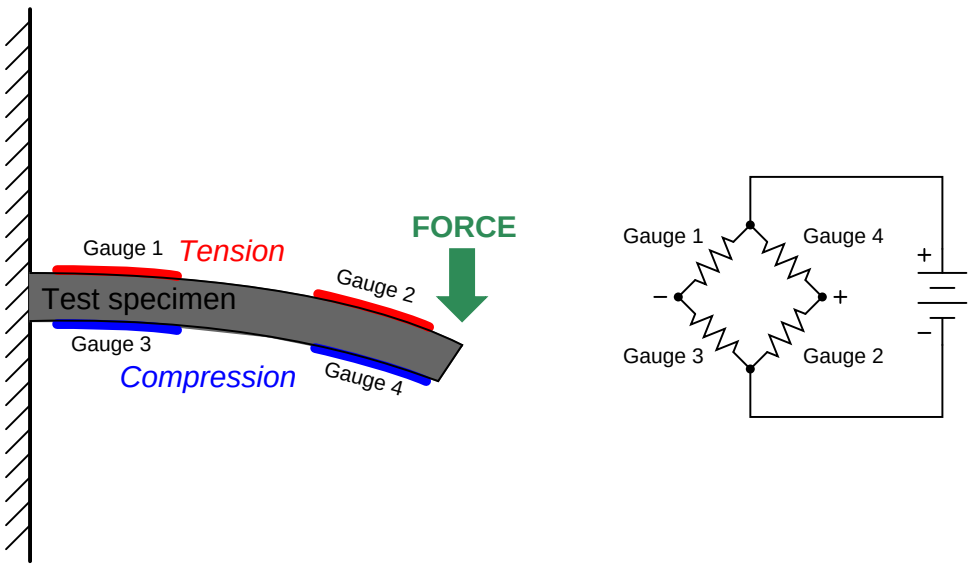


The "dummy" gauge is attached to the specimen in such a way that it maintains the same temperature as the active strain gauge, yet experiences no strain. Thus, any *difference* in gauge resistances must be due solely to specimen strain. The differential nature of the bridge circuit naturally translates the differential resistance of the two gauges into one voltage signal representing strain.

If thermistors are used instead of strain gauges, this circuit becomes a differential temperature sensor. Differential temperature sensing circuits are used in solar heating control systems, to detect when the solar collector is hotter than the room or heat storage mass being heated.

Sensing bridge circuits may have more than one active "arm" as well. The examples you have seen so far in this section have all been *quarter-active* bridge circuits. It is possible, however, to incorporate more than one sensor into the same bridge circuit. So long as the sensors' resistance changes are coordinated, their combined effect will be to increase the sensitivity (and often the linearity as well) of the measurement.

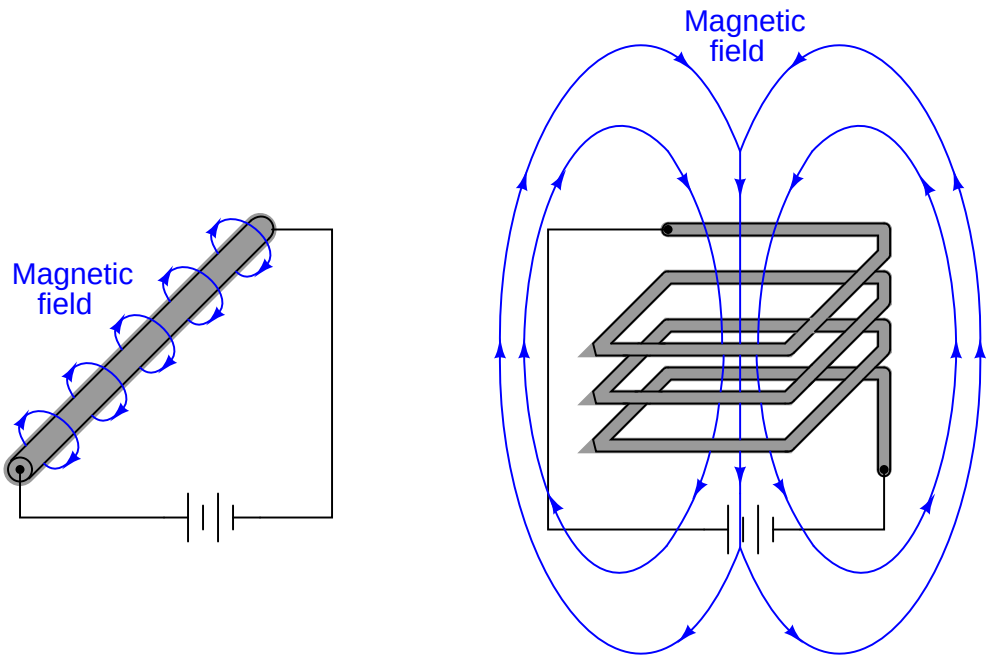
For example, *full-active bridge* circuits are sometimes built out of four strain gauges, where each strain gauge comprises one arm of the bridge. Two of the strain gauges must compress and the other two must stretch under the application of the same mechanical force, in order that the bridge will become unbalanced with strain:



Not only does a full-active bridge circuit provide greater sensitivity and linearity than a quarter-active bridge, but it also *naturally* provides temperature compensation without the need for "dummy" strain gauges, since the resistances of all four strain gauges will change by the same proportion if the specimen temperature changes.

4.9 Electromagnetism

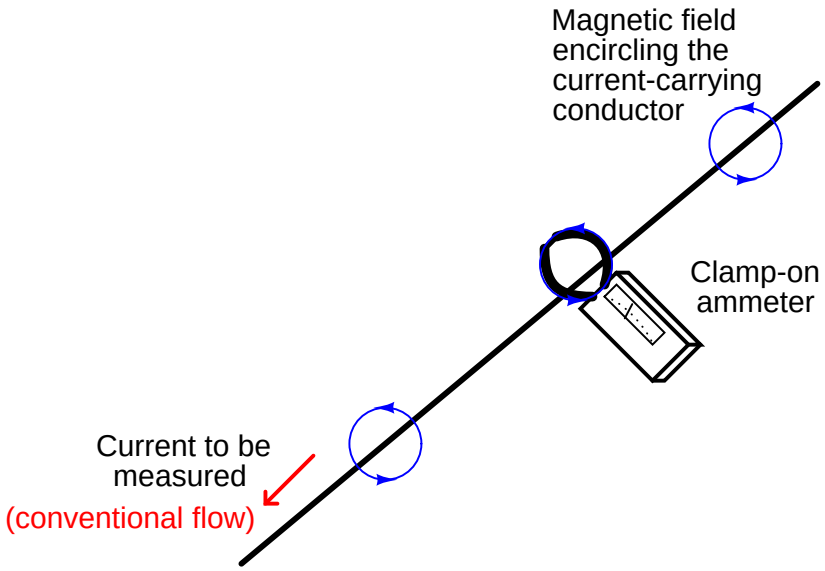
The fundamental principle of *electromagnetism* is that an electric current will create a magnetic field at right angles to the direction of the current. If the electric current travels in a straight path, the lines of magnetic flux will form concentric circles around that path. If the electric current travels in a circular path (i.e. through a loop or coil of wire), the magnetic lines of flux will form straight lines down the center of the coil, wrapping around at the ends to form a complete loop of its own:



Magnetic field strength is directly proportional to the amount of current in the conductor (and also directly proportional to the number of "turns" in a coiled wire), such that the unit of measurement[11] for magnetic field strength is the *amp-turn*.

[11] Both in the British system of measurement *and* the SI metric system of measurement! The older metric system (called "CGS" for Centimeter-Gram-Second) had a special unit of measurement called the *Gilbert* for expressing magnetic field strength, with 1 Gilbert (Gb) equal to 0.7958 Amp-turns (At).

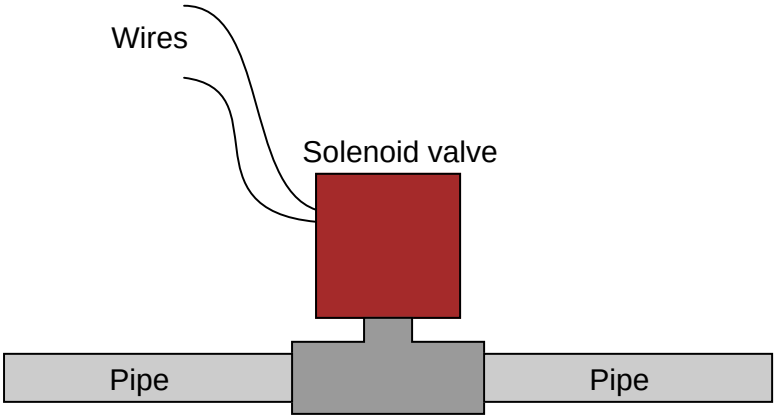
The directly proportional relationship between current intensity and magnetic field strength is exploited by *clamp-on ammeters*, which are able to measure electric current without the need for contact with the conductor:



Very strong magnetic fields may be generated with wire coils, since the magnetic fields surrounding each "turn" of wire in a coil tend to overlap constructively, supporting one another to form a stronger total field. The magnetic field from a wire coil may be so strong, in fact, that it is useful for creating an attractive force upon a ferrous object (called an *armature*) strong enough to move mechanisms. This arrangement of a wire coil and an iron armature is called a *solenoid*[12].

[12] The word "solenoid" may also be used to describe a wire coil with no armature, but the more common industrial use of the word refers to the complete arrangement of coil and movable armature.

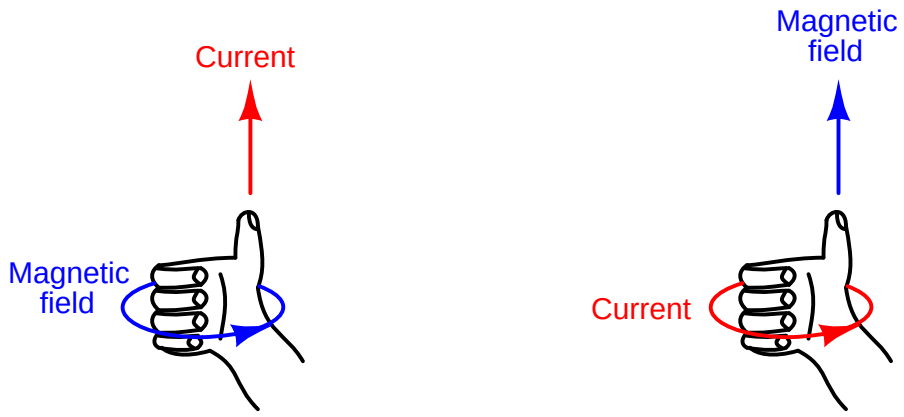
A practical example of a solenoid is a *solenoid valve*: a mechanical valve opened and/or closed by the application of electric current through the coil:



When the coil is energized by an external source of electric current, the magnetic field attracts the movable armature, thereby actuating the valve. A spring typically returns the valve mechanism back to its original position upon de-energization of the coil.

Both electric current and magnetic field lines are *vectors*, having both magnitude and direction. As we have seen already, there is a perpendicular relationship between these two vectors. This relationship may be visualized by a simple rule called the *right-hand rule*, whereby the fingers and thumb of a human right hand represent the vector orientations of current and magnetism (or visa-versa). Using the right-hand rule, digits representing current direction assume the use of *conventional flow* rather than electron flow[13], while digits representing magnetism point in the direction of "North."

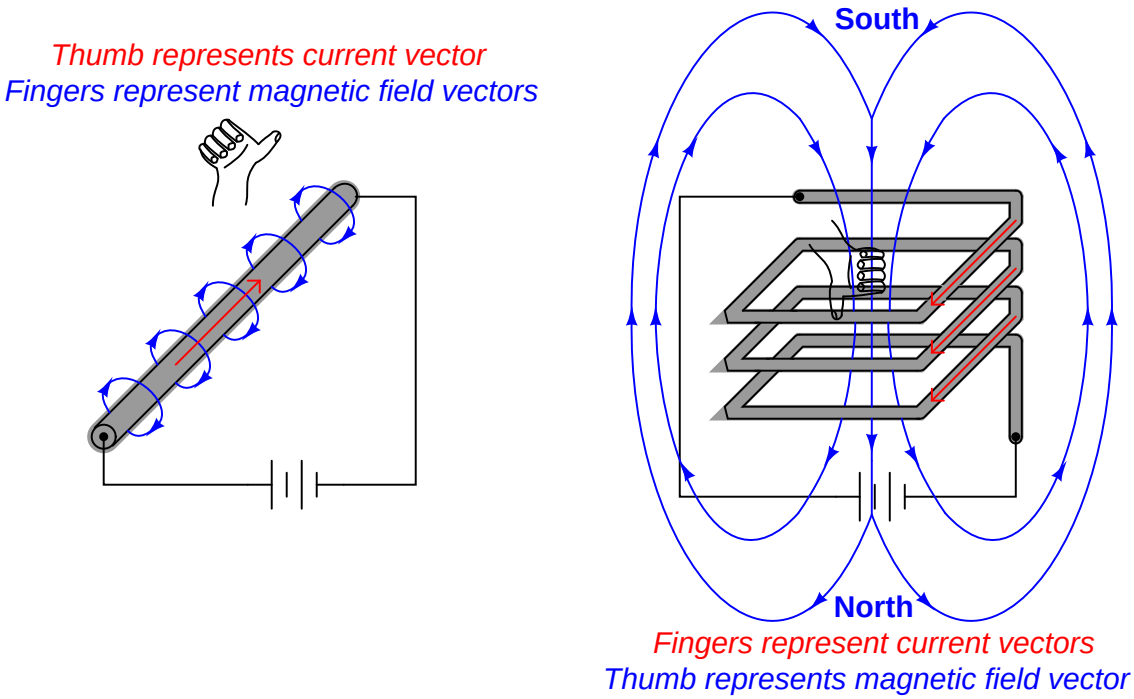
To use this rule, curl the four fingers of your right hand such that they point toward the palm of that hand, and extend your right thumb so that it points perpendicularly to your curled fingers. Your right hand should look like this:



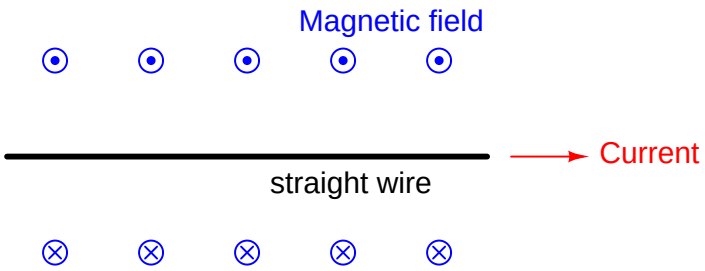
Your curled fingers will represent one set of vectors, while your thumb points in the direction of the other. Whether the fingers represent current and the thumb magnetism, or whether the fingers represent magnetism and the thumb current, is irrelevant: the rule works both ways.

[13] There is also a left-hand rule for fans of electron flow, but in this book I will default to conventional flow. For a more complete discussion on this matter, see section 4.2.1 beginning on page 193.

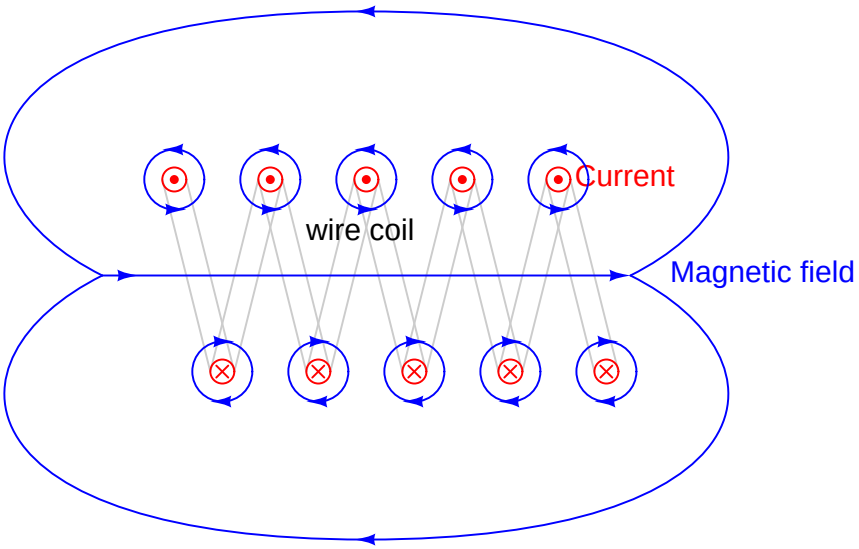
This flexibility makes the right-hand rule easy to apply to different situations such as these:



Physicists have devised a convention for clearly illustrating the directions of perpendicular vectors (arrows) without resorting to illustrations drawn in 3-dimensional perspective. The following two-dimensional illustration shows the magnetic field surrounding a straight, current-carrying wire. The magnetic field, of course, encircles the wire. This is shown by the alternating "dot" and "cross" symbols above and below the wire, representing arrow heads (circles with dots) coming "out" of the page directly at the reader, and representing arrow tails (circles with crosses) headed straight into the page away from the reader:



The same notation may be used to show the perpendicular relationship between current and magnetic flux lines for a coiled conductor. Here, the arrow "tips" and "tails" represent current (conventional flow) entering and exiting the page, while the horizontal arrow represents magnetic field direction:



Note how the individual magnetic fields surrounding each wire in the coil all have their arrows pointing to the right in the coil's interior, and to the left at the coil's exterior. This shows how the individual magnetic loops constructively add to generate a large magnetic field through the center of the coil, looping around back to the other end of the coil.

4.10 Capacitors

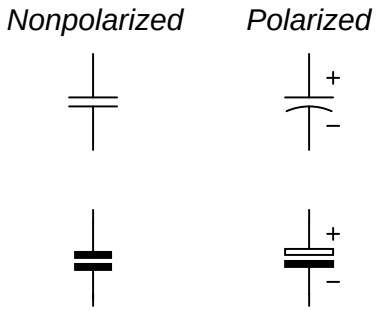
Any two electrical conductors separated by an insulating medium possess the characteristic called *capacitance*: the ability to store energy in the form of an electric field. Capacitance is symbolized by the capital letter C and is measured in the unit of the *Farad* (F). The relationship between capacitance, stored electric charge (Q), and voltage (V) is as follows:

$$Q = CV$$

For example, a capacitance having a value of 33 microfarads charged to a voltage of 5 volts would store an electric charge of 165 microcoulombs.

Capacitance is a non-dissipative quantity. Unlike resistance, a pure capacitance does not dissipate energy in the form of heat; rather, it stores and releases energy from and to the rest of the circuit.

Capacitors are devices expressly designed and manufactured to possess capacitance. They are constructed of a "sandwich" of conductive plates separated by an insulating *dielectric*. Capacitors have voltage ratings as well as capacitance ratings. Here are some schematic symbols for capacitors:



A capacitor's capacitance is related to the electric permittivity of the dielectric material (symbolized by the Greek letter "epsilon," ϵ), the cross-sectional area of the overlapping plates (A), and the distance separating the plates (d):

$$C = \frac{\epsilon A}{d}$$

Capacitance adds when capacitors are connected in parallel. It diminishes when capacitors are connected in series:

$$C_{parallel} = C_1 + C_2 + \cdots C_n \qquad C_{series} = \frac{1}{\frac{1}{C_1} + \frac{1}{C_2} + \cdots + \frac{1}{C_n}}$$

The relationship between voltage and current for a capacitor is as follows:

$$I = C\frac{dV}{dt}$$

As such, capacitors oppose changes in voltage over time by creating a current. This behavior makes capacitors useful for stabilizing voltage in DC circuits. One way to think of a capacitor in a DC circuit is as a *temporary voltage source*, always "wanting" to maintain voltage across its terminals at the same value.

The amount of potential energy (E_p, in units of joules) stored by a capacitor may be determined by altering the voltage/current/capacitance equation to express power ($P = IV$) and then applying some calculus (recall that power is defined as the time-derivative of work or energy, $P = \frac{dW}{dt} = \frac{dE}{dt}$):

$$I = C\frac{dV}{dt}$$

$$P = IV = CV\frac{dV}{dt}$$

$$\frac{dE_p}{dt} = CV\frac{dV}{dt}$$

$$\frac{dE_p}{dt}\,dt = CV\,dV$$

$$\int \frac{dE_p}{dt}\,dt = \int CV\,dV$$

$$\int dE_p = C\int V\,dV$$

$$E_p = \frac{1}{2}CV^2$$

In an AC circuit, the amount of capacitive reactance (X_C) offered by a capacitor is inversely proportional to both capacitance and frequency:

$$X_C = \frac{1}{2\pi fC}$$

This means an AC signal finds it "easier" to pass through a capacitor (i.e. less ohms of reactance) at higher frequencies than at lower frequencies.

4.11 Inductors

Any conductor possesses a characteristic called *inductance*: the ability to store energy in the form of a magnetic field. Inductance is symbolized by the capital letter L and is measured in the unit of the *Henry* (H).

Inductance is a non-dissipative quantity. Unlike resistance, a pure inductance does not dissipate energy in the form of heat; rather, it stores and releases energy from and to the rest of the circuit.

Inductors are devices expressly designed and manufactured to possess inductance. They are typically constructed of a wire coil wound around a ferromagnetic core material. Inductors have current ratings as well as inductance ratings. Due to the effect of *magnetic saturation*, inductance tends to decrease as current approaches the rated maximum value in an iron-core inductor. Here are some schematic symbols for inductors:

An inductor's inductance is related to the magnetic permeability of the core material (μ), the number of turns in the wire coil (N), the cross-sectional area of the coil (A), and the length of the coil (l):

$$L = \frac{\mu N^2 A}{l}$$

Inductance adds when inductors are connected in series. It diminishes when inductors are connected in parallel:

$$L_{series} = L_1 + L_2 + \cdots L_n \qquad\qquad L_{parallel} = \frac{1}{\frac{1}{L_1} + \frac{1}{L_2} + \cdots + \frac{1}{L_n}}$$

The relationship between voltage and current for an inductor is as follows:

$$V = L\frac{dI}{dt}$$

As such, inductors oppose changes in current over time by dropping a voltage. This behavior makes inductors useful for stabilizing current in DC circuits. One way to think of an inductor in a DC circuit is as a *temporary current source*, always "wanting" to maintain current through its coil at the same value.

The amount of potential energy (E_p, in units of joules) stored by an inductor may be determined by altering the voltage/current/inductance equation to express power ($P = IV$) and then applying some calculus (recall that power is defined as the time-derivative of work or energy, $P = \frac{dW}{dt} = \frac{dE}{dt}$):

$$V = L\frac{dI}{dt}$$

$$P = IV = LI\frac{dI}{dt}$$

$$\frac{dE_p}{dt} = LI\frac{dI}{dt}$$

$$\frac{dE_p}{dt}\,dt = LI\,dI$$

$$\int \frac{dE_p}{dt}\,dt = \int LI\,dI$$

$$\int dE_p = L\int I\,dI$$

$$E_p = \frac{1}{2}LI^2$$

In an AC circuit, the amount of inductive reactance (X_L) offered by an inductor is directly proportional to both inductance and frequency:

$$X_L = 2\pi f L$$

This means an AC signal finds it "harder" to pass through an inductor (i.e. more ohms of reactance) at higher frequencies than at lower frequencies.

References

Boylestad, Robert L., *Introductory Circuit Analysis*, 9th Edition, Prentice Hall, Upper Saddle River, NJ, 2000.

Chapter 5

AC electricity

While *direct current* (DC) refers to the flow of electrical charge carriers in a continuous direction, *alternating current* (or *AC*) refers to a periodic reversal of charge flow direction[1]. As a mode of transferring electrical power, AC is tremendously useful because it allows us to use *transformers* to easily and efficiently step voltage up or down at will. If an electro-physical sensor senses a physical quantity that oscillates, the electric signal it produces will oscillate (AC) as well. For both these reasons, an instrument technician needs to be aware of how AC circuits work, and how to understand them mathematically.

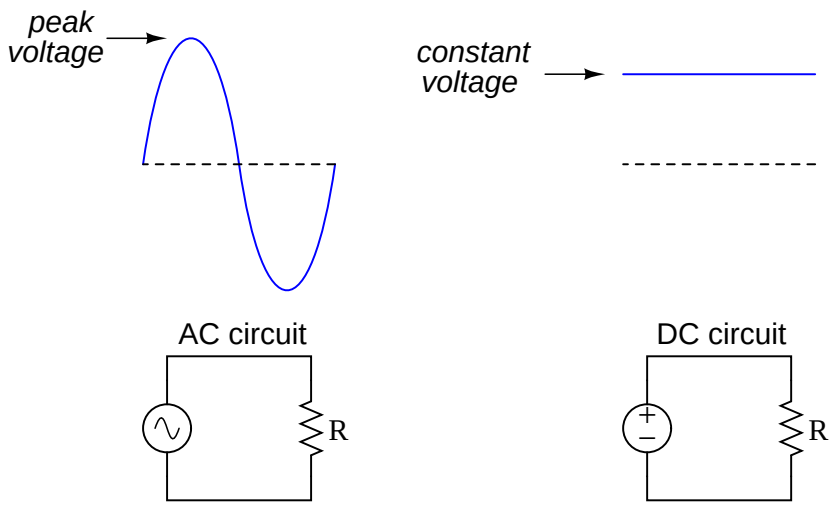
[1] It is also acceptable to refer to electrical voltages and/or currents that vary periodically over time even if their directions never alternate, as AC *superimposed* on DC.

5.1 RMS quantities

It is often useful to be able to express the amplitude of an AC quantity such as voltage or current in terms that are equivalent to direct current (DC). Doing so provides an "apples-to-apples" comparison between AC and DC quantities that makes comparative circuit analysis much easier.

The most popular standard of equivalence is based on *work* and *power*, and we call this the *root-mean-square* value of an AC waveform, or RMS for short. For example, an AC voltage of 120 volts "RMS" means that this AC voltage is capable of producing the exact same amount of power (in Watts) at an electrical load as a 120 volt DC source powering the exact same load.

The problem is exactly how to calculate this "RMS" value if all we know about the AC waveform is its peak value. If we compare a sine wave and a DC "wave" side by side, it is clear that the sine wave must peak at a greater value than the constant DC level in order to be equivalent in terms of doing the same *work* in the same amount of time:



At first, it might seem like the correct approach would be to use calculus to integrate the sine wave over one-half of a cycle (from 0 to π radians) and figure out how much area is under the curve. This is close, but not fully correct. You see, the ability of an electrical voltage to produce a power dissipation at a resistor is not directly proportional to the magnitude of that voltage, but rather proportional to the *square* of the magnitude of that voltage! In mathematical terms, power is predicted by the following equation:

$$P = \frac{V^2}{R}$$

If we double the amount of voltage applied to a resistor, the power increases four-fold. If we triple the voltage, the power goes up by a factor of nine! If we are to figure out the "RMS" equivalent value of a sine wave, we must take this nonlinearity into consideration.

First let us begin with a mathematical equivalence between the DC and AC cases. On one hand, the amount of work done by the DC voltage source will be equal to the power of that circuit multiplied by time. The unit of measurement for power is the *Watt*, which is defined as 1 Joule of work per second. So, multiplying the steady power rate in a DC circuit by the time we keep it powered will result in an answer of joules (total energy dissipated by the resistor):

$$\text{Work} = \left(\frac{V^2}{R}\right) t$$

On the other hand, the amount of work done by a sine-wave-shaped AC voltage is equal to the *square* of the sine function divided by resistance, integrated over a specified time period. In other words, we will use the calculus process of *integration* to calculate the area underneath the function $\sin^2 t$ rather than under the function $\sin t$. Since the interval from 0 to π will encompass the essence of the sine wave's shape, this will be our integration interval:

$$\text{Work} = \int_0^{\pi} \frac{\sin^2 t}{R}\, dt$$

Setting these two equations equal to each other (since we want the amount of work in each case to be equal), and making sure the DC side of the equation has π for the amount of time (being the same interval as the AC side), we get this:

$$\left(\frac{V^2}{R}\right) \pi = \int_0^{\pi} \frac{\sin^2 t}{R}\, dt$$

First, we know that R is a constant value, and so we may move it out of the integrand:

$$\left(\frac{V^2}{R}\right) \pi = \frac{1}{R} \int_0^{\pi} \sin^2 t\, dt$$

Multiplying both sides of the equation by R eliminates it completely. This should make intuitive sense, as our RMS equivalent value for a voltage is defined strictly by the ability to produce the same amount of power as the same value of DC voltage for *any* resistance value. Therefore the actual value of resistance (R) should not matter, and it should come as no surprise that it falls out:

$$V^2 \pi = \int_0^{\pi} \sin^2 t\, dt$$

Now, we may simplify the integrand by substituting the half-angle equivalence for the $\sin^2 t$ function

$$V^2 \pi = \int_0^{\pi} \frac{1 - \cos 2t}{2}\, dt$$

Factoring one-half out of the integrand and moving it outside (because it's a constant):

$$V^2 \pi = \frac{1}{2} \int_0^{\pi} 1 - \cos 2t\, dt$$

We may write this as the difference between two integrals, treating each term in the integrand as its own integration problem:

$$V^2\pi = \frac{1}{2}\left(\int_0^{\pi} 1\ dt - \int_0^{\pi} \cos 2t\ dt\right)$$

The second integral may be solved simply by using substitution, with $u = 2t$, $du = 2\ dt$, and $dt = \frac{du}{2}$:

$$V^2\pi = \frac{1}{2}\left(\int_0^{\pi} 1\ dt - \int_0^{\pi} \frac{\cos u}{2}\ du\right)$$

Moving the one-half outside the second integrand:

$$V^2\pi = \frac{1}{2}\left(\int_0^{\pi} 1\ dt - \frac{1}{2}\int_0^{\pi} \cos u\ du\right)$$

Finally, now we can integrate the silly thing:

$$V^2\pi = \frac{1}{2}\left([t]_0^{\pi} - \frac{1}{2}\left[\sin 2t\right]_0^{\pi}\right)$$

$$V^2\pi = \frac{1}{2}\left([\pi - 0] - \frac{1}{2}[\sin 2\pi - \sin 0]\right)$$

$$V^2\pi = \frac{1}{2}\left([\pi - 0] - \frac{1}{2}[0 - 0]\right)$$

$$V^2\pi = \frac{1}{2}(\pi - 0)$$

$$V^2\pi = \frac{1}{2}\pi$$

We can see that π cancels out of both sides:

$$V^2 = \frac{1}{2}$$

Taking the square root of both sides, we arrive at our final answer:

$$V = \frac{1}{\sqrt{2}}$$

So, for a sinusoidal voltage with a peak value of 1 volt, the DC equivalent or "RMS" voltage value would be $\frac{1}{\sqrt{2}}$ volts, or approximately 0.707 volts. In other words, a sinusoidal voltage of 1 volt peak will produce just as much power dissipation at a resistor as a steady DC battery voltage of 0.7071 volts applied to that same resistor. Therefore, this 1 volt peak sine wave may be properly called a 0.7071 volt RMS sine wave, or a 0.7071 volt "DC equivalent" sine wave.

This factor for sinusoidal voltages is quite useful in electrical power system calculations, where the wave-shape of the voltage is nearly always sinusoidal (or very close). In your home, for example, the voltage available at any wall receptacle is 120 volts RMS, which translates to 169.7 volts peak.

Electricians and electronics technicians often memorize the $\frac{1}{\sqrt{2}}$ conversion factor without realizing it only applies to sinusoidal voltage and current waveforms. If we are dealing with a non-sinusoidal wave-shape, the conversion factor between peak and RMS will be different! The mathematical procedure for obtaining the conversion factor will be identical, though: integrate the wave-shape's function (squared) over an interval sufficiently long to capture the essence of the shape, and set that equal to V^2 times that same interval span.

5.2 Resistance, Reactance, and Impedance

Resistance (R) is the dissipative opposition to an electric current, analogous to friction encountered by a moving object. *Reactance* (X) is the opposition to an electric current resulting from energy storage within circuit components, analogous to inertia of a moving object. *Impedance* (Z) is the combined total opposition to an electric current.

Reactance comes in two opposing types: capacitive (X_C) and inductive (X_L). Each one is a function of frequency (f) in an AC circuit:

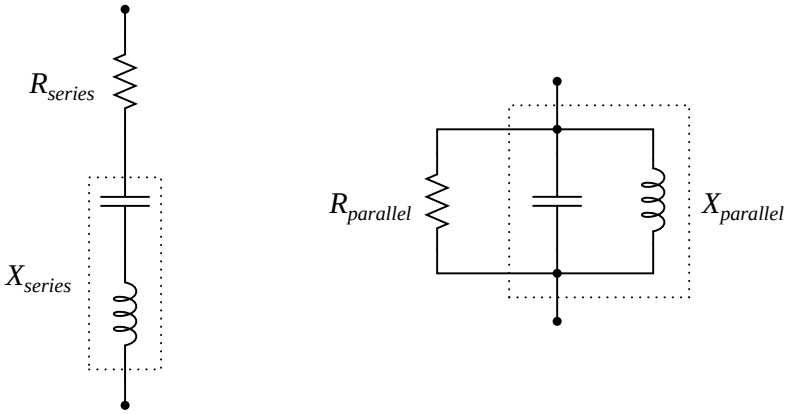
$$X_C = \frac{1}{2\pi f C} \qquad\qquad X_L = 2\pi f L$$

5.3 Series and parallel circuits

Impedance in a series circuit is the orthogonal sum of resistance and reactance:

$$Z = \sqrt{R^2 + (X_L^2 - X_C^2)}$$

Equivalent series and parallel circuits are circuits that have the exact same total impedance as one another, one with series-connected resistance and reactance, and the other with parallel-connected resistance and reactance. The resistance and reactance values of equivalent series and parallel circuits may be expressed in terms of those circuits' total impedance:



If the total impedance of one circuit (either series or parallel) is known, the component values of the equivalent circuit may be found by algebraically manipulating these equations and solving for the desired R and X values:

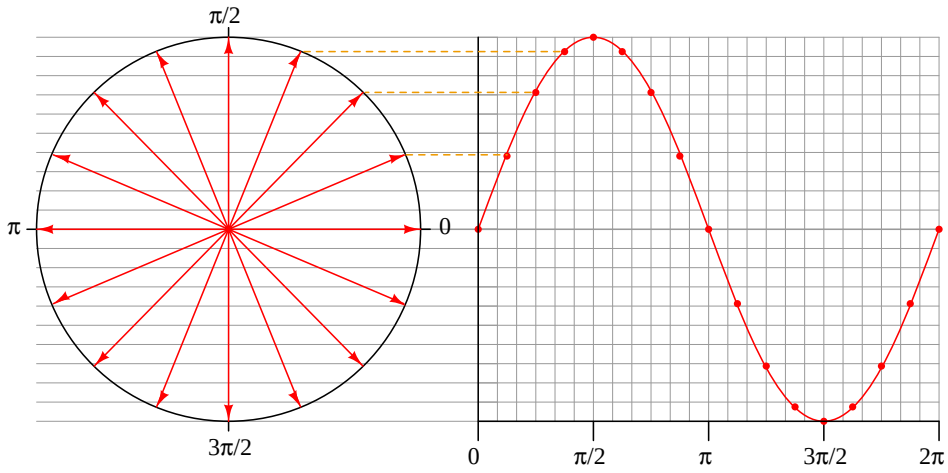
$$Z^2 = R_{series}R_{parallel} \qquad\qquad Z^2 = X_{series}X_{parallel}$$

5.4 Phasor mathematics

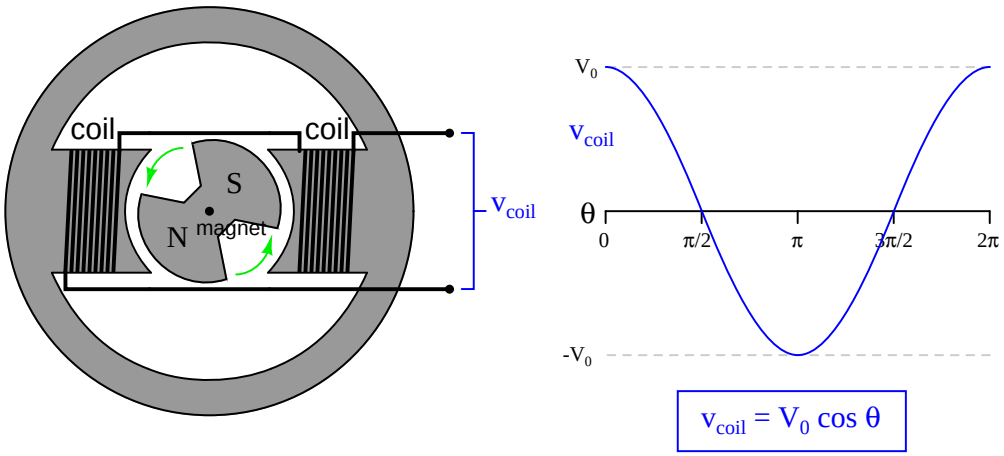
A powerful mathematical technique for analyzing AC circuits is that of *phasors*: representing AC quantities as complex numbers, a "complex" number defined as one having both a "real" and "imaginary" component. The purpose of this section is to explore how complex numbers relate to sinusoidal waveforms, and show some of the mathematical symmetry and beauty of this approach.

5.4.1 Crank diagrams and phase shifts

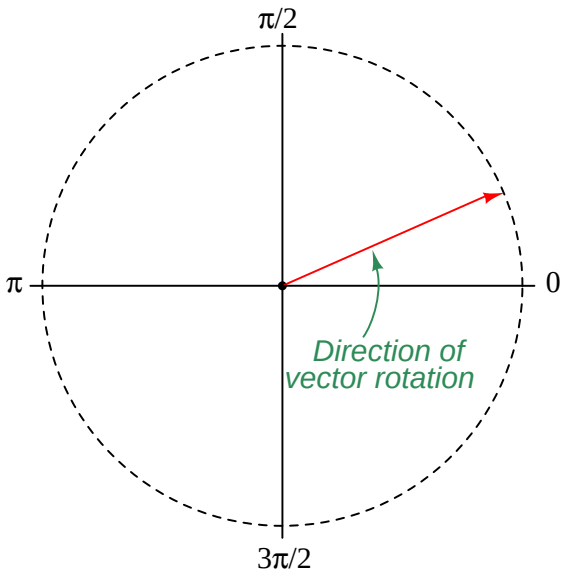
Something every beginning trigonometry student learns (or *should* learn) is how a sine wave is derived from the polar plot of a circle:



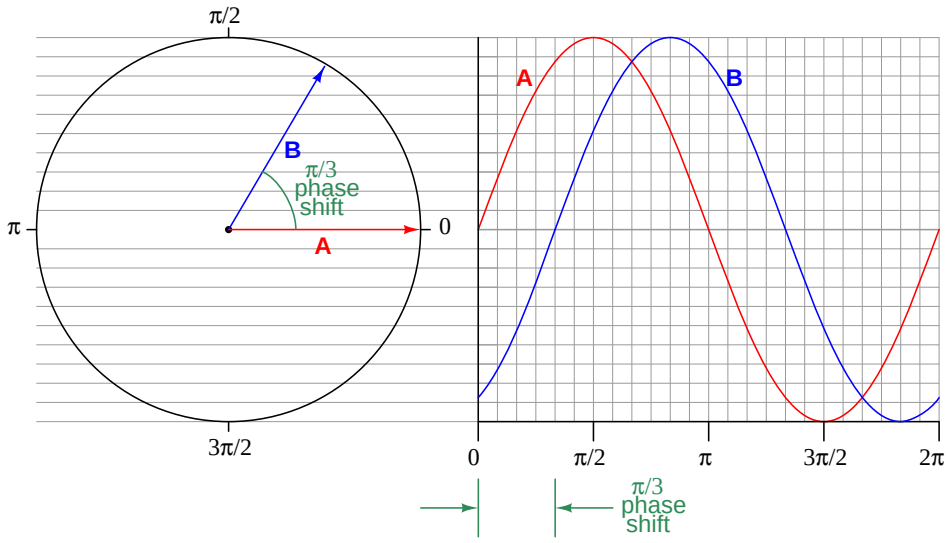
This translation from circular motion to a lengthwise plot has special significance to electrical circuits, because the circular diagram represents how *alternating current* (AC) is generated by a rotating machines, while the lengthwise plot shows how AC is generally displayed on a measuring instrument. The principle of an AC generator is that a magnet is rotated on a shaft past stationary coils of wire. When these wire coils experience the changing magnetic field produced by the rotating magnet, a sinusoidal voltage will be induced in the coils.



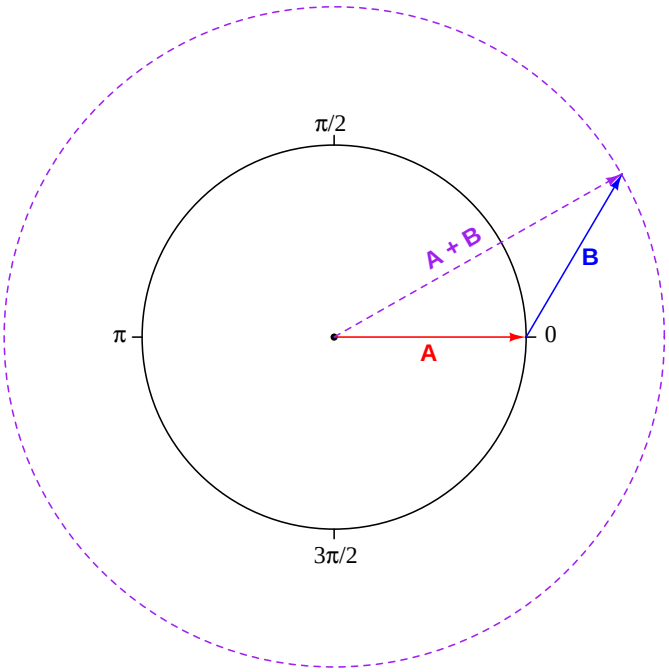
While sine and cosine wave graphs are quite descriptive, there is another type of graph that is even more descriptive for AC circuits: the so-called *crank diagram*. A "crank diagram" represents the sinusoidal wave not as a plot of instantaneous amplitude versus time, but rather as a plot of peak amplitude versus generator shaft angle. This is basically the polar-circular plot seen earlier, which beginning trigonometry students often see near the beginning of their studies:



By representing a sinusoidal voltage as a rotating vector instead of a graph over time, it is easier to see how multiple waveforms will interact with each other. Quite often in alternating-current (AC) circuits, we must deal with voltage waveforms that add with one another by virtue of their sources being connected in series. This sinusoidal addition becomes confusing if the two waveforms are not perfectly in step, which is often the case. However, out-of-step sinusoids are easy to represent and easy to sum when drawn as vectors in a crank diagram. Consider the following example, showing two sinusoidal waveforms, 60 degrees ($\frac{\pi}{3}$ radians) out of step with each other:



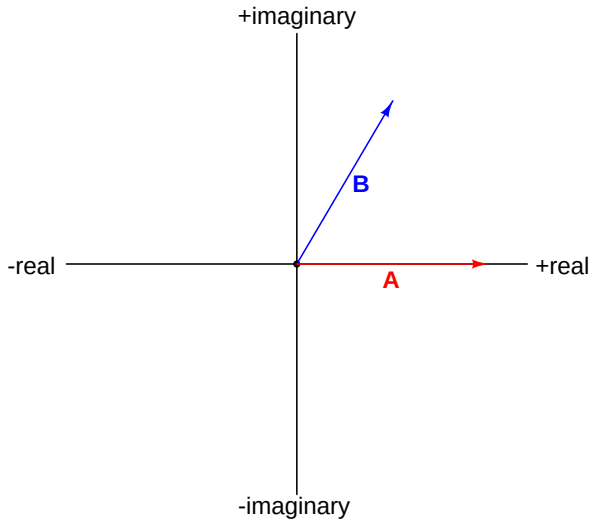
Graphically computing the sum of these two waves would be quite difficult in the standard graph (right-hand side), but it is as easy as stacking vectors tip-to-tail in the crank diagram:



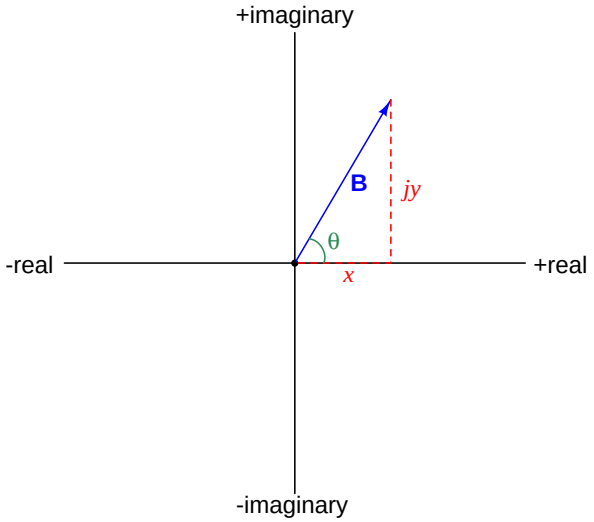
The length of the dashed-line vector $\mathbf{A} + \mathbf{B}$ (radius of the dashed-line circle) represents the amplitude of the resultant sum waveform, while the phase shift is represented by the angles between this new vector and the original vectors $\mathbf{A}$ and $\mathbf{B}$.

5.4.2 Complex numbers and phase shifts

This is all well and good, but we need to have a symbolic means of representing this same information if we are to do any real math with AC voltages and currents. There is one way to do this, if we take the leap of labeling the axes of the "crank diagram" as the axes of a complex plane (real and imaginary numbers):



If we do this, we may symbolically represent each vector as a complex number. For example, vector **B** in the above diagram could be represented as the complex number $x + jy$ (using j as the symbol for an imaginary quantity instead of i so as to not confuse it with *current*):



Vector **A** lies completely on the "real" axis, and so it could be represented as a complex number $x + jy$ where y has a value of zero.

Alternatively, we could express these complex quantities in polar form as amplitudes (A and B) and angles (θ_1 and θ_2), using the cosine and sine functions to translate this amplitude and angle into rectangular terms:

$$A(\cos\theta_1 + j\sin\theta_1)$$

$$B(\cos\theta_2 + j\sin\theta_2)$$

This is where things get really elegant. A Swiss mathematician named Leonhard Euler (1707-1783) proved that sine and cosine functions are related to imaginary powers of e. The following equation is known as *Euler's Relation*, and it is extremely useful for our purposes here:

$$e^{j\theta} = \cos\theta + j\sin\theta$$

With this critical piece of information, we have a truly elegant way to express all the information contained in the crank-diagram vector, in the form of an exponential term:

$$Ae^{j\theta_1} \qquad\qquad Be^{j\theta_2}$$

In other words, these two AC voltages, which are really sinusoidal functions over time, may be symbolized as constant amplitudes A and B (representing the peak voltages of the two waveforms) multiplied by a complex exponential ($e^{j\theta}$), where θ represents the phase of each waveform. What makes this representation very nice is that the complex exponential obeys all the mathematical laws we associate with real exponentials, including the differentiation and integration rules of calculus.

Anyone familiar with calculus knows that exponential functions are extremely easy to differentiate and integrate, which makes calculus operations on AC waveforms *much* easier to determine than if we had to represent AC voltages as trigonometric functions.

Credit for this mathematical application goes to Charles Proteus Steinmetz (1865-1923), the brilliant electrical engineer. At the time, Steinmetz simply referred to this representation of AC waveforms as *vectors*. Now, we assign them the unique name of *phasors* so as to not confuse them with other types of vectors. The term "phasor" is quite appropriate, because the angle of a phasor (θ) represents the *phase shift* between that waveform and a reference waveform.

The notation has become so popular in electrical theory that even students who have never been introduced to Euler's Relation use them. In this case the notation is altered to make it easier to understand. Instead of writing $Be^{j\theta}$, the mathematically innocent electronics student would write $B\angle\theta$.

5.4.3 Phasor expressions of impedance

However, the real purpose of phasors is to make difficult math easier, so this is what we will explore now. Consider the problem of defining electrical opposition to current in an AC circuit. In DC (direct-current) circuits, resistance (R) is defined by Ohm's Law as being the ratio between voltage (V) and current (I):

$$R = \frac{V}{I}$$

There are some electrical components, though, which do not obey Ohm's Law. *Capacitors* and *inductors* are two outstanding examples. The fundamental reason why these two components do not follow Ohm's Law is because they do not dissipate energy like resistances do. Rather than dissipate energy (in the form of heat and/or light), capacitors and inductors *store* and *release* energy from and to the circuit in which they are connected. The contrast between resistors and these components is remarkably similar to the contrast between *friction* and *inertia* in mechanical systems. Whether pushing a flat-bottom box across a floor or pushing a heavy wheeled cart across a floor, work is required to get the object moving. However, the flat-bottom box will immediately stop when you stop pushing it, while the wheeled cart will continue to coast because it has kinetic energy stored in it.

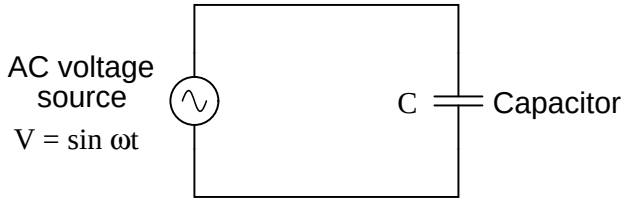
The relationships between voltage and current for capacitors (C) and inductors (L) are as follows:

$$I = C\frac{dV}{dt} \qquad\qquad V = L\frac{dI}{dt}$$

Expressed verbally, capacitors pass electric current proportional to how quickly the voltage across them *changes* over time. Conversely, inductors produce a voltage drop proportional to how quickly current through them *changes* over time. The symmetry here is beautiful: capacitors, which store energy in an electric field, oppose changes in voltage. Inductors, which store energy in a magnetic field, oppose changes in current.

When either type of component is placed in an AC circuit, and subjected to sinusoidal voltages and currents, it will appear to have a "resistance." Given the amplitude of the circuit voltage and the frequency of oscillation (how rapidly the waveforms alternate over time), each type of component will only pass so much current. It would be nice, then, to be able to express the opposition each of these components offers to alternating current in the same way we express the resistance of a resistor in ohms (Ω). To do this, we will have to figure out a way to take the above equations and manipulate them to express each component's behavior as a ratio of $\frac{V}{I}$. I will begin this process by using regular trigonometric functions to represent AC waveforms, then after seeing how ugly this gets I will switch to using phasors and you will see how much easier the math becomes.

Let's start with the capacitor. Suppose we impress an AC voltage across a capacitor as such:



It is common practice to represent the angle of an AC signal as the product ωt rather than as a static angle θ, with ω representing the *angular velocity* of the circuit in radians per second. If a circuit has a ω equal to 2π, it means the generator shaft is making one full rotation every second. Multiplying ω by time t will give the generator's shaft position at that point in time. For example, in the United States our AC power grid operates at 60 cycles per second, or 60 revolutions of our ideal generator every second. This translates into an angular velocity ω of 120π radians per second, or approximately 377 radians per second.

We know that the capacitor's relationship between voltage and current is as follows:

$$I = C\frac{dV}{dt}$$

Therefore, we may substitute the expression for voltage in this circuit into the equation and use calculus to differentiate it with respect to time:

$$I = C\frac{d}{dt}(\sin \omega t)$$

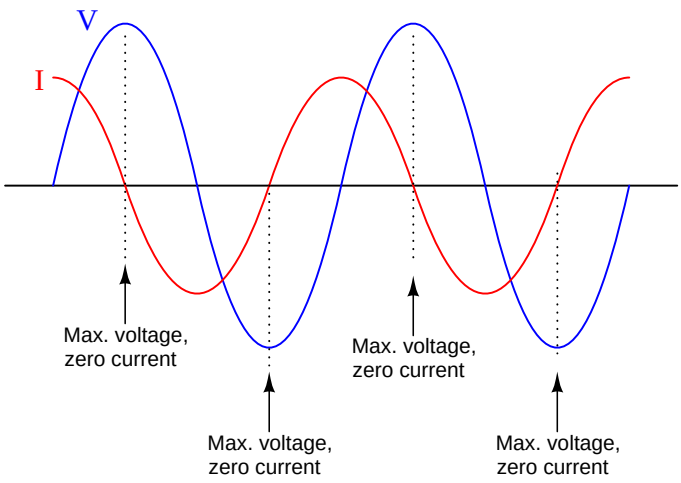
$$I = \omega C(\cos \omega t)$$

The ratio of $\frac{V}{I}$ (the opposition to electric current, analogous to resistance R) will then be:

$$\frac{V}{I} = \frac{\sin \omega t}{\omega C \cos \omega t}$$

$$\frac{V}{I} = \frac{1}{\omega C} \tan \omega t$$

This might look simple enough, until you realize that the ratio $\frac{V}{I}$ will become undefined for certain values of t, notably $\frac{\pi}{2}$ and $\frac{3\pi}{2}$. If we look at a time-domain plot of voltage and current for a capacitor, it becomes clear why this is. There are points in time where voltage is maximum and current is zero:



At these instantaneous points in time, it truly does appear as if the "resistance" of the capacitor is undefined (infinite), with multiple incidents of maximum voltage and zero current. However, this does not capture the essence of what we are trying to do: relate the *peak amplitude* of the voltage with the *peak amplitude* of the current, to see what the ratio of these two peaks are. The ratio calculated here utterly fails because those peaks never happen at the same time.

One way around this problem is to express the voltage as a complex quantity rather than as a scalar quantity. In other words, we use the sine *and* cosine functions to represent what this wave is doing, just like we used the "crank diagram" to represent the voltage as a rotating vector. By doing this, we can represent the waveforms as static amplitudes (vector lengths) rather than as instantaneous quantities that alternately peak and dip over time. The problem with this approach is that the math gets a lot tougher:

$$I = C\frac{dV}{dt} \qquad\qquad V = \cos\omega t + j\sin\omega t$$

$$I = C\frac{d}{dt}(\cos\omega t + j\sin\omega t)$$

$$I = C(-\omega\sin\omega t + j\omega\cos\omega t)$$

$$I = \omega C(-\sin\omega t + j\cos\omega t)$$

$$\frac{V}{I} = \frac{\cos\omega t + j\sin\omega t}{\omega C(-\sin\omega t + j\cos\omega t)}$$

The final result is so ugly no one would want to use it. We may have succeeded in obtaining a ratio of V to I that doesn't blow up at certain values of t, but it provides no practical insight into what the capacitor will really do when placed in the circuit.

Phasors to the rescue! Instead of representing the source voltage as a sum of trig functions ($V = \cos\omega t + j\sin\omega t$), we will use Euler's Relation to represent it as a complex exponential and differentiate from there:

$$I = C\frac{dV}{dt} \qquad\qquad V = e^{j\omega t}$$

$$I = C\frac{d}{dt}\left(e^{j\omega t}\right)$$

$$I = j\omega C e^{j\omega t}$$

$$\frac{V}{I} = \frac{e^{j\omega t}}{j\omega C e^{j\omega t}}$$

$$\frac{V}{I} = \frac{1}{j\omega C} = -j\frac{1}{\omega C}$$

Note how the exponential term completely drops out of the equation, leaving us with a clean ratio strictly in terms of capacitance (C), angular velocity (ω), and of course j. This is the power of phasors: it transforms an ugly math problem into something trivial by comparison.

Next, will will apply this same phasor analysis to inductors. Recall that an inductor's relationship between voltage and current is as follows:

$$V = L\frac{dI}{dt}$$

If an AC current is forced through an inductor (the AC current described by the expression $I = e^{j\omega t}$), we may substitute this expression for current into the inductor's characteristic equation to solve for voltage as a function of time:

$$V = L\frac{dI}{dt} \qquad\qquad I = e^{j\omega t}$$

$$V = L\frac{d}{dt}\left(e^{j\omega t}\right)$$

$$V = j\omega L e^{j\omega t}$$

$$\frac{V}{I} = \frac{j\omega L e^{j\omega t}}{e^{j\omega t}}$$

$$\frac{V}{I} = j\omega L$$

In summary, we may express the impedance (voltage-to-current ratio) of capacitors and inductors by the following equations:

$$Z_L = j\omega L \qquad\qquad Z_C = \frac{1}{j\omega C}$$

Most students familiar with electronics from an algebraic perspective (rather than calculus) find the expressions $X_L = 2\pi f L$ and $X_C = \frac{1}{2\pi f C}$ easier to grasp. Just remember that angular velocity (ω) is really "shorthand" notation for $2\pi f$, so these familiar expressions may be alternatively written as $X_L = \omega L$ and $X_C = \frac{1}{\omega C}$. Furthermore, recall that reactance (X) is a *scalar quantity*, having magnitude but no direction. Impedance (Z), on the other hand, possesses both magnitude *and* direction (phase), which is why the imaginary operator j must appear in the impedance expressions to make them complete. The impedance offered by inductors and capacitors alike are nothing more than their reactance values (X) scaled along the imaginary (j) axis (phase-shifted 90^o).

5.4.4 Euler's Relation and crank diagrams

Another detail of phasor math that is both beautiful and practical is the famous expression of Euler's Relation, the one all math teachers love because it directly relates several fundamental constants in one elegant equation (remember that i and j mean the same thing, just different notational conventions for different disciplines):

$$e^{i\pi} = -1$$

If you understand that this equation is nothing more than the fuller version of Euler's Relation with θ set to the value of π, you may draw a few more practical insights from it:

$$e^{i\theta} = \cos\theta + i\sin\theta$$

$$e^{i\pi} = \cos\pi + i\sin\pi$$

$$e^{i\pi} = -1 + i0$$

$$e^{i\pi} = -1$$

After seeing this, the natural question to ask is what happens when we set θ equal to other, common angles such as 0, $\frac{\pi}{2}$, or $\frac{3\pi}{2}$? The following examples explore these angles:

$$e^{i0} = \cos 0 + i \sin 0$$

$$e^{i0} = 1 + i0$$

$$e^{i0} = 1$$

$$e^{i\frac{\pi}{2}} = \cos\left(\frac{\pi}{2}\right) + i \sin\left(\frac{\pi}{2}\right)$$

$$e^{i\frac{\pi}{2}} = 0 + i1$$

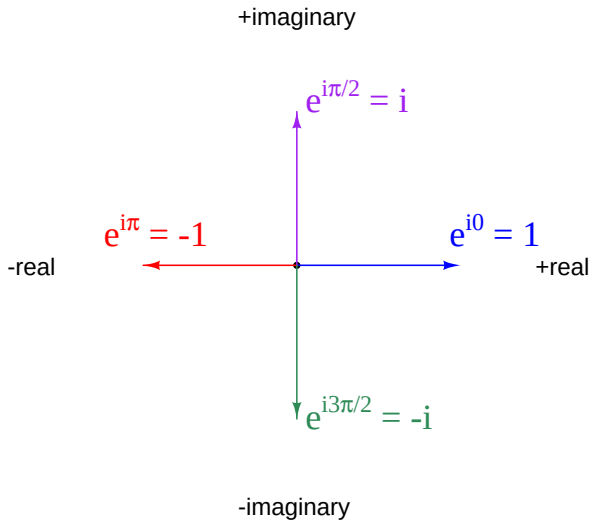
$$e^{i\frac{\pi}{2}} = i$$

$$e^{i\frac{3\pi}{2}} = \cos\left(\frac{3\pi}{2}\right) + i \sin\left(\frac{3\pi}{2}\right)$$

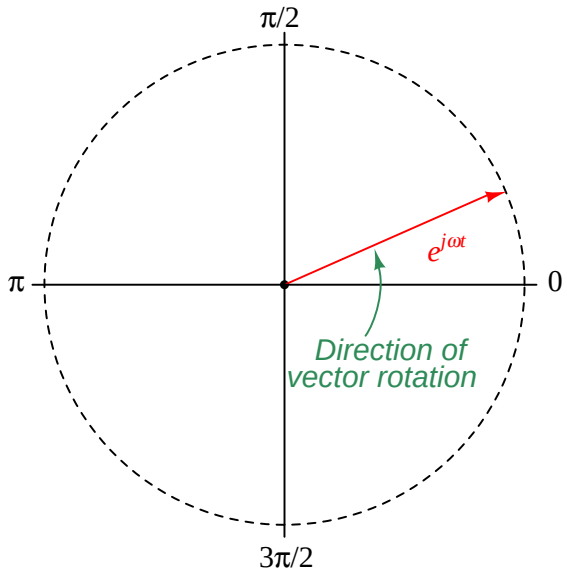
$$e^{i\frac{3\pi}{2}} = 0 - i1$$

$$e^{i\frac{3\pi}{2}} = -i$$

We may show all the equivalencies on the complex plane, as unit vectors:



If we substitute ωt for θ, describing a continually increasing angle rather than a fixed angle, we see our original "crank diagram" come to life, with the vector arrow spinning about the origin of the graph in a counter-clockwise rotation:



Going back to the result we got for the capacitor's opposition to current ($\frac{V}{I}$), we see that we can express the $-i$ term (or $-j$ term, as it is more commonly written in electronics work) as a complex exponential and gain a little more insight:

$$\frac{V}{I} = -j\frac{1}{\omega C}$$

$$\frac{V}{I} = \left(e^{\frac{j3\pi}{2}}\right)\frac{1}{\omega C}$$

What this means is that the capacitor's opposition to current takes the form of a phasor pointing *down* on the complex plane. In other words, it is a phasor with a fixed angle ($\frac{3\pi}{2}$, or $-\frac{\pi}{2}$ radians) rather than rotating around the origin like all the voltage and current phasors do. In electric circuit theory, there is a special name we give to such a quantity, being a ratio of voltage to current, but possessing a complex value. We call this quantity *impedance* rather than *resistance*, and we symbolize it using the letter Z.

When we do this, we arrive at a new form of Ohm's Law for AC circuits:

$$Z = \frac{V}{I} \qquad V = IZ \qquad I = \frac{V}{Z}$$

With all quantities expressed in the form of phasors, we may apply nearly all the rules of DC circuits (Ohm's Law, Kirchhoff's Laws, etc.) to AC circuits. What was old is new again!

5.4.5 The s variable

A concept vital to many forms of engineering is something called a *Laplace transform*. This is a mathematical technique used to convert differential equations (complicated to solve) into algebraic equations (simpler to solve). The subject of Laplace transforms is vast, and requires a solid foundation in calculus to even begin to explore, but one of the elements of Laplace transforms has application right here to our discussion of phasor mathematics and component impedance.

Recall that we may express the impedance (voltage-to-current ratio) of capacitors and inductors by the following equations:

$$Z_L = j\omega L \qquad\qquad Z_C = \frac{1}{j\omega C}$$

The product $j\omega$ keeps appearing again and again in phasor expressions such as these, because $j\omega$ is at the heart of Euler's Relation, where $e^{j\theta} = \cos\theta + j\sin\theta$ (or, where $e^{j\omega} = \cos\omega + j\sin\omega$).

In Laplace transforms, the "imaginary" concept of $j\omega$ is extended by adding a "real" portion to it symbolized by the lower-case Greek letter Sigma (σ). Thus, a complex quantity called s is born:

$$s = \sigma + j\omega$$

We have already seen how $e^{j\omega t}$ may be used to describe the instantaneous amplitude of a sinusoidal waveform with an angular velocity of ω. What would happen if we used s as the Euler exponent instead of $j\omega$? Convention dictates we place a negative sign in the exponential, so the expression will look like this:

$$e^{-st}$$

We may re-write this expression to show the meaning of s, a complex number formed of a real part (σ) and an imaginary part ($j\omega$):

$$e^{-(\sigma+j\omega)t}$$

Distributing the negative sign and the variable for time (t) through the parentheses:

$$e^{-\sigma t - j\omega t}$$

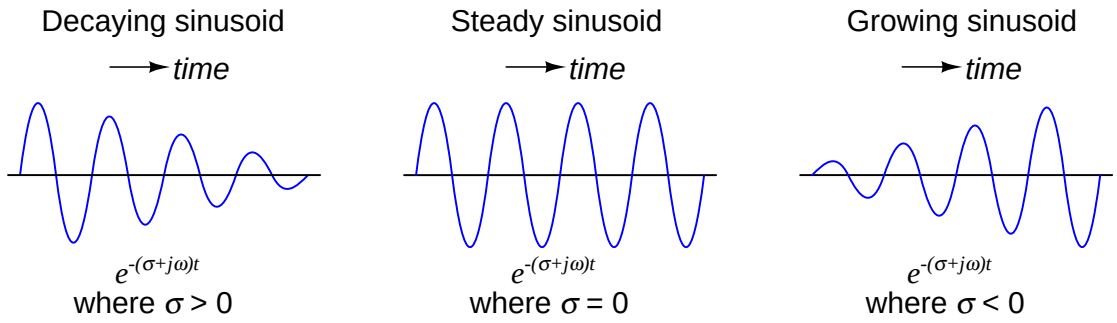
Recalling from the algebraic rules of exponents that a sum (or difference) of exponents is equivalent to a product of exponential terms:

$$e^{-\sigma t}e^{-j\omega t}$$

This expansion is useful for understanding because each of these exponential terms (both $e^{-\sigma t}$ and $e^{-j\omega t}$) have practical meaning on their own. We have already seen how the expression $e^{-j\omega t}$ defines the instantaneous value of a sinusoidal function of angular velocity (frequency) ω at any point in time t. The term $e^{-\sigma t}$, however, is even simpler than this: it defines an exponential growth (or decay) function. The familiar expression $e^{\frac{-t}{\tau}}$ from RC and RL time-constant circuits describing the decay of voltage or current is an example of an exponential decay function. $e^{-\sigma t}$ is just a general expression of this concept, where "sigma" (σ) is the decay constant, equivalent to the reciprocal of the system's "time constant" ($\frac{1}{\tau}$). If the value of sigma is positive ($\sigma > 0$), the expression $e^{-\sigma t}$

describes a process of *decay*, where the value approaches zero over time. Conversely, if the value of sigma is negative ($\sigma < 0$), the expression $e^{-\sigma t}$ describes a process of unbounded *growth*, where the value approaches infinity over time. If sigma happens to be zero ($\sigma = 0$), the value of the expression $e^{-\sigma t}$ will be a constant 1 (neither growing nor decaying over time).

Therefore, when we multiply an exponential growth/decay function ($e^{-\sigma t}$) by a sinusoidal function ($e^{-j\omega t}$), what we get is an expression describing a sinusoidal waveform that either decays, grows, or holds at a steady amplitude over time:



We can see from the expression and from the graph that $e^{-j\omega t}$ is just a special case of e^{-st}, when sigma has a value of zero. Focusing on just the exponent, it is safe to say that $j\omega$ is just a special case of s when there is no exponential growth or decay over time.

For this reason, engineers often substitute s for $j\omega$ in phasor expressions of impedance. So, instead of defining inductor and capacitor impedance in terms of j and ω, they often just define impedance in terms of s:

$$Z_L = sL \quad \text{and} \quad Z_C = \frac{1}{sC} \qquad \text{instead of} \qquad Z_L = j\omega L \quad \text{and} \quad Z_C = \frac{1}{j\omega C}$$

5.5 Transmission lines

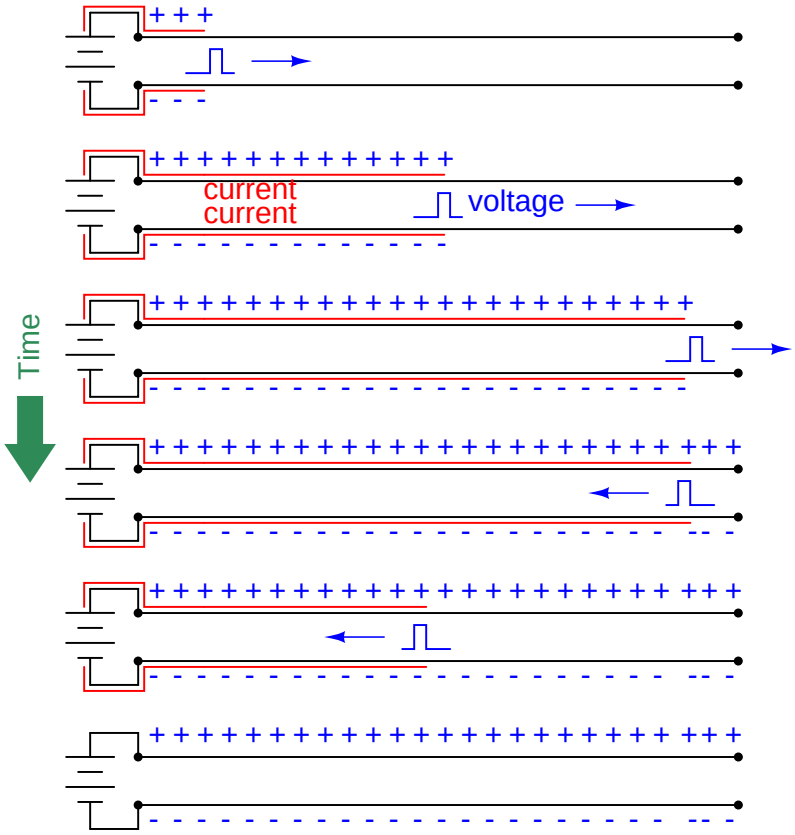
A two-conductor cable conveying an electrical signal whose period is short compared to the propagation time along the cable's length is known as a *transmission line*. In low-frequency and/or physically small circuits, the effects of signal propagation go unnoticed because they are so brief. In circuits where the time delay of signal propagation is significant compared to the period (pulse width) of the signals, however, the effects can be detrimental to circuit function.

When a pulse signal is applied to the beginning of a transmission line, the reactive elements of that cable (i.e. capacitance between the conductors, inductance along the cable length) begin to store energy. This translates to a current draw from the source of the pulse, as though the pulse source were driving a (momentarily) resistive load. If the transmission line happened to be infinitely long, it would behave exactly like a resistor from the perspective of the signal source; i.e. it would never stop "charging."

During the time when a transmission line is absorbing energy from a power source – whether this is indefinitely for a transmission line of infinite length, or momentarily for a transmission line of finite length – the current it draws will be in direct proportion to the voltage applied by the source. In other words, a transmission line behaves like a resistor, at least for a time. The amount of "resistance" presented by a transmission line is called its *characteristic impedance*, or *surge impedance*, symbolized in equations as Z_0.

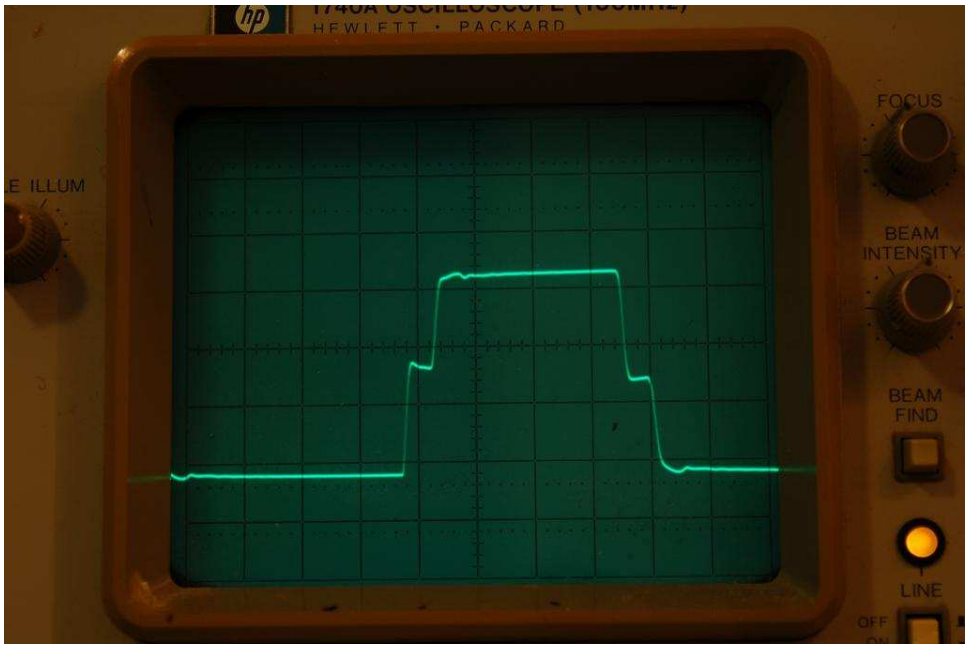
A transmission line's characteristic impedance is a function of its conductor geometry (wire diameter, spacing) and the permittivity of the dielectric separating those conductors. If the line's design is altered to increase its bulk capacitance and/or decrease its bulk inductance (e.g. decreasing the distance between conductors), the characteristic impedance will decrease. Conversely, if the transmission line is altered such that its bulk capacitance decreases and/or its bulk inductance increases, the characteristic impedance will increase. It should be noted that the length of the transmission line has absolutely no bearing on characteristic impedance. A 10-meter length of RG-58/U coaxial cable will have the exact same characteristic impedance as a 10,000 kilometer length of RG-58/U coaxial cable (50 ohms, in both cases). The only difference is the *length of time* the cable will behave like a resistor to an applied voltage.

The following sequence illustrates the propagation of a voltage pulse forward and back (reflected) on an open-ended transmission line:



The end result is a transmission line exhibiting the full source voltage, but no current. This is exactly what we would expect in an open circuit. However, during the time it took for the pulse to travel down the line's length and back, it drew current from the source equal to the source voltage divided by the cable's characteristic impedance ($I_{surge} = \frac{V_{source}}{Z_0}$). For a short amount of time, the two-conductor transmission line acted as a *load* to the voltage source rather than an open circuit.

An experiment performed with a square-wave signal generator and oscilloscope[2] connected to one end of a long wire pair cable (open on the far end) shows the effect of the reflected signal:

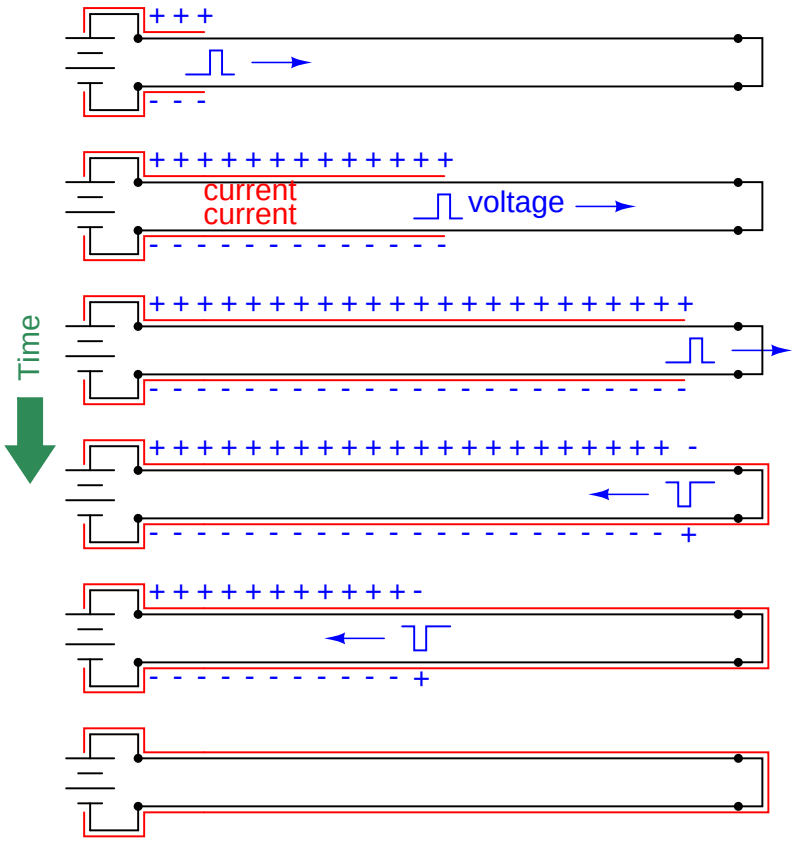


The waveform steps up for a short time, then steps up further to full source voltage. The first step represents the voltage at the source during the time the pulse traveled along the cable's length, when the cable's characteristic impedance acted as a load to the signal generator (making its output voltage "sag" to a value less than its full potential). The next step represents the reflected pulse's return to the signal generator, when the cable's capacitance is fully charged and is no longer drawing current from the signal generator (making its output voltage "rise"). A two-step "fall" appears at the trailing edge of the pulse, when the signal generator reverses polarity and sends an opposing pulse down the cable.

The duration of the first and last "steps" on the waveform represents the time taken by the signal to propagate down the length of the cable *and* return to the source. This oscilloscope's timebase was set to 0.5 microseconds per division for this experiment, indicating a pulse round-trip travel time of approximately 0.2 microseconds. Assuming a velocity factor of 0.7 (70% the speed of light), the round-trip distance calculates to be approximately 42 meters, making the cable 21 meters in length.

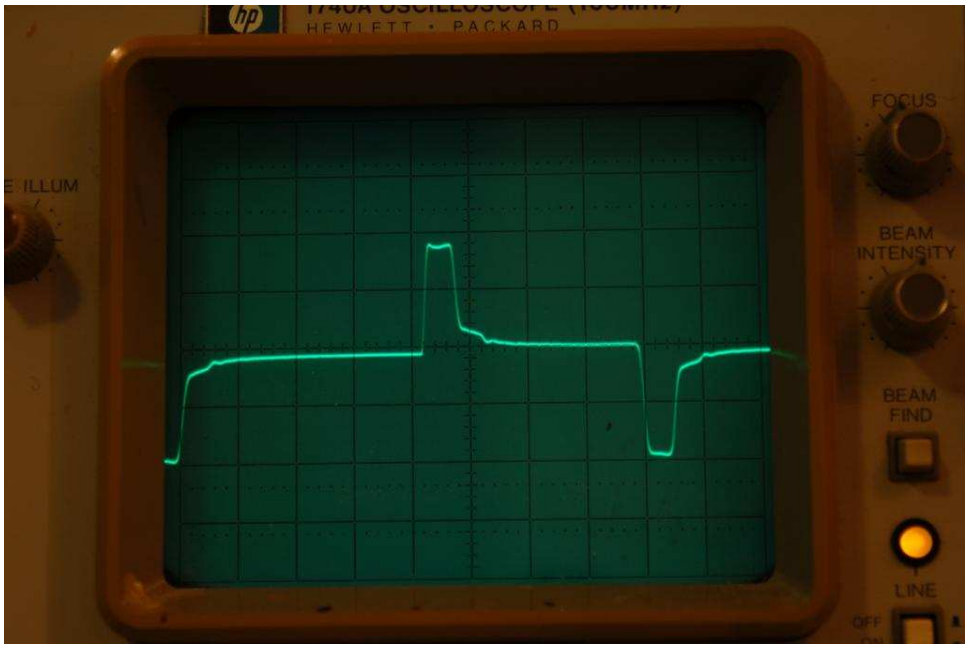
[2] The signal generator was set to a frequency of approximately 240 kHz with a Thévenin resistance of 118 ohms to closely match the cable's characteristic impedance of 120 ohms. The signal amplitude was just over 6 volts peak-to-peak.

The following sequence illustrates the propagation of a voltage pulse forward and back (reflected) on a shorted-end transmission line:



The end result is a transmission line exhibiting the full current of the source ($I_{max} = \frac{V_{source}}{R_{wire}}$), but no voltage. This is exactly what we would expect in a short circuit. However, during the time it took for the pulse to travel down the line's length and back, it drew current from the source equal to the source voltage divided by the cable's characteristic impedance ($I_{surge} = \frac{V_{source}}{Z_0}$). For a short amount of time, the two-conductor transmission line acted as a moderate *load* to the voltage source rather than a direct short.

An experiment performed with the same signal generator and oscilloscope connected to one end of the same long wire pair cable (shorted on the far end) shows the effect of the reflected signal:

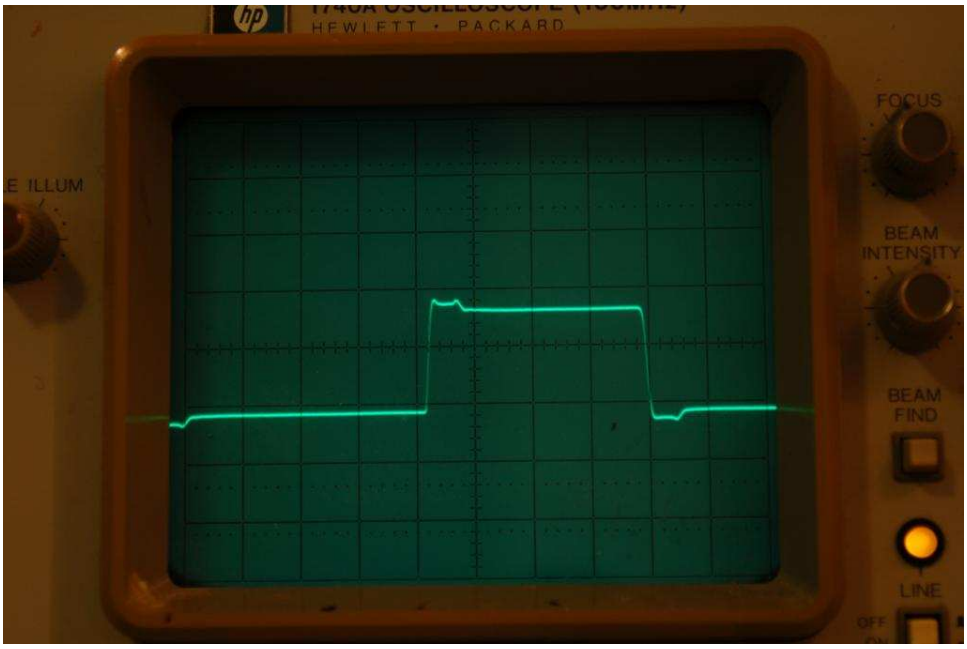


Here, the waveform steps up for a short time, then steps down toward zero. As before, the first step represents the voltage at the source during the time the pulse traveled along the cable's length, when the cable's characteristic impedance acted as a load to the signal generator (making its output voltage "sag" to a value less than its full potential). The step down represents the (inverted) reflected pulse's return to the signal generator, nearly canceling the incident voltage and causing the signal to fall toward zero. A similar pattern appears appears at the trailing edge of the pulse, when the signal generator reverses polarity and sends an opposing pulse down the cable.

Note the duration of the pulse on this waveform, compared to the first and last "steps" on the open-circuited waveform. This pulse width represents the time taken by the signal to propagate down the length of the cable *and* return to the source. This oscilloscope's timebase remained at 0.5 microseconds per division for this experiment as well, indicating the same pulse round-trip travel time of approximately 0.2 microseconds. This stands to reason, as the cable length was not altered between tests; only the type of termination (short versus open).

Proper "termination" of a transmission line consists of connecting a resistance to the end(s) of the line so that the pulse "sees" the exact same amount of impedance at the end as it did while propagating along the line's length. The purpose of the termination resistor is to completely absorb the pulse's energy so that none of it will be reflected back to the source.

Finally, we see the oscilloscope plot of the voltage signal (measured at the source end of the cable) with a resistor of the (nearly) correct value connected to the far end of the cable:



The pulse looks much more like the square wave it should be, now that the cable has been properly terminated[3]. With the termination resistor in place, a transmission line *always* presents the same impedance to the source, no matter what the signal level or the time of signal application. Another way to think of this is from the perspective of cable length. With the proper size of termination resistor in place, *the cable appears infinitely long* from the perspective of the power source because it never reflects any signals back to the source and it always consumes power from the source.

A transmission line's characteristic impedance will be constant throughout its length so long as its conductor geometry and dielectric properties are consistent throughout its length. Abrupt changes in either of these parameters, however, will create a *discontinuity* in the cable capable of producing signal reflections. This is why transmission lines must never be sharply bent, crimped, pinched, twisted, or otherwise deformed.

[3]The termination shown here is imperfect, as evidenced by the irregular amplitude of the square wave. The cable used for this experiment was a length of twin-lead speaker cable, with a characteristic impedance of approximately 120 ohms. I used a 120 ohm (+/- 5%) resistor to terminate the cable, which apparently was not close enough to eliminate all reflections.

The speed at which an electrical signal propagates down a transmission line is never as fast as the speed of light in a vacuum. A value called the *velocity factor* expresses the propagation velocity as a ratio to light, and its value is always less than one:

$$\text{Velocity factor} = \frac{v}{c}$$

Where,

v = Propagation velocity of signal traveling along the transmission line

c = Velocity of light in a vacuum ($\approx 3.0 \times 10^8$ meters per second)

Velocity factor is a function of dielectric constant, but not conductor geometry. A greater permittivity value results in a slower velocity (lesser velocity factor).

Data communication cables for digital instruments behave as transmission lines, and must be terminated at both ends to prevent signal reflections. Reflected signals (or "echoes") may cause errors in received data in a communications network, which is why proper termination can be so important. For point-to-point networks (networks formed by exactly two electronic devices, one at either end of a single cable), the proper termination resistance is often designed into the transmission and receiving circuitry, and so no external resistors need be connected. For "multi-drop" networks where multiple electronic devices tap into the same electrical cable, excessive signal loading would occur if each and every device had its own built-in termination resistance, and so the devices are built with no internal termination, and the installer must place two termination resistors in the network (one at each far end of the cable).

The probe for a guided-wave radar (GWR) level transmitter is another example of a transmission line, one where the vapor/liquid interface creates a discontinuity: there will be an abrupt change in characteristic impedance between the transmission line in vapor space versus the transmission line submerged in a liquid due to the differing dielectric permittivities of the two substances. This sudden change in characteristic impedance sends a reflected signal back to the transmitter. The time delay measured between the signal's transmission and the signal's reception by the transmitter represents the vapor space distance, or *ullage*.

For more detail on the theory and function of radar level measurement, see section 19.5.2 beginning on page 904.

References

Boylestad, Robert L., *Introductory Circuit Analysis*, 9th Edition, Prentice Hall, Upper Saddle River, NJ, 2000.

Kaplan, Wilfred, *Advanced Mathematics for Engineers*, Addison-Wesley Publishing Company, Reading, MA, 1981.

Smith, Steven W., *The Scientist and Engineer's Guide to Digital Signal Processing*, California Technical Publishing, San Diego, CA, 1997.

Steinmetz, Charles P., *Theory and Calculation of Alternating Current Phenomena*, Third Edition, McGraw Publishing Company, New York, NY, 1900.

Chapter 6

Introduction to Industrial Instrumentation

Instrumentation is the science of automated measurement and control. Applications of this science abound in modern research, industry, and everyday living. From automobile engine control systems to home thermostats to aircraft autopilots to the manufacture of pharmaceutical drugs, automation surrounds us. This chapter explains some of the fundamental principles of industrial instrumentation.

The first step, naturally, is measurement. If we can't measure something, it is really pointless to try to control it. This "something" usually takes one of the following forms in industry:

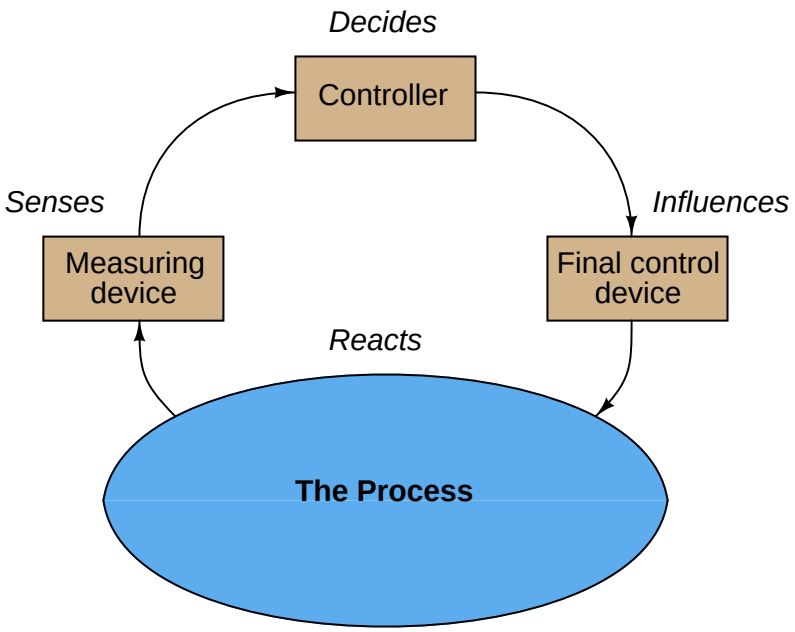
- Fluid pressure
- Fluid flow rate
- The temperature of an object
- Fluid volume stored in a vessel
- Chemical concentration
- Machine position, motion, or acceleration
- Physical dimension(s) of an object
- Count (inventory) of objects
- Electrical voltage, current, or resistance

Once we measure the quantity we are interested in, we usually transmit a signal representing this quantity to an indicating or computing device where either human or automated action then takes place. If the controlling action is automated, the computer sends a signal to a final controlling device which then influences the quantity being measured.

This final control device usually takes one of the following forms:

- Control valve (for throttling the flow rate of a fluid)
- Electric motor
- Electric heater

Both the measurement device and the final control device connect to some physical system which we call the *process*. To show this as a general block diagram:



The common home thermostat is an example of a measurement and control system, with the home's internal air temperature being the "process" under control. In this example, the thermostat usually serves two functions: sensing and control, while the home's heater adds heat to the home to increase temperature, and/or the home's air conditioner extracts heat from the home to decrease temperature. The job of this control system is to maintain air temperature at some comfortable level, with the heater or air conditioner taking action to correct temperature if it strays too far from the desired value (called the *setpoint*).

Industrial measurement and control systems have their own unique terms and standards, which is the primary focus of this lesson. Here are some common instrumentation terms and their definitions:

Process: The physical system we are attempting to control or measure. *Examples: water filtration system, molten metal casting system, steam boiler, oil refinery unit, power generation unit.*

Process Variable, or **PV**: The specific quantity we are measuring in a process. *Examples: pressure, level, temperature, flow, electrical conductivity, pH, position, speed, vibration.*

Setpoint, or **SP**: The value at which we desire the process variable to be maintained at. In other words, the "target" value of the process variable.

Primary Sensing Element, or **PSE**: A device that directly senses the process variable and translates that sensed quantity into an analog representation (electrical voltage, current, resistance; mechanical force, motion, etc.). *Examples: thermocouple, thermistor, bourdon tube, microphone, potentiometer, electrochemical cell, accelerometer.*

Transducer: A device that converts one standardized instrumentation signal into another standardized instrumentation signal, and/or performs some sort of processing on that signal. Often referred to as a *converter* and sometimes as a "relay." *Examples: I/P converter (converts 4-20 mA electric signal into 3-15 PSI pneumatic signal), P/I converter (converts 3-15 PSI pneumatic signal into 4-20 mA electric signal), square-root extractor (calculates the square root of the input signal).*

Note: in general science parlance, a "transducer" is any device that converts one form of energy into another, such as a microphone or a thermocouple. In industrial instrumentation, however, we generally use "primary sensing element" to describe this concept and reserve the word "transducer" to specifically refer to a conversion device for standardized instrumentation signals.

Transmitter: A device that translates the signal produced by a primary sensing element (PSE) into a *standardized* instrumentation signal such as 3-15 PSI air pressure, 4-20 mA DC electric current, Fieldbus digital signal packet, etc., which may then be conveyed to an indicating device, a controlling device, or both.

Lower- and Upper-range values, abbreviated **LRV** and **URV**, respectively: the values of process measurement deemed to be 0% and 100% of a transmitter's calibrated range. For example, if a temperature transmitter is calibrated to measure a range of temperature starting at 300 degrees Celsius and ending at 500 degrees Celsius, 300 degrees would be the LRV and 500 degrees would be the URV.

Zero and **Span**: alternative descriptions to LRV and URV for the 0% and 100% points of an instrument's calibrated range. "Zero" refers to the beginning-point of an instrument's range (equivalent to LRV), while "span" refers to the width of its range (URV − LRV). For example, if a temperature transmitter is calibrated to measure a range of temperature starting at 300 degrees Celsius and ending at 500 degrees Celsius, its zero would be 300 degrees and its span would be 200 degrees.

Controller: A device that receives a process variable (PV) signal from a primary sensing element (PSE) or transmitter, compares that signal to the desired value for that process variable (called the setpoint), and calculates an appropriate output signal value to be sent to a final control element (FCE) such as an electric motor or control valve.

Final Control Element, or **FCE**: A device that receives the signal from a controller to directly influence the process. *Examples: variable-speed electric motor, control valve, electric heater.*

Manipulated Variable, or **MV**: Another term to describe the output signal generated by a controller. This is the signal commanding ("manipulating") the final control element to influence the process.

Automatic mode: When the controller generates an output signal based on the relationship of process variable (PV) to the setpoint (SP).

Manual mode: When the controller's decision-making ability is bypassed to let a human operator directly determine the output signal sent to the final control element.

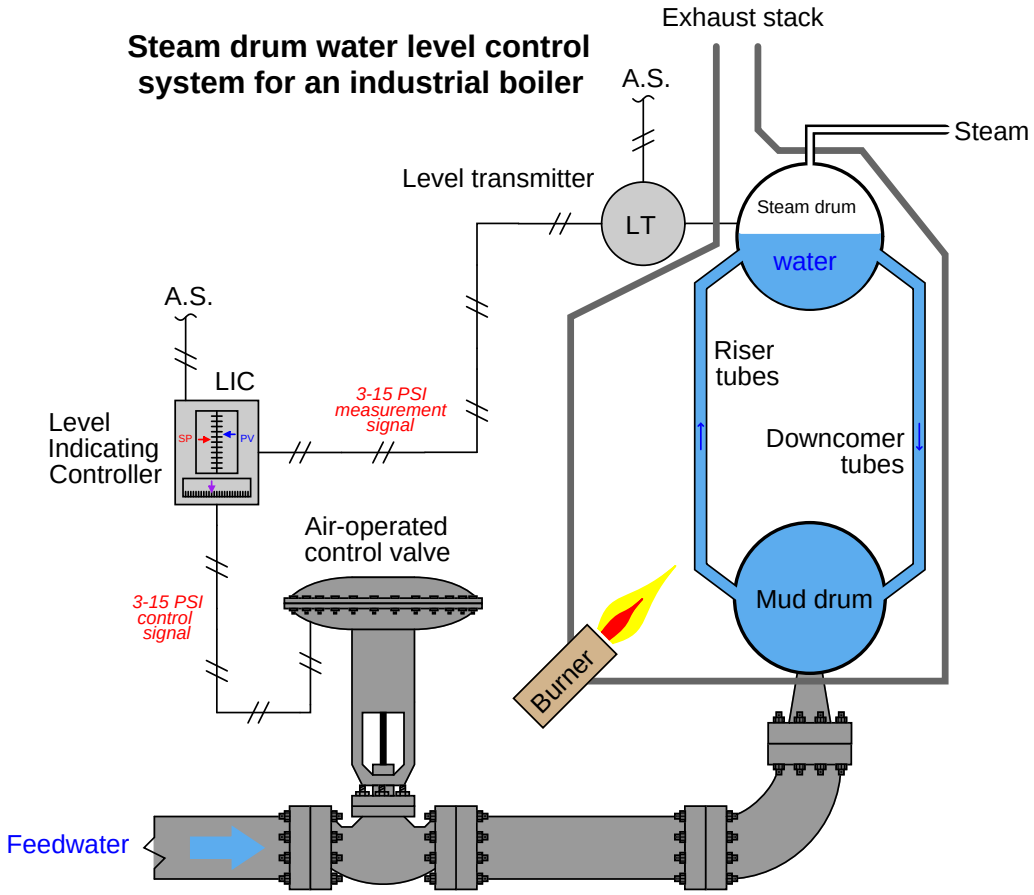
Now I will show you some practical examples of measurement and control systems so you can get a better idea of these fundamental concepts.

6.1 Example: boiler water level control system

Steam boilers are very common in industry, principally because steam power is so useful. Common uses for steam in industry include doing mechanical work (e.g. a steam engine moving some sort of machine), heating, producing vacuums (through the use of "steam eductors"), and augmenting chemical processes (e.g. reforming of natural gas into hydrogen and carbon dioxide).

The process of converting water into steam is quite simple: heat up the water until it boils. Anyone who has ever boiled a pot of water for cooking knows how this process works. Making steam continuously, however, is a little more complicated. An important variable to measure and control in a continuous boiler is the level of water in the "steam drum" (the upper vessel in a water-tube boiler). In order to safely and efficiently produce a continuous flow of steam, we must ensure the steam drum never runs too low on water, or too high. If there is not enough water in the drum, the water tubes may run dry and burn through from the heat of the fire. If there is too much water in the drum, liquid water may be carried along with the flow of steam, causing problems downstream.

In this next illustration, you can see the essential elements of a water level control system, showing transmitter, controller, and control valve:



The first instrument in this control system is the *level transmitter*, or "LT". The purpose of this device is to sense the water level in the steam drum and report that measurement to the controller in the form of an instrument signal. In this case, the type of signal is *pneumatic*: a variable air pressure sent through metal or plastic tubes. The greater the water level in the drum, the more air pressure output by the level transmitter. Since the transmitter is pneumatic, it must be supplied with a source of clean, compressed air on which to run. This is the meaning of the "A.S." tube (Air Supply) entering the top of the transmitter.

This pneumatic signal is sent to the next instrument in the control system, the *level indicating controller*, or "LIC". The purpose of this instrument is to compare the level transmitter's signal with a *setpoint* value entered by a human operator (the desired water level in the steam drum). The controller then generates an *output* signal telling the control valve to either introduce more or less water into the boiler to maintain the steam drum water level at setpoint. As with the transmitter, the controller in this system is pneumatic, operating entirely on compressed air. This means the

output of the controller is also a variable air pressure signal, just like the signal output by the level transmitter. Naturally, the controller requires a constant supply of clean, compressed air on which to run, which explains the "A.S." (Air Supply) tube connecting to it.

The last instrument in this control system is the control valve, being operated directly by the air pressure signal generated by the controller. This particular control valve uses a large diaphragm to convert the air pressure signal into a mechanical force to move the valve open and closed. A large spring inside the valve mechanism provides the force necessary to return the valve to its normal position, while the force generated by the air pressure on the diaphragm works against the spring to move the valve the other direction.

When the controller is placed in the "automatic" mode, it will move the control valve to whatever position it needs to be in order to maintain a constant steam drum water level. The phrase "whatever position it needs to be" suggests that the relationship between the controller output signal, the process variable signal (PV), and the setpoint (SP) can be quite complex. If the controller senses a water level above setpoint, it will take whatever action is necessary to bring that level back down to setpoint. Conversely, if the controller senses a water level below setpoint, it will take whatever action is necessary to bring that level up to setpoint. What this means in a practical sense is that the controller's output signal (equating to valve position) is just as much a function of process load (i.e. how much steam is being used from the boiler) as it is a function of setpoint.

Consider a situation where the steam demand from the boiler is very low. If there isn't much steam being drawn off the boiler, this means there will be little water boiled into steam and therefore little need for additional feedwater to be pumped into the boiler. Therefore, in this situation, one would expect the control valve to hover near the fully-closed position, allowing just enough water into the boiler to keep the steam drum water level at setpoint.

If, however, there is great demand for steam from this boiler, the rate of evaporation will be much higher. This means the control system will have to add feedwater to the boiler at a much greater flow rate in order to maintain the steam drum water level at setpoint. In this situation we would expect to see the control valve much closer to being fully-open as the control system "works harder" to maintain a constant water level in the steam drum.

A human operator running this boiler has the option of placing the controller into "manual" mode. In this mode, the control valve position is under direct control of the human operator, with the controller essentially ignoring the signal sent from the water level transmitter. Being an indicating controller, the controller faceplate will still show how much water is in the steam drum, but it is now the human operator's sole responsibility to move the control valve to the appropriate position to hold water level at setpoint.

Manual mode is useful to the human operator(s) during start-up and shut-down conditions. It is also useful to the instrument technician for troubleshooting a misbehaving control system. When a controller is in automatic mode, the output signal (sent to the control valve) changes in response to the process variable (PV) and setpoint (SP) values. Changes in the control valve position, in turn, naturally affect the process variable signal through the physical relationships of the process. What we have here is a situation where causality is uncertain. If we see the process variable changing erratically over time, does this mean we have a faulty transmitter (outputting an erratic signal), or does it mean the controller output is erratic (causing the control valve to shift position unnecessarily), or does it mean the steam demand is fluctuating and causing the water level to vary as a result? So long as the controller remains in automatic mode, we can never be completely sure what is

causing what to happen, because the chain of causality is actually a *loop*, with everything affecting everything else in the system.

A simple way to diagnose such a problem is to place the controller in manual mode. Now the output signal to the control valve will be fixed at whatever level the human operator or technician sets it to. If we see the process variable signal suddenly stabilize, we know the problem has something to do with the controller output. If we see the process variable signal suddenly become even more erratic once we place the controller in manual mode, we know the controller was actually trying to do its job properly in automatic mode and the cause of the problem lies within the process itself.

As was mentioned before, this is an example of a *pneumatic* (compressed air) control system, where all the instruments operate on compressed air, and use compressed air as the signaling medium. Pneumatic instrumentation is an old technology, dating back many decades. While most modern instruments are electronic in nature, pneumatic instruments still find application within industry. The most common industry standard for pneumatic pressure signals is 3 to 15 PSI, with 3 PSI representing low end-of-scale and 15 PSI representing high end-of-scale. The following table shows the meaning of different signal pressures are they relate to the level transmitter's output:

Transmitter air signal pressure	Steam drum water level
3 PSI	0% (Empty)
6 PSI	25%
9 PSI	50%
12 PSI	75%
15 PSI	100% (Full)

Likewise, the controller's pneumatic output signal to the control valve uses the same 3 to 15 PSI standard to command different valve positions:

Controller output signal pressure	Control valve position
3 PSI	0% open (Fully shut)
6 PSI	25% open
9 PSI	50% open
12 PSI	75% open
15 PSI	100% (Fully open)

It should be noted the previously shown transmitter calibration table assumes the transmitter measures the *full range* of water level possible in the drum. Usually, this is not the case. Instead, the transmitter will be calibrated so it only senses a narrow range of water level near the middle of the drum. Thus, 3 PSI (0%) will not represent an empty drum, and neither will 15 PSI (100%) represent a completely full drum. Calibrating the transmitter like this helps avoid the possibility of actually running the drum completely empty or completely full in the case of an operator incorrectly setting the setpoint value near either extreme end of the measurement scale.

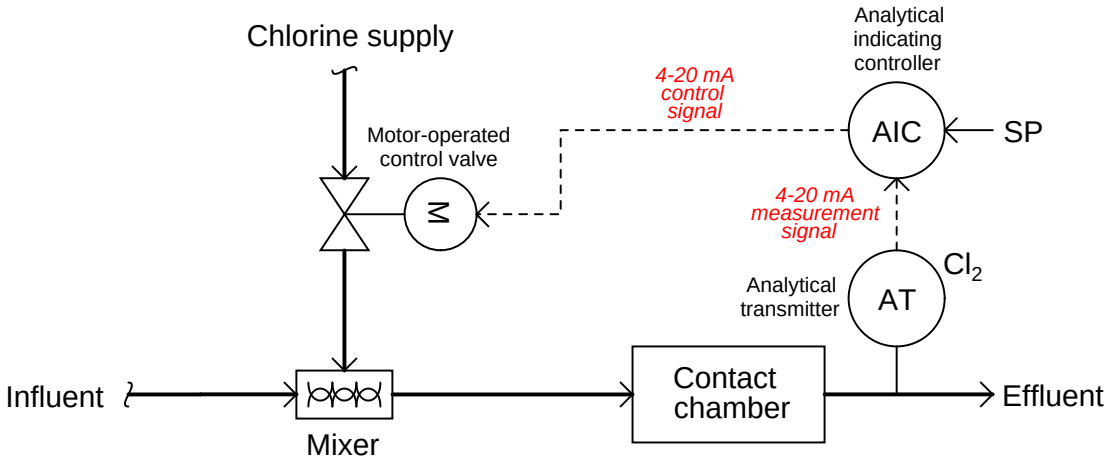
An example table showing this kind of realistic transmitter calibration is shown here:

Transmitter air signal pressure	Actual steam drum water level
3 PSI	40%
6 PSI	45%
9 PSI	50%
12 PSI	55%
15 PSI	60%

6.2 Example: wastewater disinfection

The final step in treating wastewater before releasing it into the natural environment is to kill any harmful bacteria in it. This is called *disinfection*, and chlorine gas is a very effective disinfecting agent. However, just as it is not good to mix too little chlorine in the outgoing water (effluent) because we might not disinfect the water thoroughly enough, there is also danger of injecting too much chlorine in the effluent because then we might begin poisoning animals and beneficial microorganisms in the natural environment.

To ensure the right amount of chlorine injection, we must use a dissolved chlorine analyzer to measure the chlorine concentration in the effluent, and use a controller to automatically adjust the chlorine control valve to inject the right amount of chlorine at all times. The following P&ID (Process and Instrument Diagram) shows how such a control system might look:



Chlorine gas coming through the control valve mixes with the incoming water (influent), then has time to disinfect in the contact chamber before exiting out to the environment.

The transmitter is labeled "AT" (Analytical Transmitter) because its function is to *analyze* the concentration of chlorine dissolved in the water and *transmit* this information to the control system. The "Cl_2" (chemical notation for a chlorine molecule) written near the transmitter bubble declares this to be a chlorine analyzer. The dashed line coming out of the transmitter tells us the signal is electronic in nature, not pneumatic as was the case in the previous (boiler control system) example. The most common and likely standard for electronic signaling in industry is 4 to 20 milliamps DC, which represents chlorine concentration in much the same way as the 3 to 15 PSI pneumatic signal standard represented steam drum water level in the previous system:

Transmitter signal current	Chlorine concentration
4 mA	0% (no chlorine)
8 mA	25%
12 mA	50%
16 mA	75%
20 mA	100% (Full concentration)

The controller is labeled "AIC" because it is an Analytical Indicating Controller. Controllers are always designated by the process variable they are charged with controlling, in this case the chlorine analysis of the effluent. "Indicating" means there is some form of display that a human operator or technician can read showing the chlorine concentration. "SP" refers to the setpoint value entered by the operator, which the controller tries to maintain by adjusting the position of the chlorine injection valve.

A dashed line going from the controller to the valve indicates another electronic signal, most likely 4 to 20 mA DC again. Just as with the 3 to 15 PSI pneumatic signal standard in the pneumatic boiler control system, the amount of electric current in this signal path directly relates to a certain valve position:

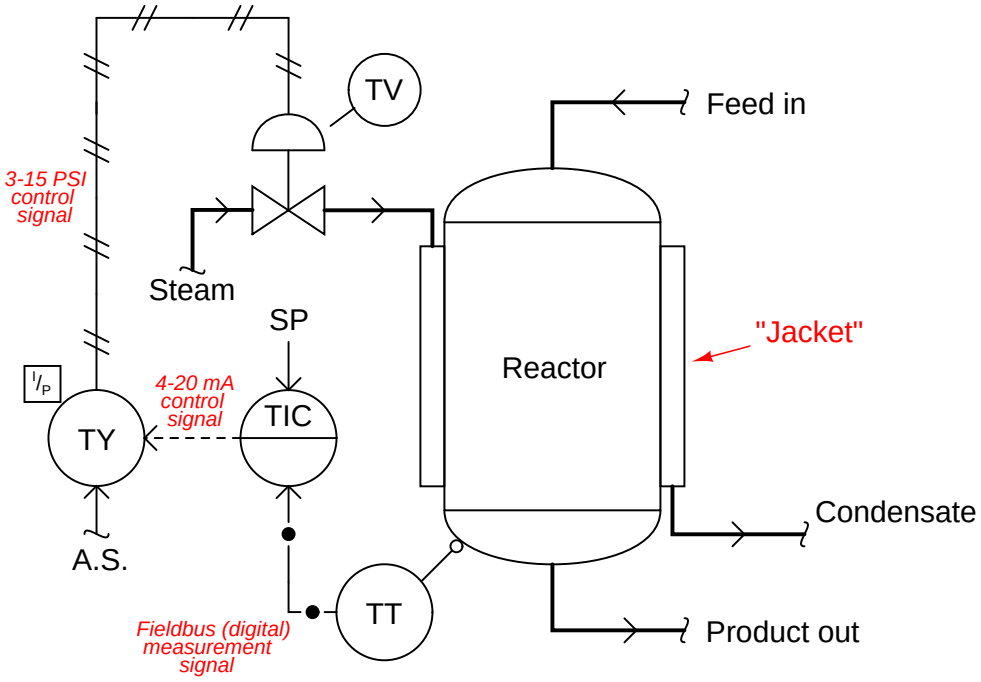
Controller output signal current	Control valve position
4 mA	0% open (Fully shut)
8 mA	25% open
12 mA	50% open
16 mA	75% open
20 mA	100% (Fully open)

Note: it is possible, and in some cases even preferable, to have either a transmitter or a control valve that responds in reverse fashion to an instrument signal such as 3 to 15 PSI or 4 to 20 milliamps. For example, this valve could have been set up to be wide open at 4 mA and fully shut at 20 mA. The main point to recognize here is that both the process variable sensed by the transmitter and the position of the control valve are proportionately represented by an analog signal.

The letter "M" inside the control valve bubble tells us this is a motor-actuated valve. Instead of using compressed air pushing against a spring-loaded diaphragm as was the case in the boiler control system, this valve is actuated by an electric motor turning a gear-reduction mechanism. The gear reduction mechanism allows slow motion of the control valve stem even though the motor spins at a fast rate. A special electronic control circuit inside the valve actuator modulates electric power to the electric motor in order to ensure the valve position accurately matches the signal sent by the controller. In effect, this is another control system in itself, controlling valve position according to a "setpoint" signal sent by another device (in this case, the AIT controller which is telling the valve what position to go to).

6.3 Example: chemical reactor temperature control

Sometimes we see a mix of instrument signal standards in one control system. Such is the case for this particular chemical reactor temperature control system, where three different signal standards are used to convey information between the instruments. A P&ID (Process and Instrument Diagram) shows the inter-relationships of the process piping, vessels, and instruments:



The purpose of this control system is to ensure the chemical solution inside the reactor vessel is maintained at a constant temperature. A steam-heated "jacket" envelops the reactor vessel, transferring heat from the steam into the chemical solution inside. The control system maintains a constant temperature by measuring the temperature of the reactor vessel, and throttling steam from a boiler to the steam jacket to add more or less heat as needed.

We begin as usual with the temperature transmitter, located near the bottom of the vessel. Note the different line type used to connect the temperature transmitter (TT) with the temperature-indicating controller (TIC): solid dots with lines in between. This signifies a *digital electronic instrument signal* – sometimes referred to as a *fieldbus* – rather than an analog type (such as 4 to 20 mA or 3 to 15 PSI). The transmitter in this system is actually a computer, and so is the controller. The transmitter reports the process variable (reactor temperature) to the controller using digital bits of information. Here there is no analog scale of 4 to 20 milliamps, but rather electric voltage/current pulses representing the 0 and 1 states of binary data.

Digital instrument signals are not only capable of transferring simple process data, but they can also convey device status information (such as self-diagnostic test results). In other words, the

digital signal coming from this transmitter not only tells the controller how hot the reactor is, but it can also tell the controller how well the transmitter is functioning!

The dashed line exiting the controller shows it to be analog electronic: most likely 4 to 20 milliamps DC. This electronic signal does not go directly to the control valve, however. It passes through a device labeled "TY", which is a *transducer* to convert the 4 to 20 mA electronic signal into a 3 to 15 PSI pneumatic signal which then actuates the valve. In essence, this signal transducer acts as an electrically-controlled air pressure regulator, taking the supply air pressure (usually 20 to 25 PSI) and regulating it down to a level commanded by the controller's electronic output signal.

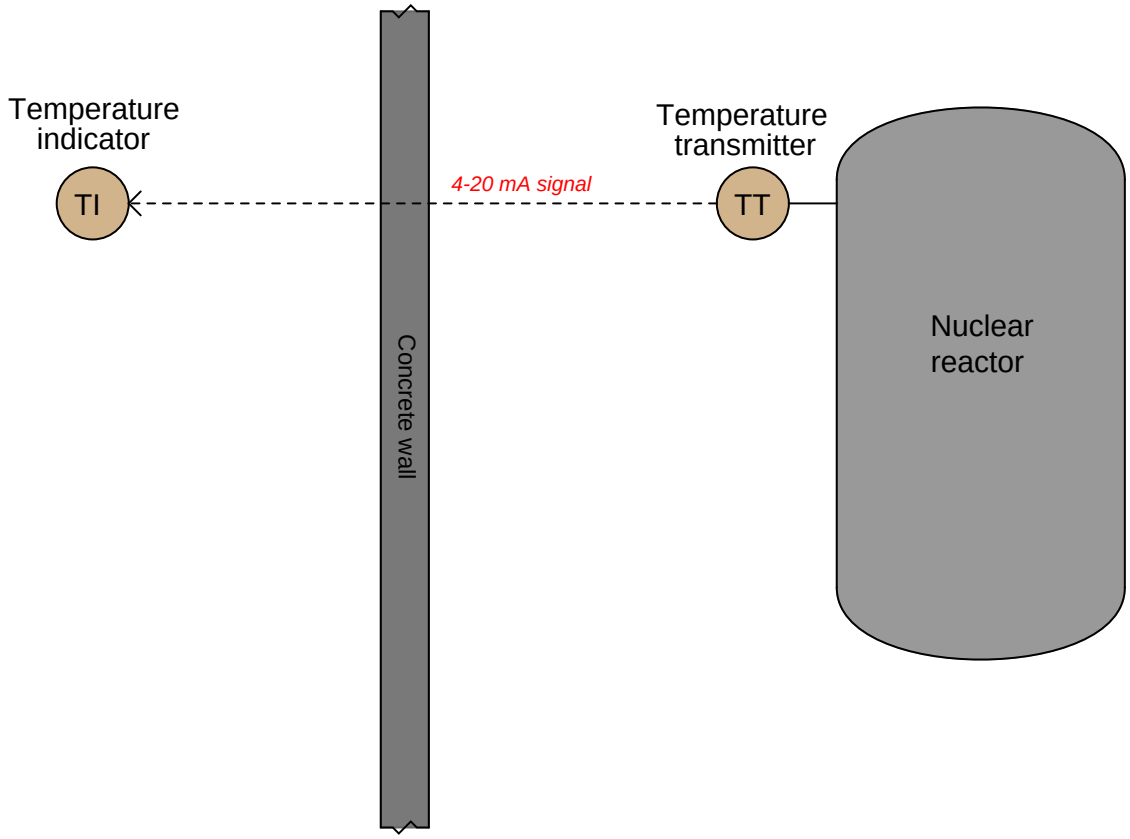
At the temperature control valve (TV) the 3 to 15 PSI pneumatic pressure signal applies a force on a diaphragm to move the valve mechanism against the restraining force of a large spring. The construction and operation of this valve is the same as for the feedwater valve in the pneumatic boiler water control system.

6.4 Other types of instruments

So far we have just looked at instruments that sense, control, and influence process variables. Transmitters, controllers, and control valves are respective examples of each instrument type. However, other instruments exist to perform useful functions for us.

6.4.1 Indicators

One common "auxiliary" instrument is the *indicator*, the purpose of which is to provide a human-readable indication of an instrument signal. Quite often process transmitters are not equipped with readouts for whatever variable they measure: they just transmit a standard instrument signal (3 to 15 PSI, 4 to 20 mA, etc.) to another device. An indicator gives a human operator a convenient way of seeing what the output of the transmitter is without having to connect test equipment (pressure gauge for 3-15 PSI, ammeter for 4-20 mA) and perform conversion calculations. Moreover, indicators may be located far from their respective transmitters, providing readouts in locations more convenient than the location of the transmitter itself. An example where remote indication would be practical is shown here, in a nuclear reactor temperature measurement system:



No human can survive near the nuclear reactor when it is in full-power operation, due to the strong radiation flux it emits. The temperature transmitter is built to withstand the radiation, though, and it transmits a 4 to 20 milliamp electronic signal to an indicating recorder located outside of the containment building where it is safe for a human operator to be. There is nothing preventing us from connecting multiple indicators, at multiple locations, to the same 4 to 20 milliamp signal wires coming from the temperature transmitter. This allows us to display the reactor temperature

in as many locations as we desire, since there is no absolute limitation on how far we may conduct a DC milliamp signal along copper wires.

A numerical and bargraph panel-mounted indicator appears in this next photograph:

This particular indicator, manufactured by Weschler, shows the position of a flow-control gate in a wastewater treatment facility, both by numerical value (98.06%) and by the height of a bargraph (very near full open – 100%).

A less sophisticated style of panel-mounted indicator shows only a numeric display, such as this Red Lion Controls model shown here:

Indicators may also be used in "field" (process) areas to provide direct indication of measured variables if the transmitter device lacks a human-readable indicator of its own. The following photograph shows a Rosemount brand field-mounted indicators, operating directly from the electrical power available in the 4-20 mA loop:

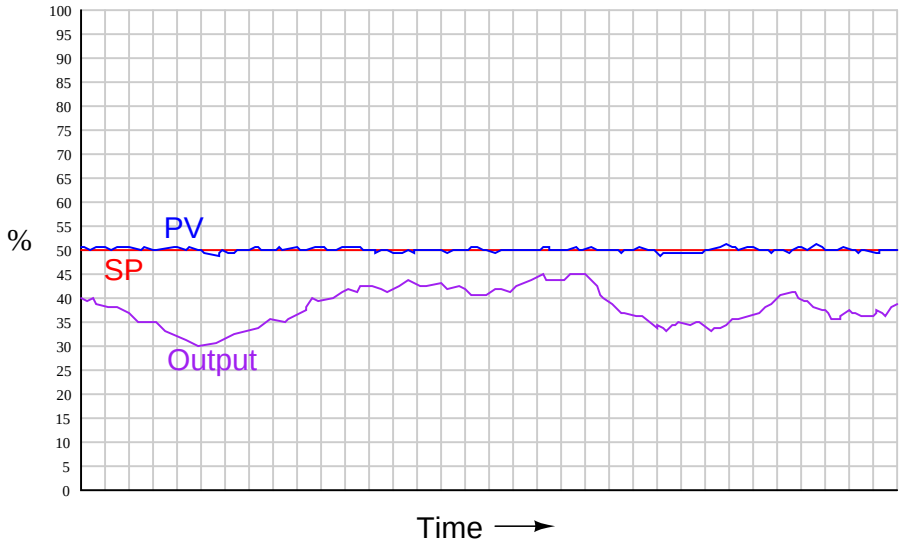
6.4.2 Recorders

Another common "auxiliary" instrument is the *recorder* (sometimes specifically referred to as a *chart recorder* or a *trend recorder*), the purpose of which is to draw a graph of process variable(s) over time. Recorders usually have indications built into them for showing the instantaneous value of the instrument signal(s) simultaneously with the historical values, and for this reason are usually designated as *indicating* recorders. A temperature indicating recorder for the nuclear reactor system shown previously would be designated as a "TIR" accordingly.

A *circular* chart recorder uses a round sheet of paper, rotated beneath a pen moved side-to-side by a servomechanism driven by the instrument signal. Two such chart recorders are shown in the following photograph:

Two more chart recorders appear in the next photograph, a *strip* chart recorder on the right and a *paperless* chart recorder on the left. The strip chart recorder uses a scroll of paper drawn past one or more lateral-moving pens, while the paperless recorder does away with paper entirely by drawing graphic trend lines on a computer screen:

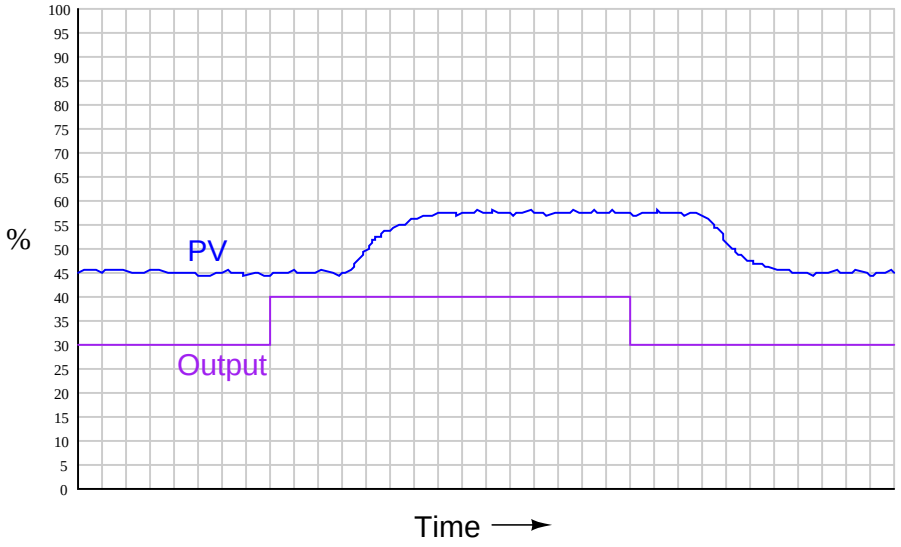
Recorders are extremely helpful for troubleshooting process control problems. This is especially true when the recorder is configured to record not just the process variable, but also the controller's setpoint and output variables as well. Here is an example of a typical "trend" showing the relationship between process variable, setpoint, and controller output in automatic mode, as graphed by a recorder:



Here, the setpoint (SP) appears as a perfectly straight (red) line, the process variable as a slightly bumpy (blue) line, and the controller output as a very bumpy (purple) line. We can see from this trend that the controller is doing exactly what it should: holding the process variable value close to setpoint, manipulating the final control element as far as necessary to do so. The erratic appearance of the output signal is not really a problem, contrary to most peoples' first impression. The fact that the process variable never deviates significantly from the setpoint tells us the control system is operating quite well. What accounts for the erratic controller output, then? Variations in process load. As other variables in the process vary, the controller is forced to compensate for these variations

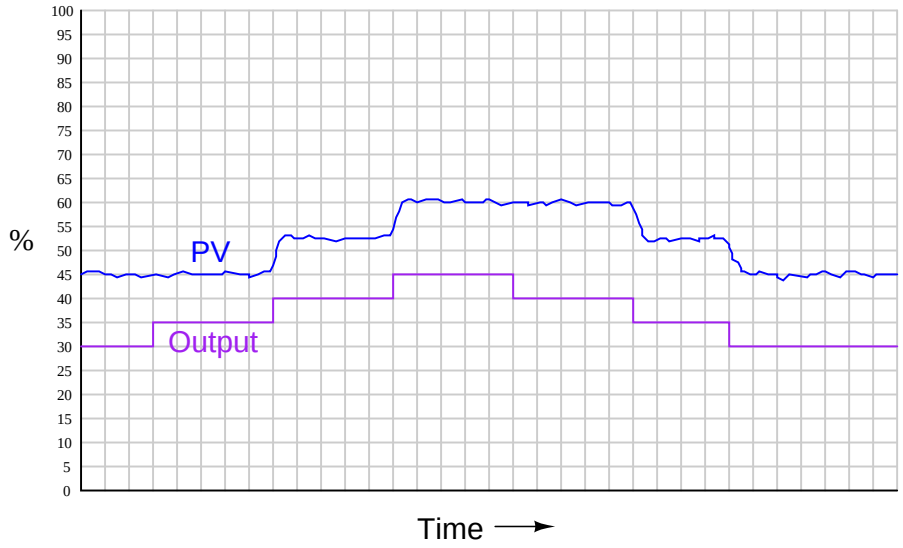
in order that the process variable does not drift from setpoint. Now, maybe this does indicate a problem somewhere else in the process, but there is certainly no problem in this control system.

Recorders become powerful diagnostic tools when coupled with the controller's manual control mode. By placing a controller in "manual" mode and allowing direct human control over the final control element (valve, motor, heater), we can tell a lot about a process. Here is an example of a trend recording for a process in manual mode, where the process variable response is seen graphed in relation to the controller output as that output is increased and decreased in steps:



Notice the time delay between when the output signal is "stepped" to a new value and when the process variable responds to the change. This sort of delay is generally not good for a control system. Imagine trying to steer an automobile whose front wheels respond to your input at the steering wheel only after a 5-second delay! This would be a very challenging car to drive, because the steering is grossly delayed. The same problem plagues any industrial control system with a time lag between the final control element and the transmitter. Typical causes of this problem include *transport delay* (where there is a physical delay resulting from transit time of a process medium from the point of control to the point of measurement) and mechanical problems in the final control element.

This next example shows another type of problem revealed by a trend recording during manual-mode testing:

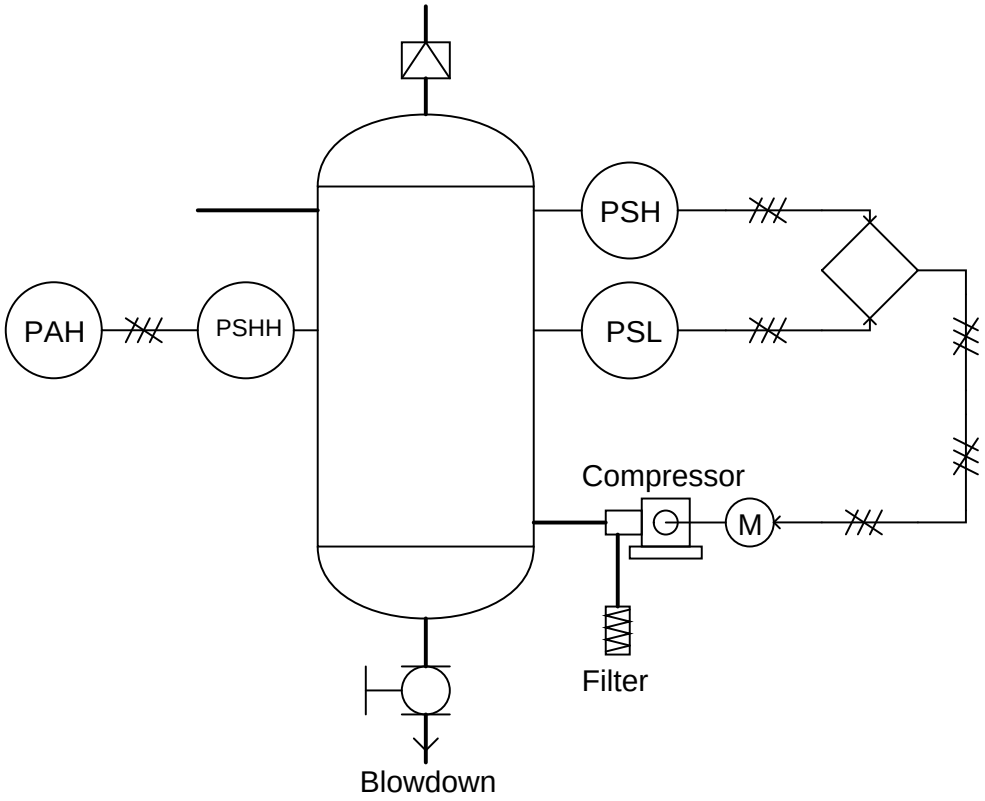


Here, we see the process quickly responding to all step-changes in controller output except for those involving a change in direction. This problem is usually caused by mechanical friction in the final control element (e.g. sticky valve stem packing in a pneumatically-actuated control valve), and is analogous to "loose" steering in an automobile, where the driver must turn the steering wheel a little bit extra after reversing steering direction. Anyone who has ever driven an old farm tractor knows what this phenomenon is like, and how it detrimentally affects one's ability to steer the tractor in a straight line.

6.4.3 Process switches and alarms

Another type of instrument commonly seen in measurement and control systems is the *process switch*. The purpose of a switch is to turn on and off with varying process conditions. Usually, switches are used to activate alarms to alert human operators to take special action. In other situations, switches are directly used as control devices.

The following P&ID of a compressed air control system shows both uses of process switches:

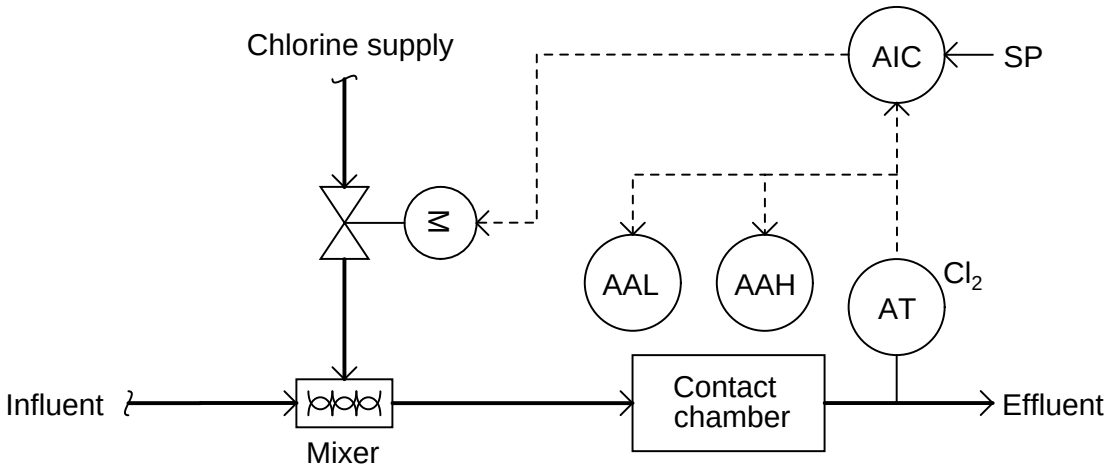


The "PSH" (*pressure switch, high*) activates when the air pressure inside the vessel reaches its high control point. The "PSL" (*pressure switch, low*) activates when the air pressure inside the vessel drops down to its low control point. Both switches feed discrete (on/off) electrical signals to a logic control device (signified by the diamond) which then controls the starting and stopping of the electric motor-driven air compressor.

Another switch in this system labeled "PSHH" (*pressure switch, high-high*) activates only if the air pressure inside the vessel exceeds a level beyond the high shut-off point of the high pressure control switch (PSH). If this switch activates, something has gone wrong with the compressor control system, and the high pressure alarm (PAH, or *pressure alarm, high*) activates to notify a human operator.

All three switches in this air compressor control system are directly actuated by the air pressure in the vessel. In other words these are process-sensing switches. It is possible to build switch devices that interpret standardized instrumentation signals such as 3 to 15 PSI (pneumatic) or 4 to 20 milliamps (analog electronic), which allows us to build on/off control systems and alarms for any type of process variable we can measure with a transmitter.

For example, the chlorine wastewater disinfection system shown earlier may be equipped with a couple of alarm switches to alert an operator if the chlorine concentration ever exceeds predetermined high or low limits:



The labels "AAL" and "AAH" refer to *analytical alarm low* and *analytical alarm high*, respectively. Since both alarms work off the 4 to 20 milliamp electronic signal output by the chlorine analytical transmitter (AT) rather than directly sensing the process, their construction is greatly simplified. If these were process-sensing switches, each one would have to be equipped with the capability of directly sensing chlorine concentration. In other words, each switch would have to be its own chlorine concentration analyzer, with all the inherent complexity of such a device!

An example of such an alarm module (operating off a 4-20 mA current signal) is the Moore Industries model SPA ("Site Programmable Alarm"), shown here:

Like all current-operated alarm modules, the Moore Industries SPA may be configured to "trip" electrical contacts when the current signal reaches a variety of different programmed thresholds. Some of the alarm types provided by this unit include high process, low process, out-of-range, and high rate-of-change.

Process alarm switches may be used to trigger a special type of indicator device known as an *annunciator*. An annunciator is an array of indicator lights and associated circuitry designed to secure a human operator's attention[1] by blinking and sounding an audible buzzer when a process switch actuates into an abnormal state. The alarm state may be then "acknowledged" by an operator pushing a button, causing the alarm light to remain on (solid) rather than blink, and silencing the buzzer. The indicator light does not turn off until the actual alarm condition (the process switch) has returned to its regular state.

[1] D.A. Strobhar, writing in *The Instrument Engineer's Handbook* on the subject of alarm management, makes the interesting observation that alarms are the only form of instrument "whose sole purpose is to alter the operator's behavior." Other instrument devices may automatically respond to process changes by directly influencing the process, but only alarms work to *control the operator*.

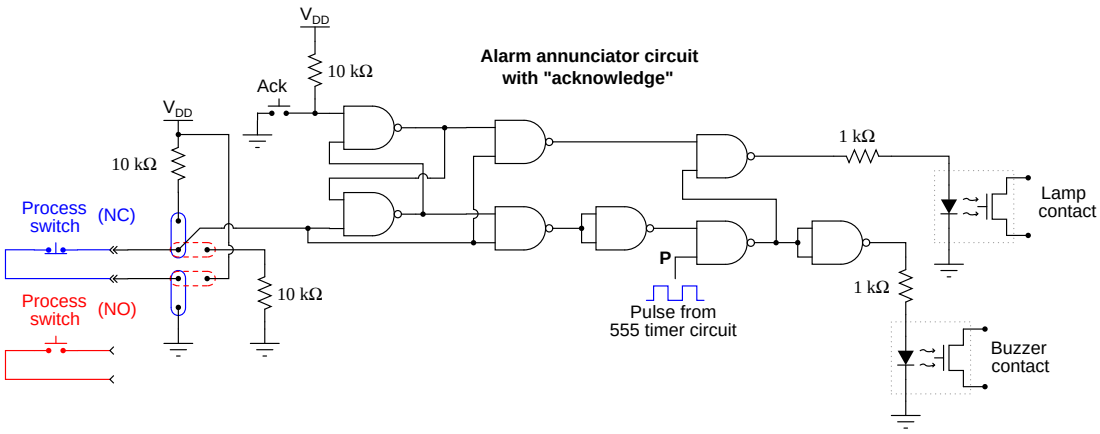
This photograph shows an annunciator located on a control panel for a large engine-driven pump. Each white plastic square with writing on it is a translucent pane covering a small light bulb. When an alarm condition occurs, the respective light bulb flashes, causing the translucent white plastic to glow, highlighting to the operator which alarm is active:

Note the two pushbutton switches below labeled "Test" and "Acknowledge." Pressing the "Acknowledge" button will silence the audible buzzer and also turn any blinking alarm light into a steady (solid) alarm light until the alarm condition clears, at which time the light turns off completely. Pressing the "Test" button turns all alarm lights on, to ensure all light bulbs are still functional.

Opening the front panel of this annunciator reveals modular relay units controlling the blinking and acknowledgment latch functions, one for each alarm light:

This modular design allows each alarm channel to be serviced without necessarily interrupting the function of all others in the annunciator panel.

A simple logic gate circuit illustrates the acknowledgment latching feature (here implemented by an S-R latch circuit) common to all process alarm annunciators:



Panel-mounted annunciators are becoming a thing of the past, as computer-based alarm displays replace them with advanced capabilities such as time logging, first-event recording, and multiple layers of ackowledgement/access. Time logging is of particular importance in the process industries, as the sequence of events is often extremely important in investigations following an abnormal operating condition. Knowing *what went wrong, in what order* is much more informative than simply knowing which alarms have tripped.

6.5 Summary

Instrument technicians maintain the safe and efficient operation of industrial measurement and control systems. As this chapter shows, this requires a broad command of technical skill. Instrumentation is more than just physics or chemistry or mathematics or electronics or mechanics or control theory alone. An instrument technician must understand all these subject areas to some degree, and more importantly how these knowledge areas relate to each other.

The all-inclusiveness of this profession makes it very challenging and interesting. Adding to the challenge is the continual introduction of new technologies. The advent of new technologies, however, does not necessarily relegate legacy technologies to the scrap heap. It is quite common to find state-of-the-art instruments in the very same facility as decades-old instruments; digital fieldbus networks running alongside 3 to 15 PSI pneumatic signal tubes; microprocessor-based sensors mounted right next to old mercury tilt-switches. Thus, the competent instrument technician must be comfortable working with both old and new technologies, understanding the relative merits and weaknesses of each.

This is why the most important skill for an instrument technician is the ability to teach oneself. It is impossible to fully prepare for a career like this with any amount of preparatory schooling. The profession is so broad and the responsibility so great, and the landscape so continuously subject to change, that life-long learning for the technician is a matter of professional survival.

Perhaps the single greatest factor determining a person's ability to independently learn is their skill at *reading*. Being able to "digest" the written word is *the* key to learning what is difficult or impractical to directly experience. In an age where information is readily accessible, the skilled reader has the advantage of leveraging generations of experts in any subject. Best of all, reading is a skill anyone can master, and everyone should.

My advice to all those desiring to become self-directed learners is to build a library of reading material on subjects that interest you (hopefully, instrumentation is one of those subjects!), and then immerse yourself in those writings. Feel free to "mark up" your books, or take notes in a separate location, so as to actively engage in your reading. Try as much as possible to approach reading as though you were having a *conversation* with the author: pose questions, challenge concepts and ideas, and do not stop doing so until you can clearly see what the author is trying to say.

I also advise *writing* about what you have learned, because re-phrasing key ideas in your own words helps you consolidate the learning, and "makes it your own" in a way few other activities do. You don't necessarily have to write your own book, but the act of expressing what you have learned to the best of your ability is a powerful tool not only for building confidence in what you know, but also for raising your own awareness of what you do not (yet) know.

References

Lipták, Béla G., *Instrument Engineers' Handbook – Process Software and Digital Networks*, Third Edition, CRC Press, New York, NY, 2002.

Chapter 7

Instrumentation documents

Every technical discipline has its own standardized way(s) of making descriptive diagrams, and instrumentation is no exception. The scope of instrumentation is so broad, however, that no one form of diagram is sufficient to capture all we might need to represent. This chapter will discuss three different types of instrumentation diagrams:

- Process Flow Diagrams (PFDs)
- Process and Instrument diagrams (P&IDs)
- Loop diagrams ("loop sheets")
- SAMA diagrams

At the highest level, the instrument technician is interested in the interconnections of process vessels, pipes, and flow paths of process fluids. The proper form of diagram to represent the "big picture" of a process is called a *process flow diagram*. Individual instruments are sparsely represented in a PFD, because the focus of the diagram is the process itself.

At the lowest level, the instrument technician is interested in the interconnections of individual instruments, including all the wire numbers, terminal numbers, cable types, instrument calibration ranges, etc. The proper form of diagram for this level of fine detail is called a *loop diagram*. Here, the process vessels and piping are sparsely represented, because the focus of the diagram is the instruments themselves.

Process and instrument diagrams (P&IDs) lie somewhere in the middle between process flow diagrams and loop diagrams. A P&ID shows the layout of all relevant process vessels, pipes, and machinery, but with instruments superimposed on the diagram showing what gets measured and what gets controlled. Here, one can view the flow of the process as well as the "flow" of information between instruments measuring and controlling the process.

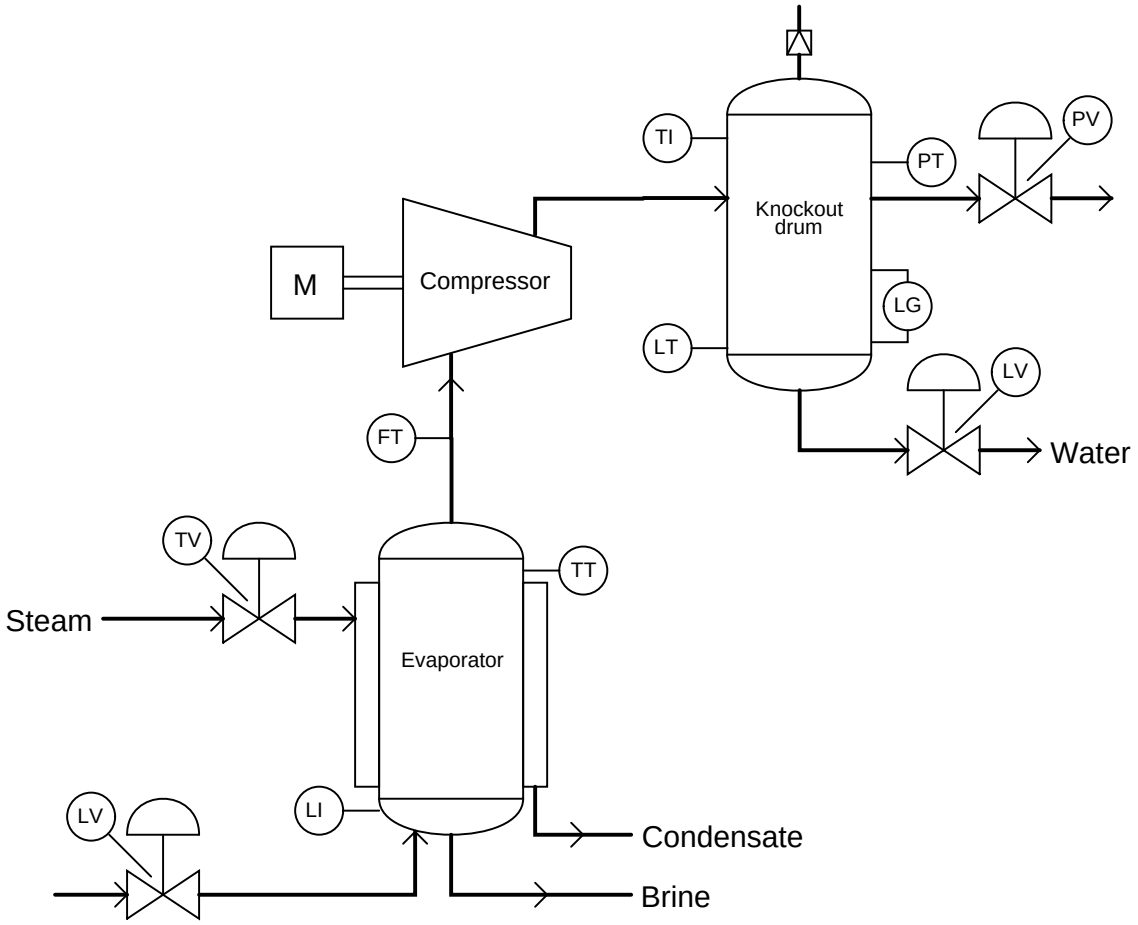
SAMA diagrams are used for an entirely different purpose: to document the *strategy* of a control system. In a SAMA diagram, emphasis is placed on the algorithms used to control a process, as opposed to piping, wiring, or instrument connections. These diagrams are commonly found within the power generation industry, but are sometimes used in other industries as well.

An instrument technician must often switch between different diagrams when troubleshooting a complex control system. There is simply too much detail for any one diagram to show everything. Even if the page were large enough, a "show everything" diagram would be so chock-full of details that it would be difficult to follow any one line of details you happened to be interested in at any particular time. The narrowing of scope with the progression from PFD to loop diagram may be visualized as a process of "zooming in," as though one were viewing a process through the lens of a microscope at different powers. First you begin with a PFD or P&ID to get an overview of the process, to see how the major components interact. Then, once you have identified which instrument "loop" you need to investigate, you go to the appropriate loop diagram to see the interconnection details of that instrument system so you know where to connect your test equipment and what signals you expect to find when you do.

Another analogy for this progression of documents is a map, or more precisely, a globe, an atlas, and a city street map. The globe gives you the "big picture" of the Earth, countries, and major cities. An atlas allows you to "zoom in" to see details of particular provinces, states, and principalities, and the routes of travel connecting them all. A city map shows you major and minor roads, canals, alleyways, and perhaps even some addresses in order for you to find your way to a particular destination. It would be impractical to have a globe large enough to show you all the details of every city! Furthermore, a globe comprehensive enough to show you all these details would have to be updated *very* frequently to keep up with all cities' road changes. There is a certain economy inherent to the omission of fine details, both in ease of use and in ease of maintenance.

7.1 Process Flow Diagrams

To show a practical process example, let's examine three diagrams for a compressor control system. In this fictitious process, water is being evaporated from a process solution under partial vacuum (provided by the compressor). The compressor then transports the vapors to a "knockout drum" where some of them condense into liquid form. As a typical PFD, this diagram shows the major interconnections of process vessels and equipment, but omits details such as instrument signal lines and auxiliary instruments:



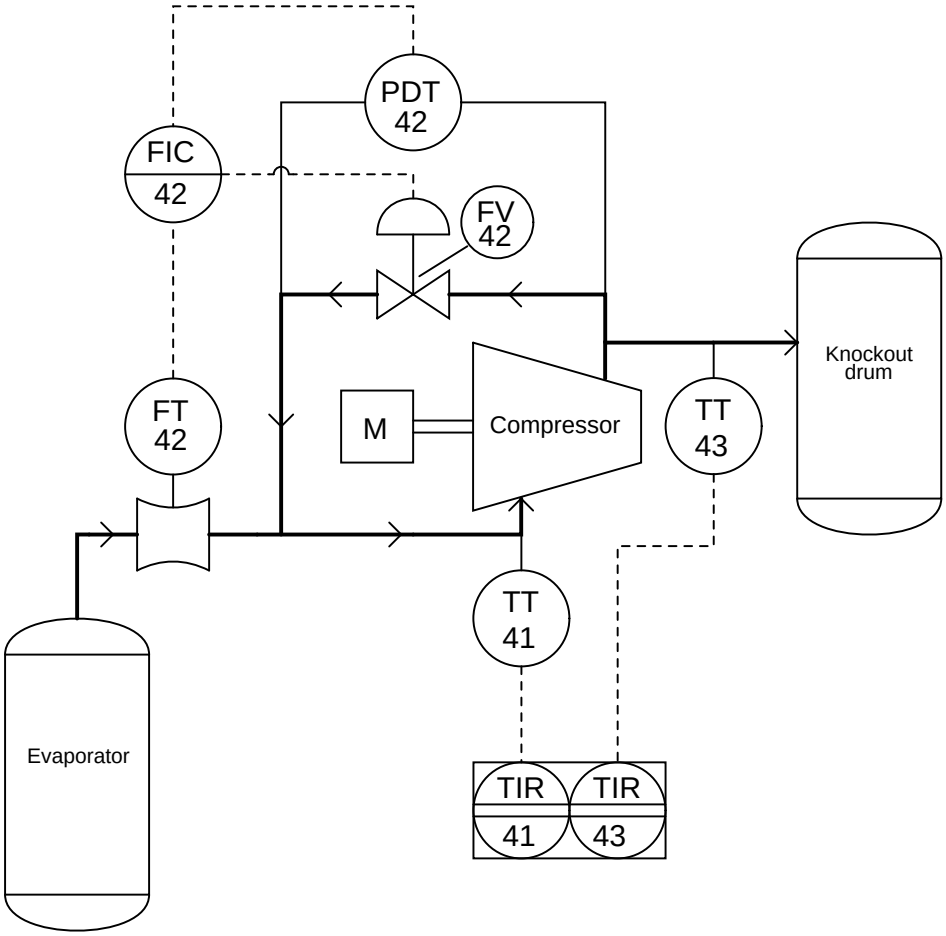
One might guess the instrument interconnections based on the instruments' labels. For instance, a good guess would be that the level transmitter (LT) on the bottom of the knockout drum might send the signal that eventually controls the level valve (LV) on the bottom of that same vessel. One might also guess that the temperature transmitter (TT) on the top of the evaporator might be part of the temperature control system that lets steam into the heating jacket of that vessel.

Based on this diagram alone, one would be hard-pressed to determine what control system, if

any, controls the compressor itself. All the PFD shows relating directly to the compressor is a flow transmitter (FT) on the suction line. This level of uncertainty is perfectly acceptable for a PFD, because its purpose is merely to show the general flow of the process itself, and only a bare minimum of control instrumentation.

7.2 Process and Instrument Diagrams

The next level of detail is the Process and Instrument Diagram[1], or P&ID. Here, we see a "zooming in" of scope from the whole evaporator process to the compressor as a unit. The evaporator and knockout vessels almost fade into the background, with their associated instruments absent from view:



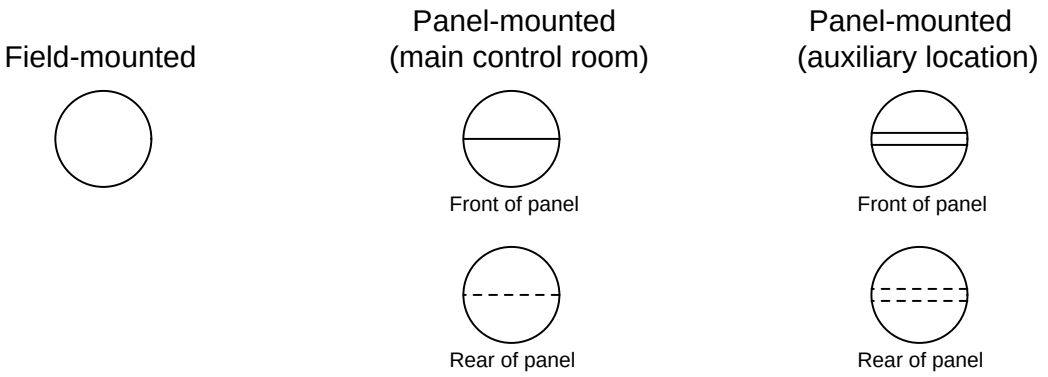
Now we see there is more instrumentation associated with the compressor than just a flow transmitter. There is also a differential pressure transmitter (PDT), a flow indicating controller (FIC), and a "recycle" control valve that allows some of the vapor coming out of the compressor's discharge line to go back around into the compressor's suction line. Additionally, we have a pair of temperature transmitters that report suction and discharge line temperatures to an indicating recorder.

Some other noteworthy details emerge in the P&ID as well. We see that the flow transmitter, flow

[1]Sometimes P&ID stands for *Piping* and Instrument Diagram. Either way, it means the same thing.

controller, pressure transmitter, and flow valve all bear a common number: 42. This common "loop number" indicates these four instruments are all part of the same control system. An instrument with any other loop number is part of a different control system, measuring and/or controlling some other function in the process. Examples of this include the two temperature transmitters and their respective recorders, bearing the loop numbers 41 and 43.

Please note the differences in the instrument "bubbles" as shown on this P&ID. Some of the bubbles are just open circles, where others have lines going through the middle. Each of these symbols has meaning according to the ISA (Instrumentation, Systems, and Automation society) standard:



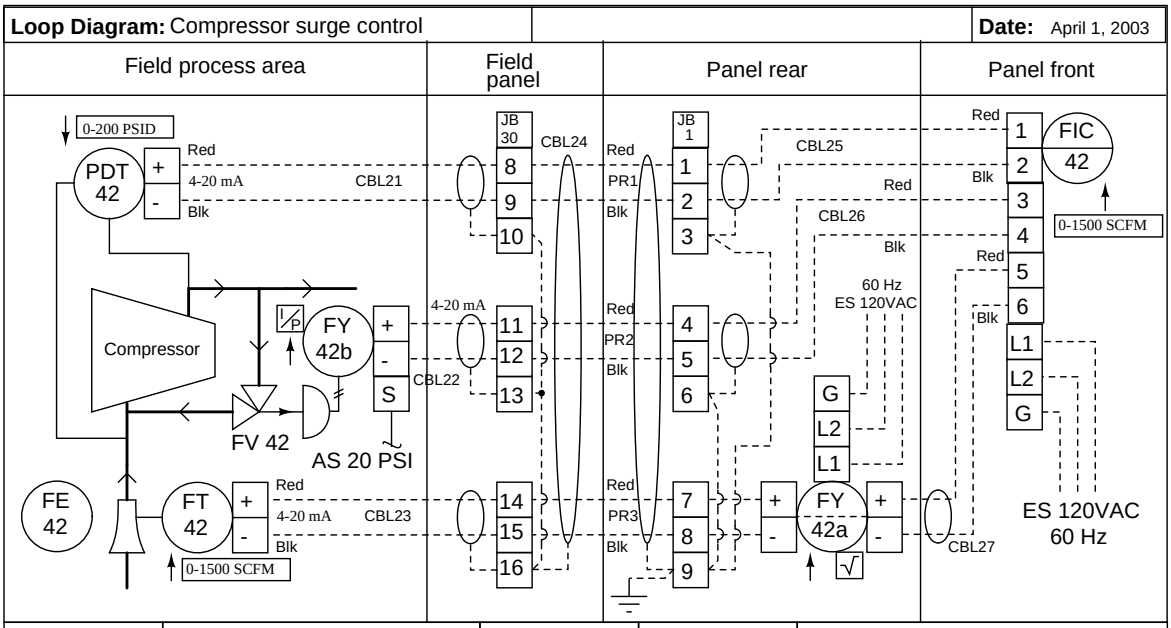
The type of "bubble" used for each instrument tells us something about its location. This, obviously, is quite important when working in a facility with many thousands of instruments scattered over acres of facility area, structures, and buildings.

The rectangular box enclosing both temperature recorders shows they are part of the same physical instrument. In other words, this indicates there is really only one temperature recorder instrument, and that it plots both suction and discharge temperatures (most likely on the same trend graph). This suggests that each bubble may not necessarily represent a discrete, physical instrument, but rather an instrument *function* that may reside in a multi-function device.

Details we do not see on this P&ID include cable types, wire numbers, terminal blocks, junction boxes, instrument calibration ranges, failure modes, power sources, and the like. To examine this level of detail, we must go to the loop diagram we are interested in.

7.3 Loop diagrams

Finally, we arrive at the loop diagram (sometimes called a *loop sheet*) for the compressor surge control system (loop number 42):



Tag number	Description	Input cal.	Output cal.	Notes
FE 42	Venturi tube	0-1500 SCFM	0-100 "WC	
FT 42	Suction flow transmitter	0-100 "WC	4-20 mA	
FY 42a	Square root extractor	4-20 mA	4-20 mA	
FY 42b	Current-to-pressure converter	4-20 mA	3-15 PSI	
FV 42	Anti-surge control valve	3-15 PSI	100%-0%	Air-to-close
PDT 42	Differential pressure transmitter	0-200 PSI	20-4 mA	Reverse action
FIC 42	Anti-surge controller	4-20 mA	4-20 mA	

Here we see that the P&ID didn't show us all the instruments in this control "loop." Not only do we have two transmitters, a controller, and a valve; we also have two signal transducers. Transducer 42a modifies the flow transmitter's signal before it goes into the controller, and transducer 42b converts the electronic 4 to 20 mA signal into a pneumatic 3 to 15 PSI air pressure signal. Each instrument "bubble" in a loop diagram represents an individual device, with its own terminals for connecting wires.

Note that dashed lines now represent individual copper wires instead of whole cables. Terminal blocks where these wires connect to are represented by squares with numbers in them. Cable numbers, wire colors, junction block numbers, panel identification, and even grounding points are all shown in loop diagrams. The only type of diagram at a lower level of abstraction than a loop diagram would be an electronic schematic diagram for an individual instrument, which of course

would only show details pertaining to that one instrument. Thus, the loop diagram is the most detailed form of diagram for a control system as a whole, and thus it must contain all details omitted by PFDs and P&IDs alike.

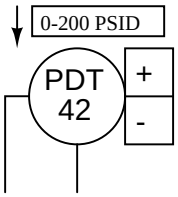
To the novice it may seem excessive to include such trivia as wire colors in a loop diagram. To the experienced instrument technician who has had to work on systems lacking such documented detail, this information is highly valued. The more detail you put into a loop diagram, the easier it makes the inevitable job of maintaining that system at some later date. When a loop diagram shows you exactly what wire color to expect at exactly what point in an instrumentation system, and exactly what terminal that wire should connect to, it becomes much easier to proceed with any troubleshooting, calibration, or upgrade task.

An interesting detail seen on this loop diagram is an entry specifying "input calibration" and "output calibration" for each and every instrument in the system. This is actually a very important concept to keep in mind when troubleshooting a complex instrumentation system: every instrument has at least one input and at least one output, with some sort of mathematical relationship between the two. Diagnosing where a problem lies within a measurement or control system often reduces to testing various instruments to see if their output responses appropriately match their input conditions.

For example, one way to test the flow transmitter in this system would be to subject it to a number of different pressures within its range (specified in the diagram as 0 to 100 inches of water column differential) and seeing whether or not the current signal output by the transmitter was consistently proportional to the applied pressure (e.g. 4 mA at 0 inches pressure, 20 mA at 100 inches pressure, 12 mA at 50 inches pressure, etc.).

Given the fact that a calibration error or malfunction in any one of these instruments can cause a problem for the control system as a whole, it is nice to know there is a way to determine which instrument is to blame and which instruments are not. This general principle holds true regardless of the instrument's type or technology. You can use the same input-versus-output test procedure to verify the proper operation of a pneumatic (3 to 15 PSI) level transmitter or an analog electronic (4 to 20 mA) flow transmitter or a digital (fieldbus) temperature transmitter alike. Each and every instrument has an input and an output, and there is always a predictable (and testable) correlation from one to the other.

Another interesting detail seen on this loop diagram is the *action* of each instrument. You will notice a box and arrow (pointing either up or down) next to each instrument bubble. An "up" arrow (↑) represents a *direct-acting* instrument: one whose output signal increases as the input stimulus increases. A "down" arrow (↓) represents a *reverse-acting* instrument: one whose output signal decreases as the input stimulus increases. All the instruments in this loop are direct-acting with the exception of the pressure differential transmitter PDT-42:

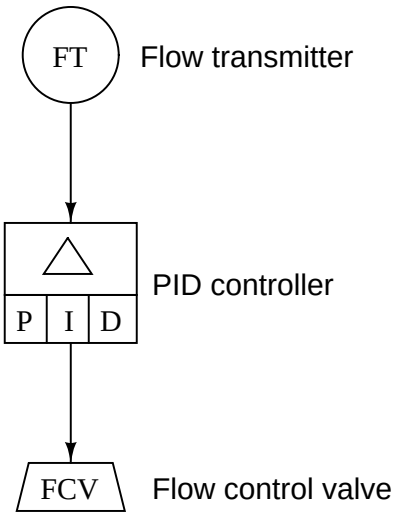


Here, the "down" arrow tells us the transmitter will output a full-range signal (20 mA) when it senses zero differential pressure, and a 0% signal (4 mA) when sensing a full 200 PSI differential. While this calibration may seem confusing and unwarranted, it serves a definite purpose in this particular control system. Since the transmitter's current signal decreases as pressure increases, and the controller must be correspondingly configured, a decreasing current signal will be interpreted by the controller as a high differential pressure. If any wire connection fails in the 4-20 mA current loop for that transmitter, the resulting 0 mA signal will be naturally "seen" by the controller as a pressure over-range condition. This is considered dangerous in a compressor system because it predicts a condition of surge. Thus, the controller will naturally take action to prevent surge by commanding the anti-surge control valve to open, because it "thinks" the compressor is about to surge. In other words, the transmitter is intentionally calibrated to be reverse-acting such that any break in the signal wiring will naturally bring the system to its safest condition.

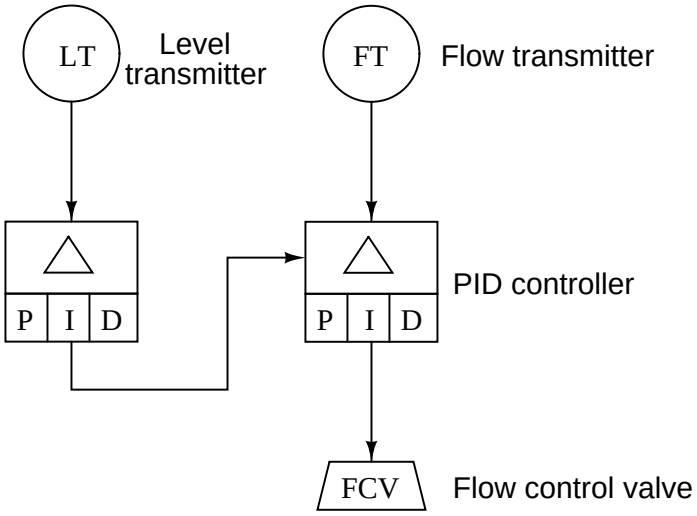
7.4 SAMA diagrams

SAMA is an acronym standing for *Scientific Apparatus Makers Association*, referring to a unique form of diagram used primary in the power generation industry to document control strategies. These diagrams focus on the flow of information within a control system rather than on the process piping or instrument interconnections (wires, tubes, etc.). The general flow of a SAMA diagram is top-to-bottom, with the process sensing instrument (transmitter) located at the top and the final control element (valve or variable-speed motor) located at the bottom. No attempt is made to arrange symbols in a SAMA diagram to correlate with actual equipment layout: these diagrams are all about the *algorithms* used to make control decisions, and nothing more.

A sample SAMA diagram appears here, showing a flow transmitter (FT) sending a process variable signal to a PID controller, which then sends a manipulated variable signal to a flow control valve (FCV):

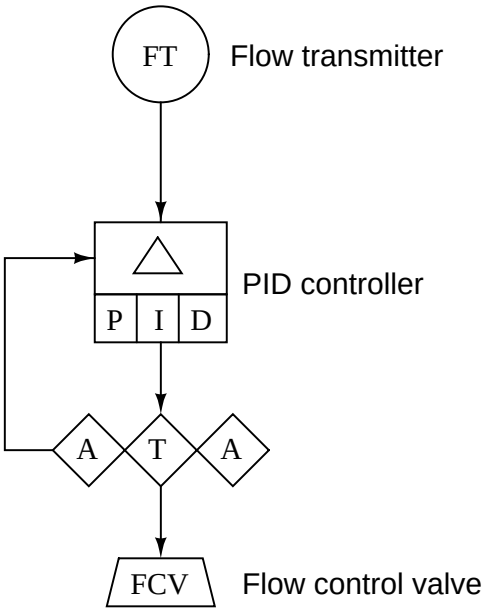


A cascaded control system, where the output of one controller acts as the setpoint for another controller to follow, appears in SAMA diagram form like this:

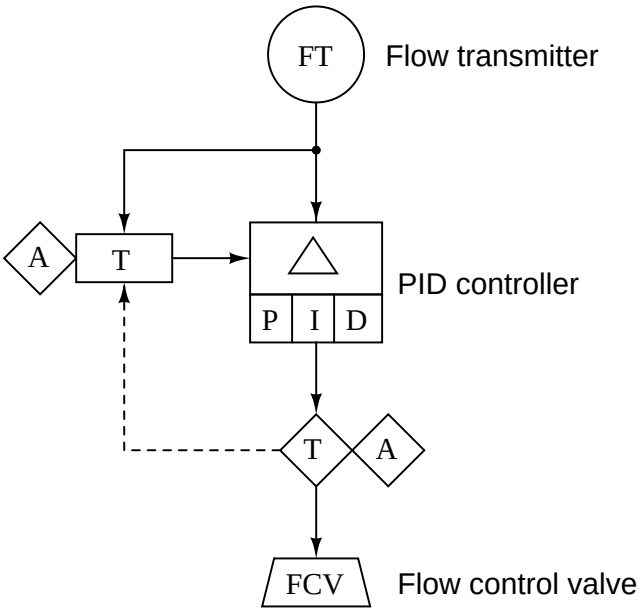


In this case, the primary controller senses the level in a vessel, commanding the secondary (flow) controller to maintain the necessary amount of flow either in or out of the vessel as needed to maintain level at some setpoint.

SAMA diagrams may show varying degrees of detail about the control strategies they document. For example, you may see the auto/manual controls represented as separate entities in a SAMA diagram, apart from the basic PID controller function. In the following example, we see a transfer block (T) and two manual adjustment blocks (A) providing a human operator the ability to separately adjust the controller's setpoint and output (manipulated) variables, and to transfer between automatic and manual modes:



Rectangular blocks such as the Δ, P, I, and D shown in this diagram represent automatic functions. Diamond-shaped blocks such as the A and T blocks are manual functions which must be set by a human operator. Showing even more detail, the following SAMA diagram indicates the presence of *setpoint tracking* in the controller algorithm, a feature that forces the setpoint value to equal the process variable value any time the controller is in manual mode:

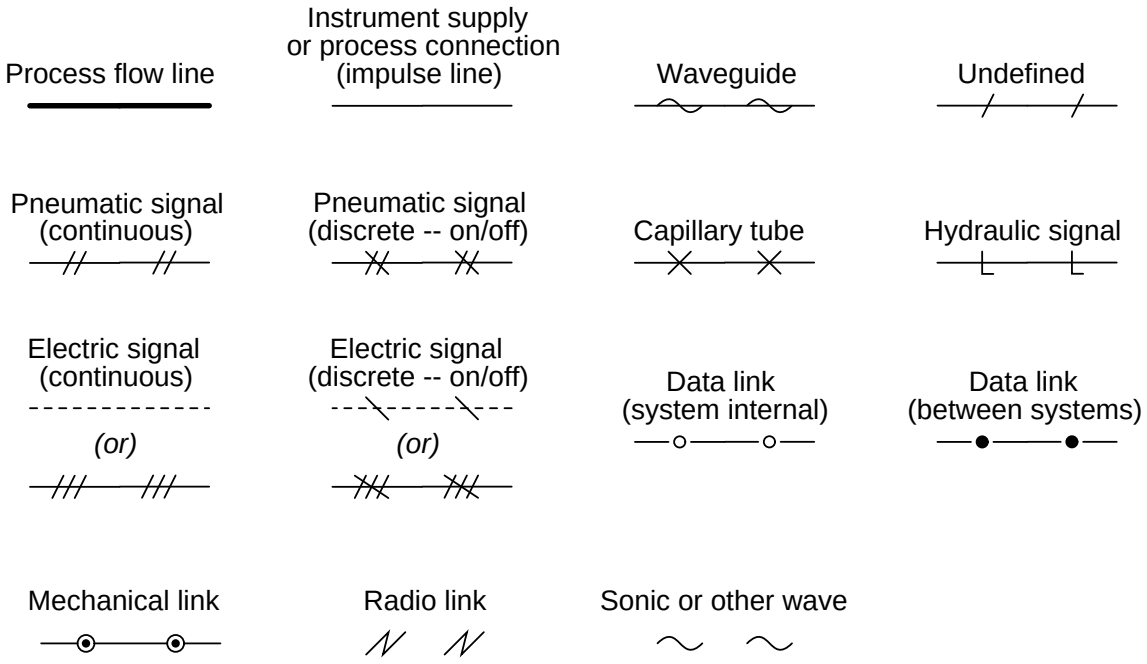


Here we see a new type of line: dashed instead of solid. This too has meaning in the world of SAMA diagrams. Solid lines represent analog (continuously variable) signals such as process variable, setpoint, and manipulated variable. Dashed lines represent discrete (on/off) signal paths, in this case the auto/manual state of the controller commanding the PID algorithm to get its setpoint either from the operator's input (A) or from the process variable input (the flow transmitter: FT).

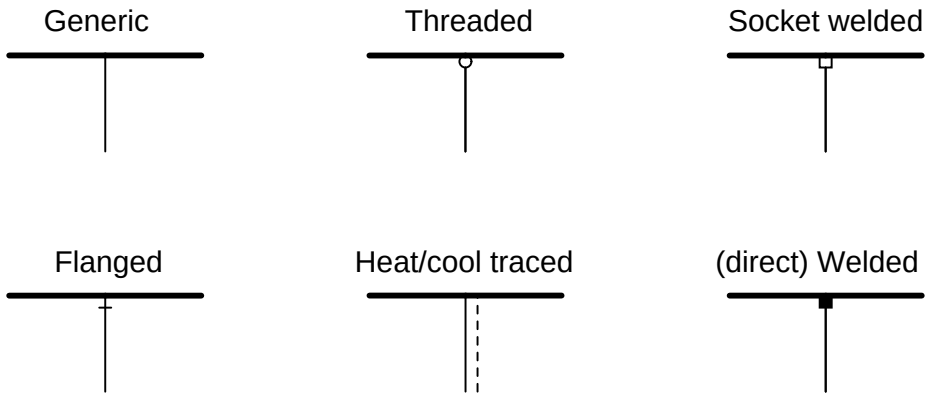
7.5 Instrument and process equipment symbols

This section shows some of the many instrument symbols included in the ISA 5.1 standard. These symbols find application in Process Flow Diagrams (PFDs), Process and Instrument Diagrams (P&IDs), and loop diagrams alike.

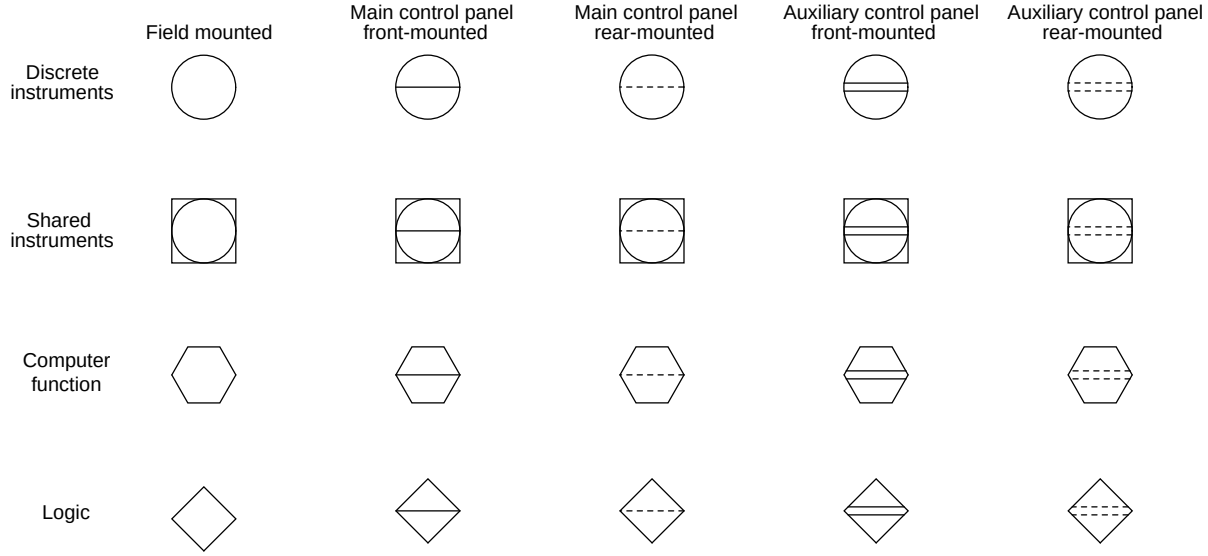
7.5.1 Line types



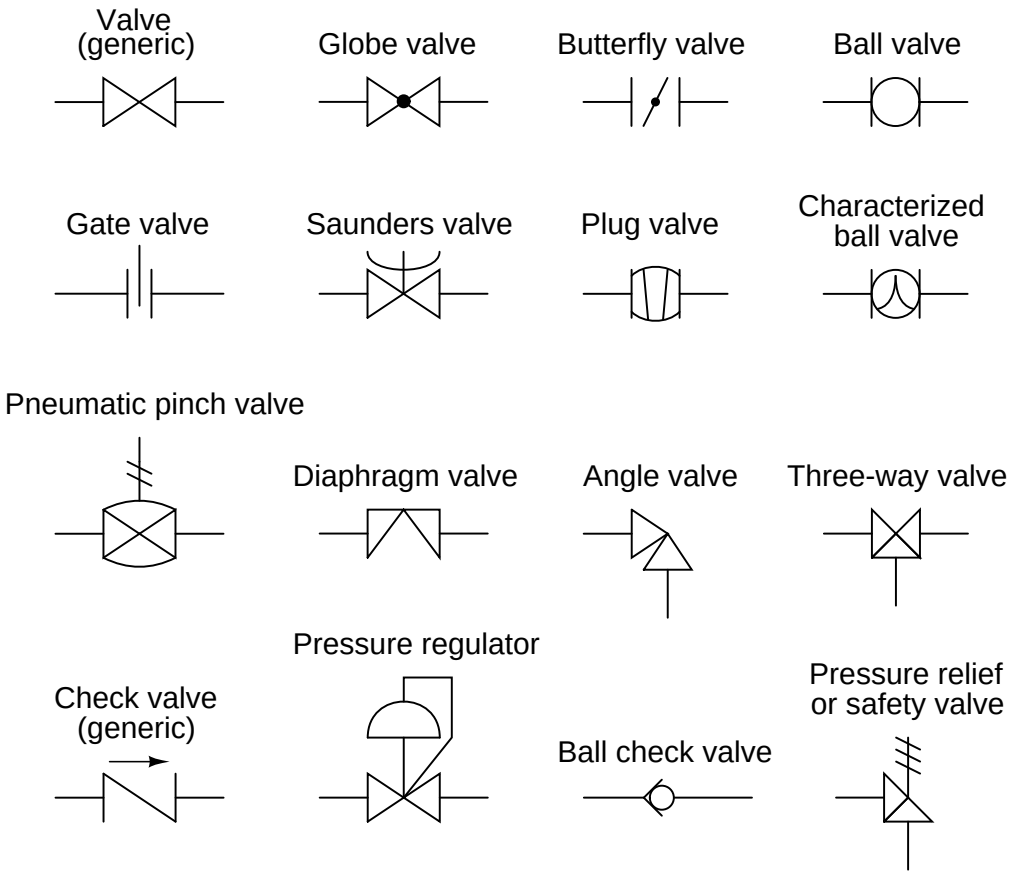
7.5.2 Process/Instrument line connections



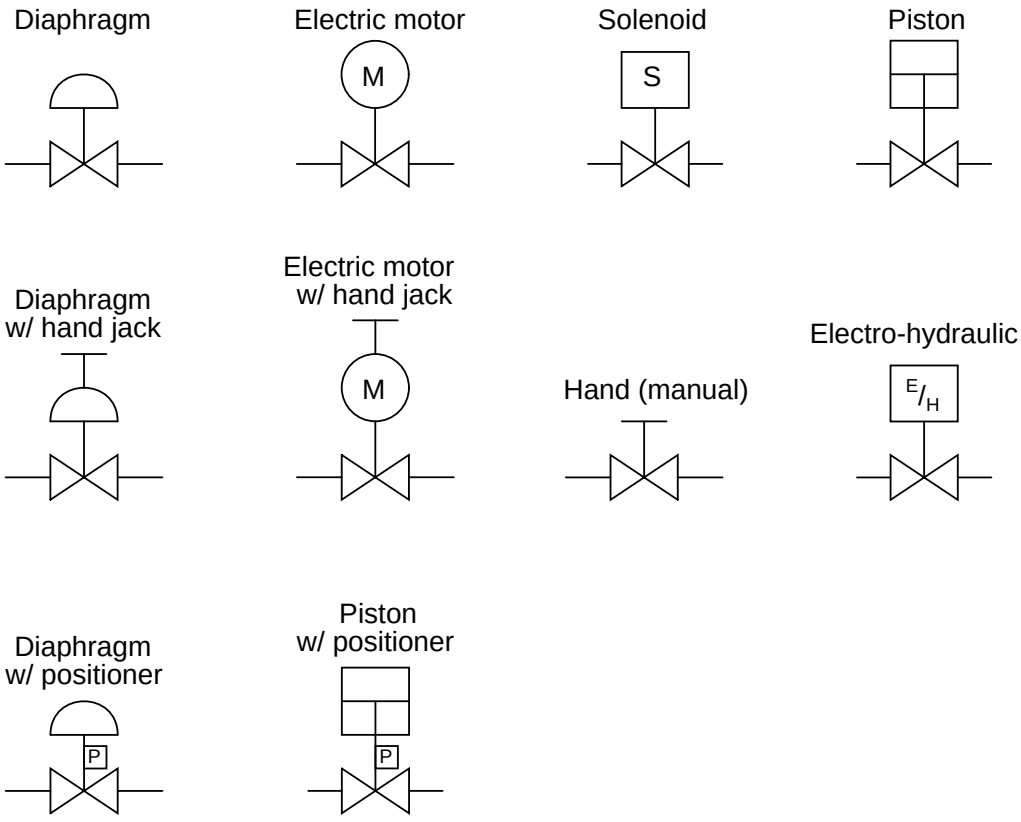
7.5.3 Instrument bubbles



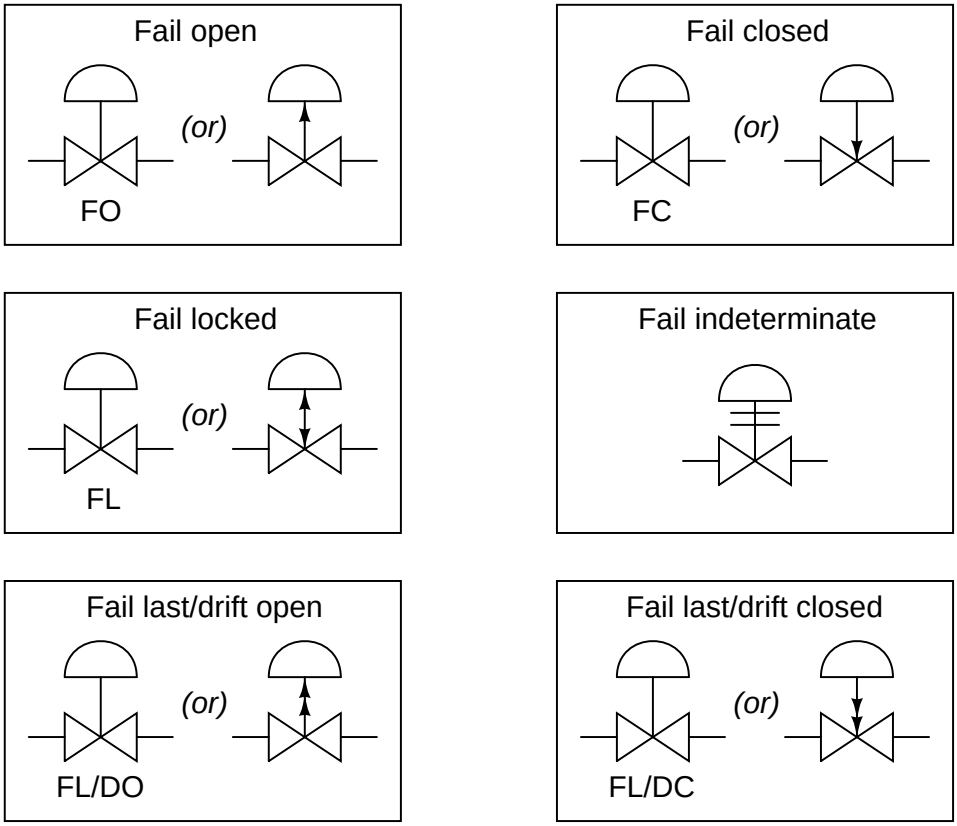
7.5.4 Process valve types



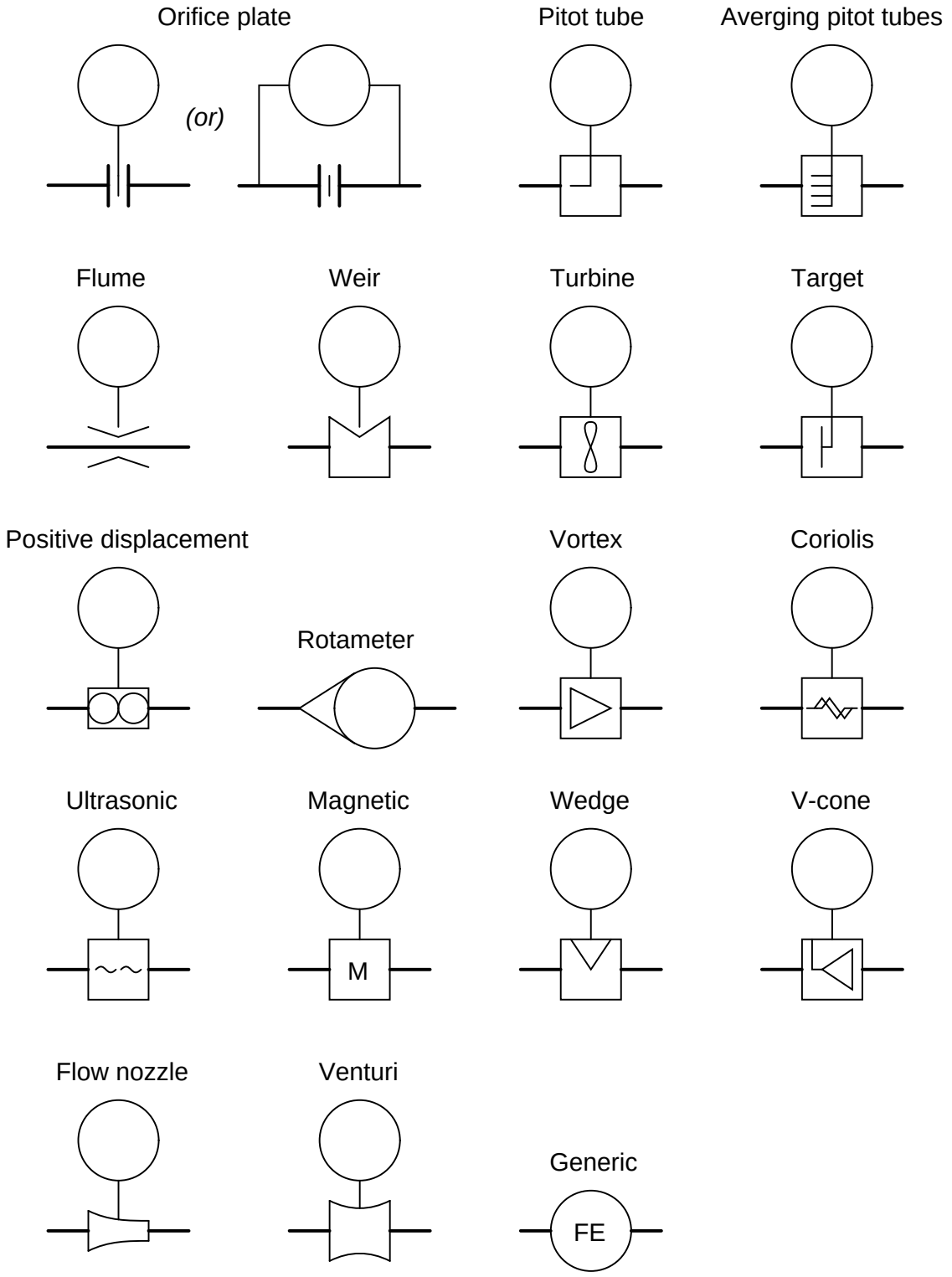
7.5.5 Valve actuator types



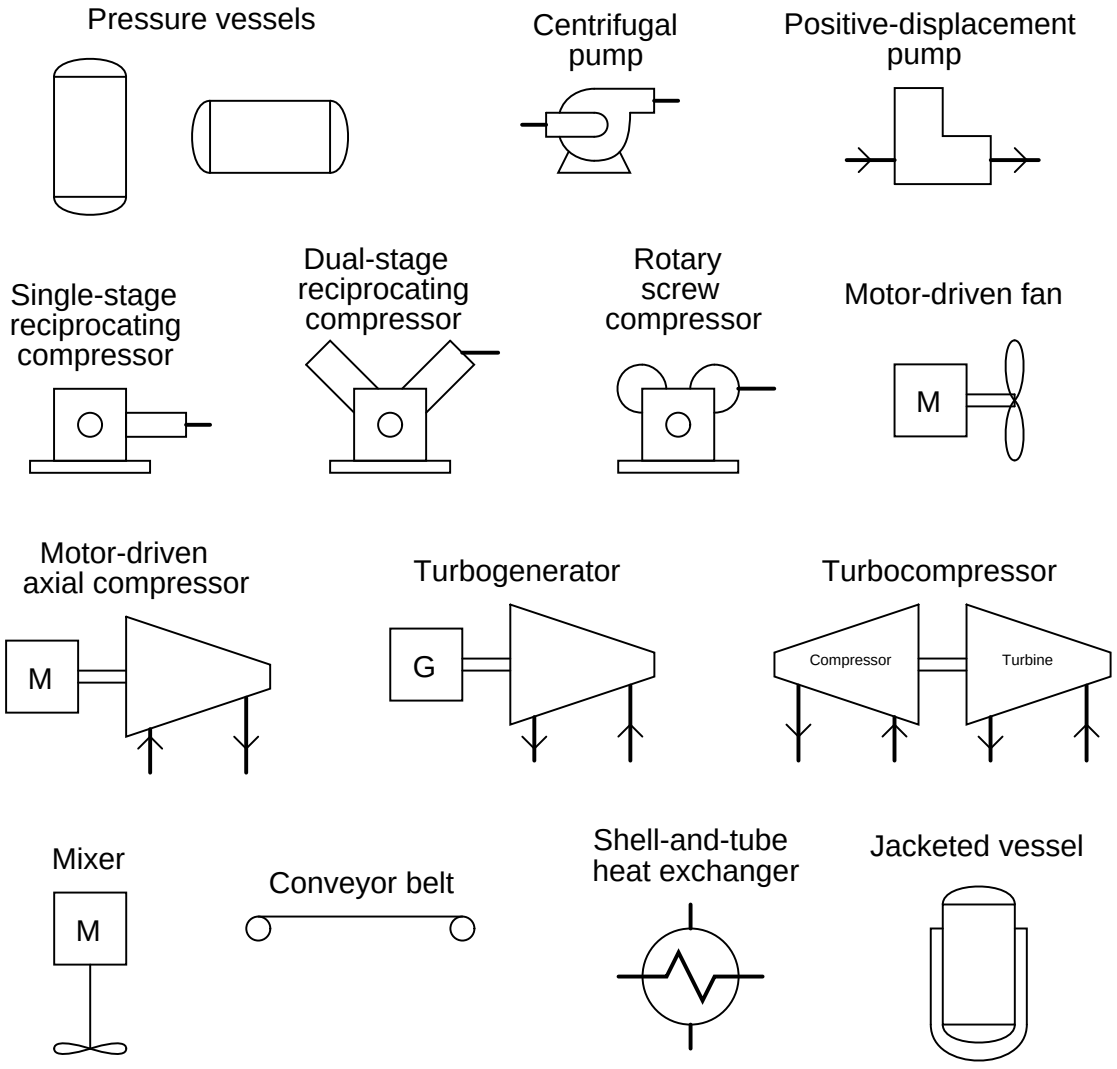
7.5.6 Valve failure mode



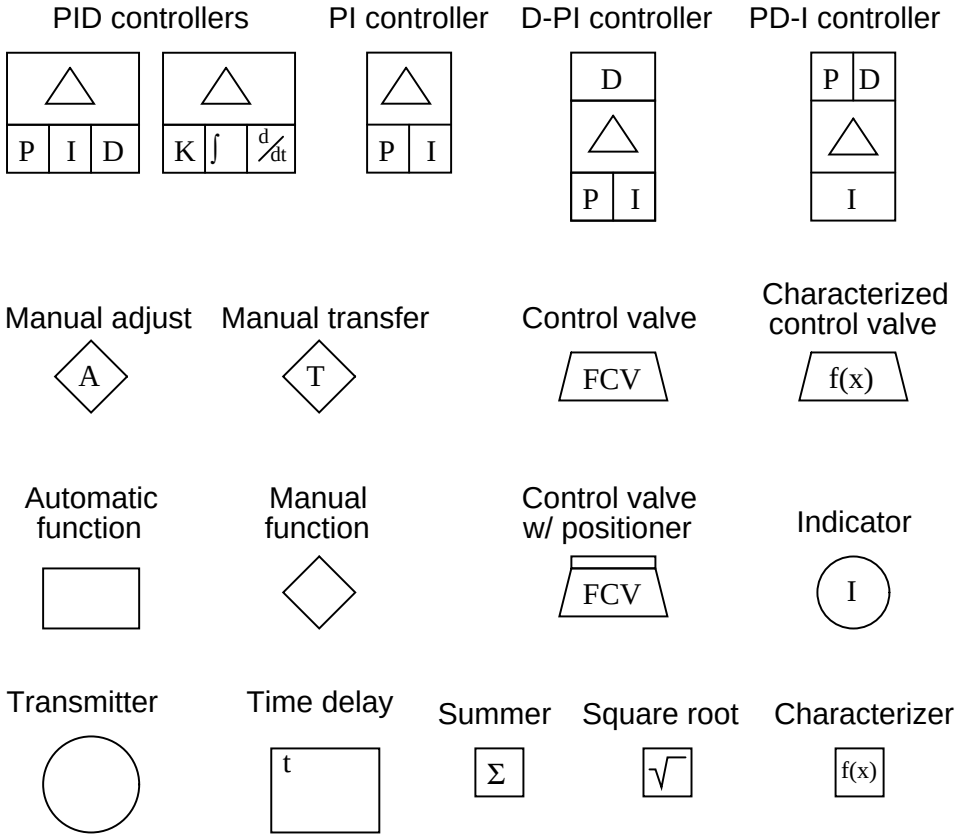
7.5.7 Flow measurement devices (flowing left-to-right)



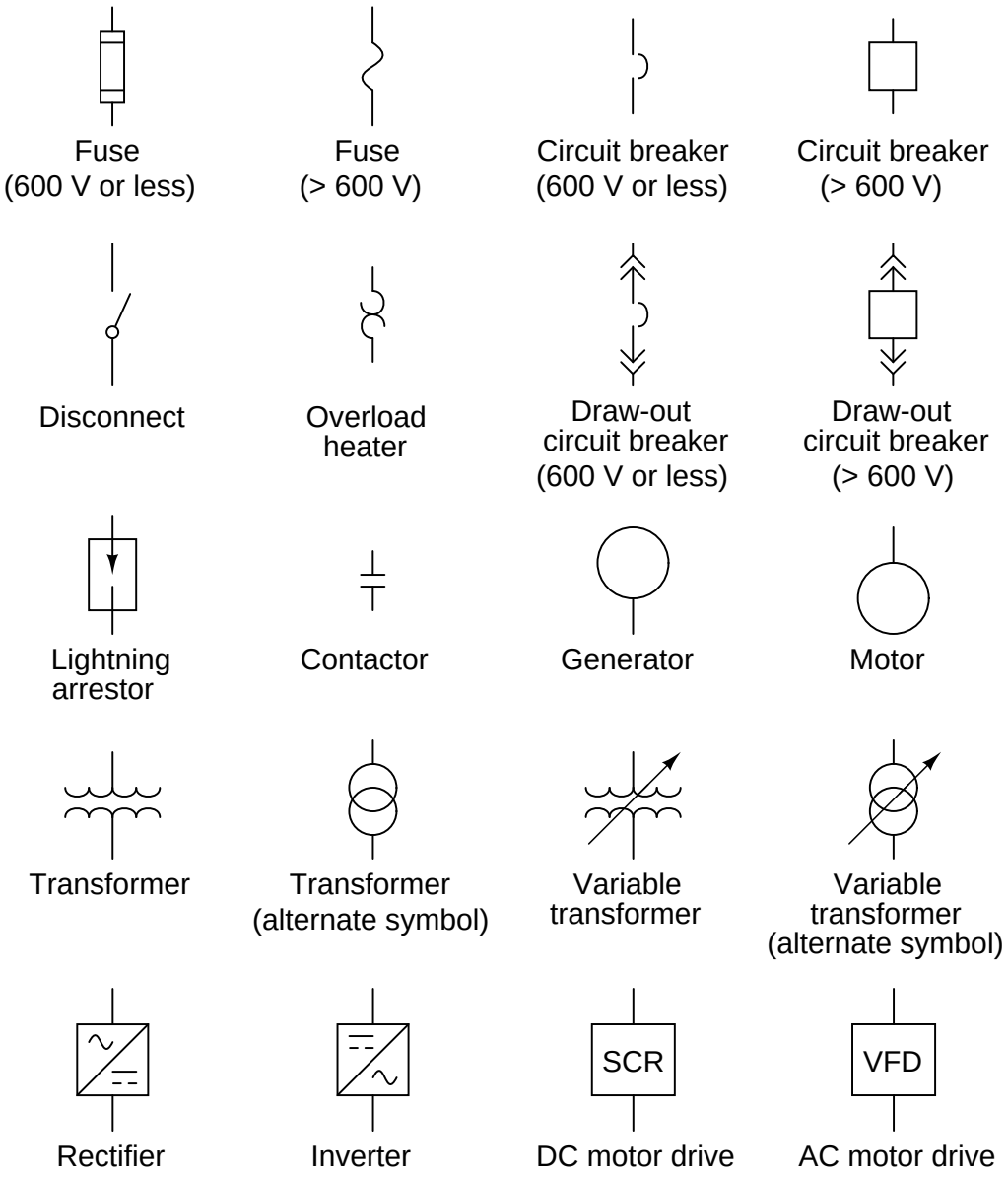
7.5.8 Process equipment

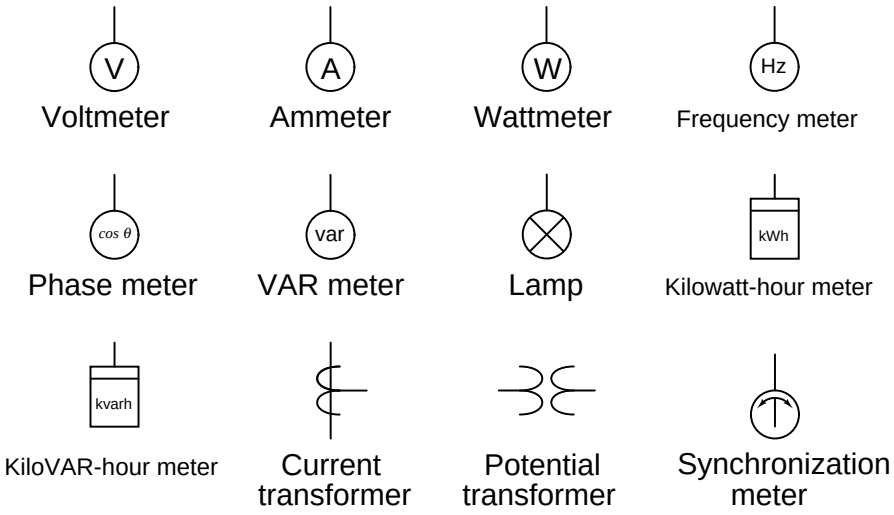


7.5.9 SAMA diagram symbols



7.5.10 Single-line electrical diagram symbols





7.6 Instrument identification tags

Up until this point, we have explored various types of instrumentation diagram, each one making reference to different instruments by lettered identifiers such as TT (Temperature Transmitter), PDT (Pressure Differential Transmitter), or FV (Flow Valve), without formally defining all the letters used to identify instruments. Part of the ISA 5.1 standard does exactly this, which is what we will now investigate.

Each instrument within an instrumented facility should have its own unique identifying *tag* consisting of a series of letters describing that instrument's *function*, as well as a number identifying the particular *loop* it belongs to. An optional numerical prefix typically designates the larger area of the facility in which the loop resides, and an optional alphabetical suffix designates multiple instances of instruments within one loop.

For example, if we were to see an instrument bearing the tag `FC-135`, we would know it was a *flow controller* (FC) for loop number *135*. In a large manufacturing facility with multiple processing "unit" areas, a tag such as this might be preceded by another number designating the unit area. For example, our hypothetical flow controller might be labeled 12-FC-135 (flow controller for loop #135, located in unit 12). If this loop happened to contain multiple controllers, we would need to distinguish them from each other by the use of suffix letters appended to the loop number (e.g. 12-FC-135A, 12-FC-135B, 12-FC-135C).

Each and every instrument within a particular loop is first defined by the variable that loop seeks to sense or control, regardless of the physical construction of the instrument itself. Our hypothetical flow controller FC-135, for example, may be physically identical to the level controller in loop #72 (LC-72), or to the temperature controller in loop #288 (TC-288). What makes FC-135 a *flow* controller is the fact that the transmitter sensing the main process variable measures *flow*. Likewise, the identifying tag for every other instrument within that loop[2] must begin with the letter "F" as well. This includes the final control element as well: in a level control loop, the transmitter is identified as an "LT" even if the actual sensing element works on *pressure* (because the variable that the loop strives to sense or control is actually level, even if indirectly sensed by pressure), the controller is identified as an "LC", and the control valve throttling fluid *flow* is identified as an "LV": every instrument in that level-controlling loop serves to help control *level*, and so its primary function is to be a "level" instrument.

[2]Exceptions do exist to this rule. For example, in a cascade or feedforward loop where multiple transmitters feed into one or more controllers, each transmitter is identified by the type of process variable *it* senses, and each controller's identifying tag follows suit.

Valid letters recognized by the ISA for defining the primary process variable of an instrument within a loop are shown in the following table. Please note that the use of a modifier defines a unique variable: for example, a "PT" is a transmitter measuring *pressure* at a single point in a process, whereas a "PDT" is a transmitter measuring a *pressure difference* between two points in a process. Likewise, a "TC" is a controller controlling temperature, whereas a "TKC" is a controller controlling the *rate-of-change of temperature*:

Letter	**Variable**	**Modifier**
A	Analytical (composition)	
B	Burner or Combustion	
C	*User-defined*	
D	*User-defined*	Differential
E	Voltage	
F	Flow	Ratio or Fraction
G	*User-defined*	
H	Hand (manual)	
I	Current	
J	Power	Scan
K	Time or Schedule	Time rate-of-change
L	Level	
M	*User-defined*	Momentary
N	*User-defined*	
O	*User-defined*	
P	Pressure or Vacuum	
Q	Quantity	Time-Integral or Total
R	Radiation	
S	Speed or Frequency	Safety
T	Temperature	
U	Multi-function	
V	Vibration	
W	Weight or Force	
X	*Unclassified*	X-axis
Y	Event, State, or Presence	Y-axis
Z	Position or Dimension	Z-axis

A "user-defined" letter represents a non-standard variable used multiple times in an instrumentation system. For example, an engineer designing an instrument system for measuring and controlling the *refractive index* of a liquid might choose to use the letter "C" for this variable. Thus, a refractive-index transmitter would be designated "CT" and a control valve for the refractive-index loop would be designated "CV". The meaning of a user-defined variable need only be defined in one location (e.g. in a legend for the diagram).

An "unclassified" letter represents one or more non-standard variables, each used only once (or a very limited number of times) in an instrumentation system. The meaning of an unclassified variable is best described immediately near the instrument's symbol rather than in a legend.

Succeeding letters in an instrument tag describe the function that instrument performs relative to the process variable. For example, a "PT" is an instrument *transmitting* a signal representing pressure, while a "PI" is an *indicator* for pressure and a "PC" is a *controller* for pressure. Many instruments have multiple functions designated by multiple letters, such as a TRC (Temperature *Recording Controller*). In such cases, the first function letter represents the "passive" function (usually provided to a human operator) while the second letter represents the "active" (automated) control function.

Letter	**Passive function**	**Active function**	**Modifier**
A	Alarm		
B	*User-defined*	*User-defined*	*User-defined*
C		Control	
E	Element (sensing)		
G	Glass or Viewport		
H			High
I	Indicate		
K		Control station	
L	Light		Low
M			Middle or Intermediate
N	*User-defined*	*User-defined*	*User-defined*
O	Orifice		
P	Test point		
R	Record		
S		Switch	
T		Transmit	
U	Multi-function	Multi-function	Multi-function
V		Valve, Damper, Louver	
W	Well		
X	*Unclassified*	*Unclassified*	*Unclassified*
Y		Relay, Compute, Convert	
Z		Driver, Actuator, or unclassified final control element	

References

"Commonly Used Electrical Symbols", Eaton Electrical Inc., Eaton Corporation, Moon Township, PA, 2005.

Instrumentation, Systems, and Automation Society Standards, 5.1-1984 (R1992), Instrumentation Symbols and Identification, Research Triangle Park, NC, 1984.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Measurement and Analysis Volume I*, Fourth Edition, CRC Press, New York, NY, 2003.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Software and Digital Networks*, Third Edition, CRC Press, New York, NY, 2002.

Chapter 8

Instrument connections

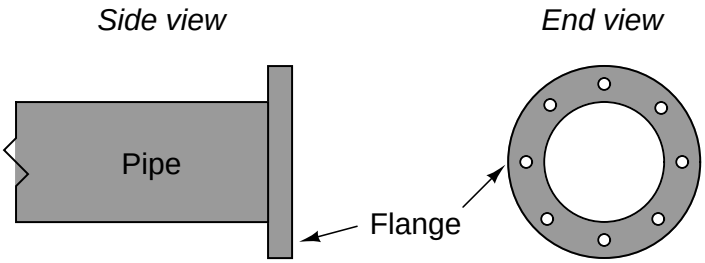
All instruments connect to their respective processes and to each other by means of pipe, tube, and/or wires. Improper installation of these connective lines can make the difference between success or failure in an installation. Safety is also impacted by improper connections between instruments and the process, and from instrument to instrument.

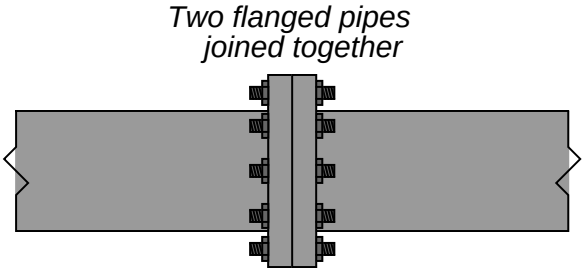
8.1 Pipe and pipe fittings

Pipe is a hollow structure designed to provide an enclosed pathway for fluids to flow, usually manufactured from cast metal (although plastic is a common pipe material for many industrial applications). This section discusses some of the more common methods for joining pipes together (and joining pipe ends to equipment such as pressure instruments).

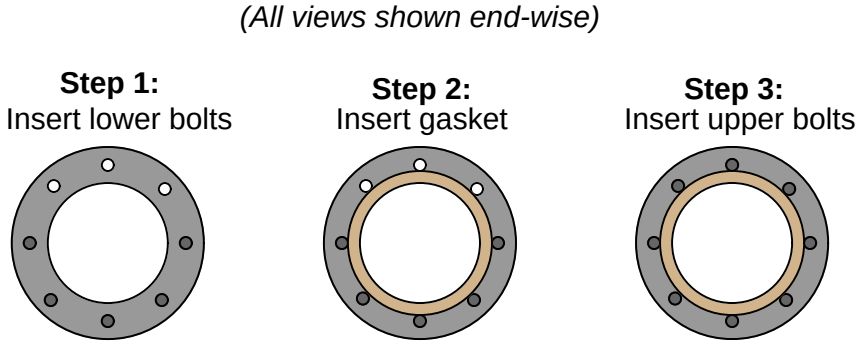
8.1.1 Flanged pipe fittings

In the United States of America, most large industrial pipes are joined together by *flanges*. A pipe "flange" is a ring of metal, usually welded to the end of a pipe, with holes drilled in it parallel to the pipe centerline to accept several bolts:

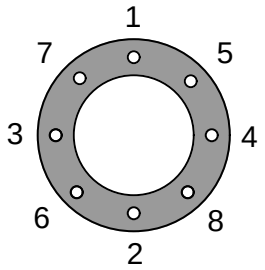




Flange joints are made pressure-tight by inserting a donut-shaped gasket between the flange pairs prior to tightening the bolts. A common method of installing such a flange gasket is to first install only half of the bolts (in the holes lower than the centerline of the pipe), drop the gasket between the flanges, then insert the rest of the bolts:



A very important procedure to observe when tightening the bolts holding two flanges together is to evenly distribute the bolt pressure, so that no single region of the flange receives significantly more bolt pressure than any other region. In an ideal world, you would tighten all bolts to the same torque limit *simultaneously*. However, since this is impossible with just a single wrench, the best alternative is to tighten the bolts in alternating sequence, in stages of increasing torque. An illustrative torque sequence is shown in the following diagram (the numbers indicate the order in which the bolts should be tightened):



With one wrench, you would tighten each bolt to a preliminary torque in the sequence shown. Then, you would repeat the tightening sequence with additional torque for a couple more cycles until all bolts had been tightened to the recommended torque value. Note how the torque sequence alternates between four quadrants of the flange, ensuring the flanges are evenly compressed together as all bolts are gradually tightened. This technique of alternating quadrants around the circle is often referred to as *cross-torquing*.

Special wrenches called *torque wrenches* exist for the purpose of measuring applied torque during the tightening process. In critical, high-pressure applications, special tools exist to infer bolt pressure by measuring how far each bolt *stretches* as it is tightened.

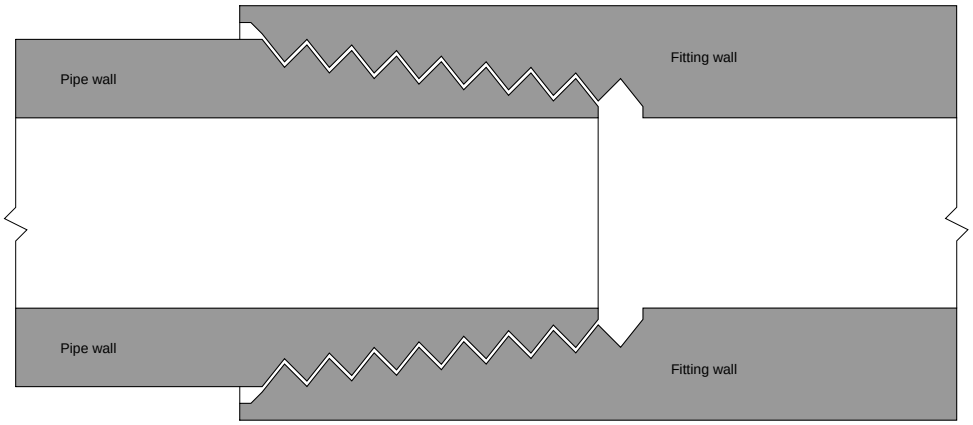
Another important procedure to observe when working with flanged pipe connections is to loosen the bolts on the *far* side of the flange before loosening the bolts on the side of the flange nearest you. This is strictly a precautionary measure against the spraying of process fluid toward your face or body in the event of stored pressure inside of a flanged pipe. By reaching over the pipe to first loosen flange bolts on the far side, if any pressure happens to be inside the pipe, it should leak there first, venting the pressure in a direction away from you.

8.1.2 Tapered thread pipe fittings

For smaller pipe sizes, *threaded fittings* are more commonly used to create connections between pipes and between pipes and equipment (including some instruments). A very common design of threaded pipe fitting is the *tapered* pipe thread design. The intent of a tapered thread is to allow the pipe and fitting to "wedge" together when engaged, creating a joint that is both mechanically rugged and leak-free.

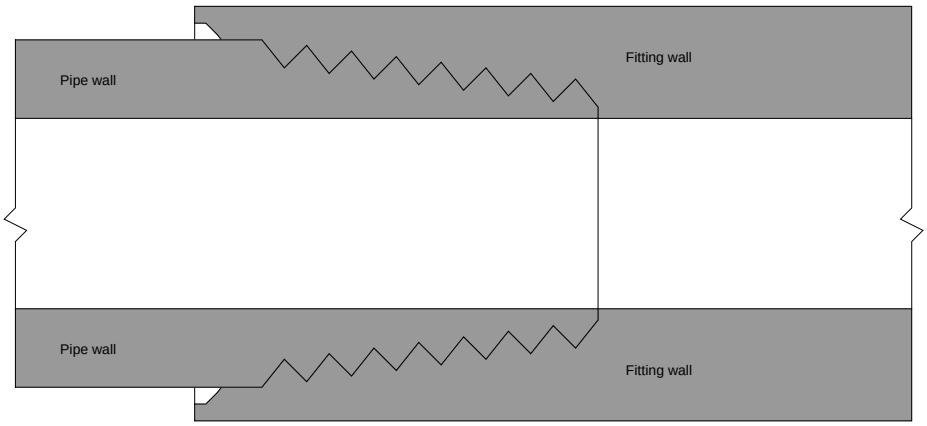
When male and female tapered pie threads are first engaged, they form a loose junction:

Threads loosely engaged



After tightening, however, the tapered profile of the threads acts to wedge both male and female pieces tightly together as such:

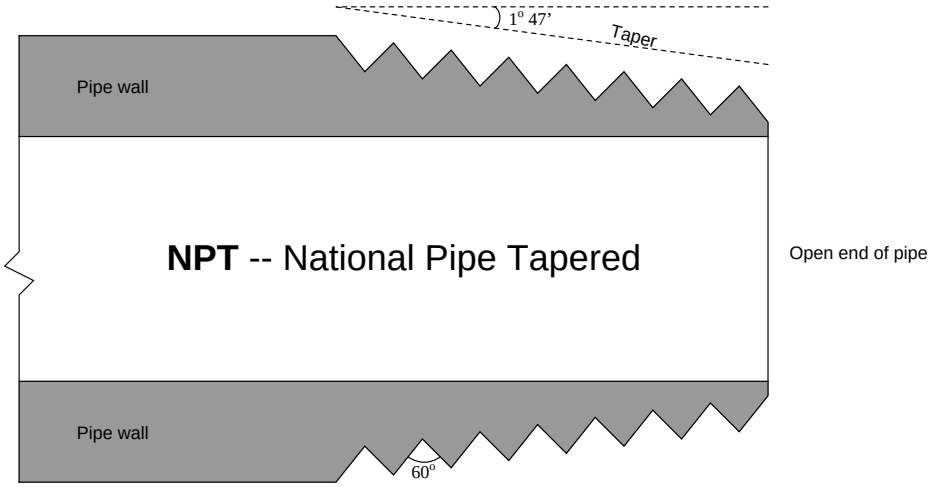
Threads fully engaged -- pressure-tight seal established



Several different standards exist for tapered-thread pipe fittings. For each standard, the angle of the thread is fixed, as is the angle of taper. Thread *pitch* (the number of threads per unit length)

varies with the diameter of the pipe fitting[1].

In the United States, the most common tapered thread standard for general-purpose piping is the *NPT*, or *National Pipe Taper* design. NPT threads have an angle of 60^o and a taper of 1^o 47' (1.7833^o):



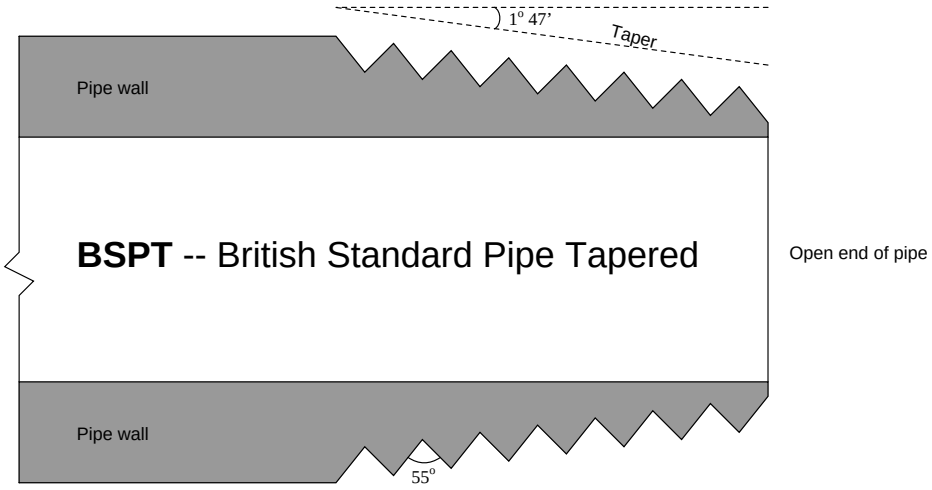
NPT pipe threads must have some form of *sealant* applied prior to assembly to ensure pressure-tight sealing between the threads. Teflon tape and various liquid pipe "dope" compounds work well for this purpose. Sealants are necessary with NPT threads for two reasons: to lubricate the male and female pieces (to guard against galling the metal surfaces), and also to fill the spiral gap formed between the root of the female thread and the crest of the male thread (and visa-versa).

NPTF (National Pipe Thread) pipe threads are engineered with the same thread angle and pitch as NPT threads, but carefully machined to avoid the spiral leak path inherent to NPT threads. This design – at least in theory – avoids the need to use sealant with NPTF threads to achieve a pressure-tight seal between male and female pieces, which is why NPTF threads are commonly referred to as *dryseal*. However, in practice it is still recommended that some form of sealant be used (or at the very least some form of thread *lubricant*) in order to achieve reliable sealing.

ANPT (Aeronautical National Pipe Tapered) is identical to NPT, except with a greater level of precision and quality for its intended use in aerospace and military applications.

[1] For example, 1/8 inch NPT pipe fittings have a thread pitch of 27 threads per inch. 1/4 inch and 3/8 inch NPT fittings are 18 threads per inch, 1/2 inch and 3/4 inch NPT fittings are 14 threads per inch, and 1 inch through 2 inch NPT fittings are 11.5 threads per inch.

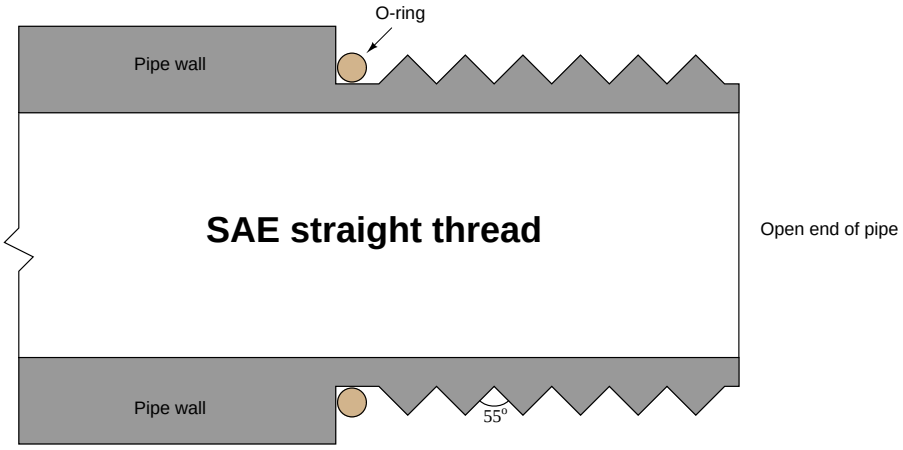
Another tapered-thread standard is the *BSPT*, or *British Standard Pipe Tapered*. BSPT threads have a narrower thread angle than NPT threads (55^o instead of 60^o) but the same taper of 1^o 47' (1.7833^o):



8.1.3 Parallel thread pipe fittings

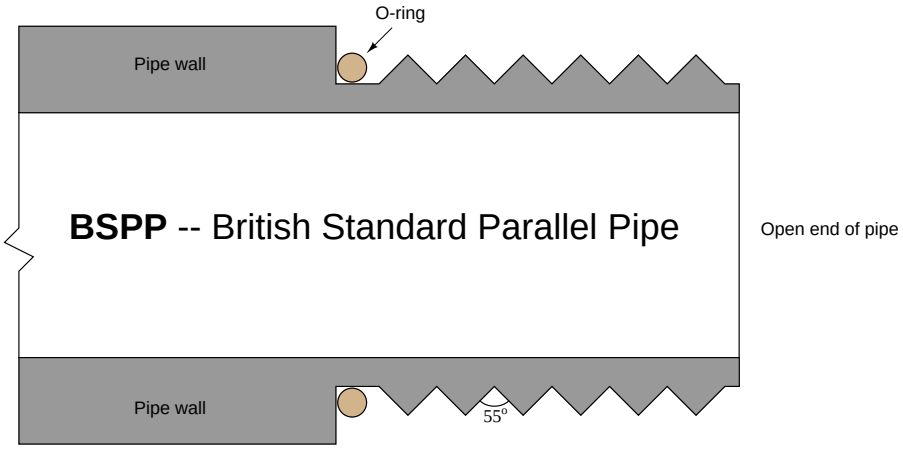
An alternative to tapered threads in pipe joints is the use of parallel threads, similar to the threads of machine screws and bolts. Since parallel threads are incapable of forming a pressure-tight seal on their own, the sealing action of a parallel thread pipe fitting must be achieved some other way. This function is usually met with an O-ring or gasket.

In the United States, a common design of parallel-thread pipe fitting is the *SAE straight thread*, named after the *Society of Automotive Engineers*:



Sealing is accomplished as the O-ring is compressed against the shoulder of the female fitting. The threads serve only to provide force (not fluid sealing), much like the threads of a fastener.

Another parallel-thread pipe standard is the *BSPP*, or *British Standard Pipe Parallel*. Like the BSPT (tapered) standard, the thread angle of BSPP is 55^o. Like the SAE parallel-thread standard, sealing is accomplished by means of an O-ring which compresses against the shoulder of the matching female fitting:

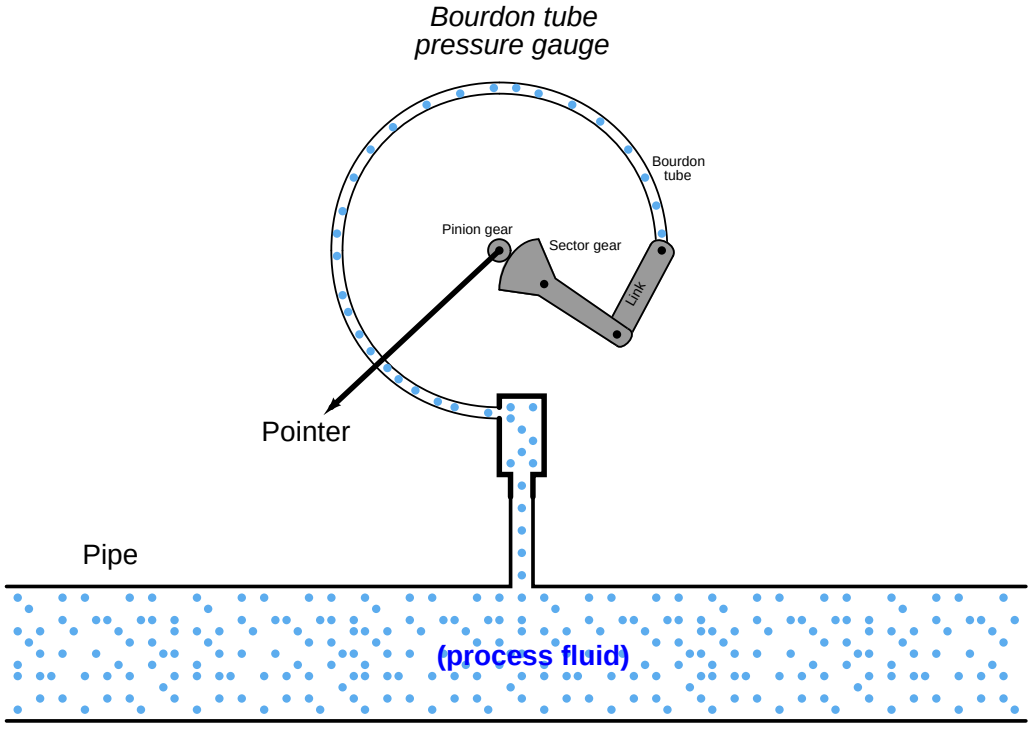


8.1.4 Sanitary pipe fittings

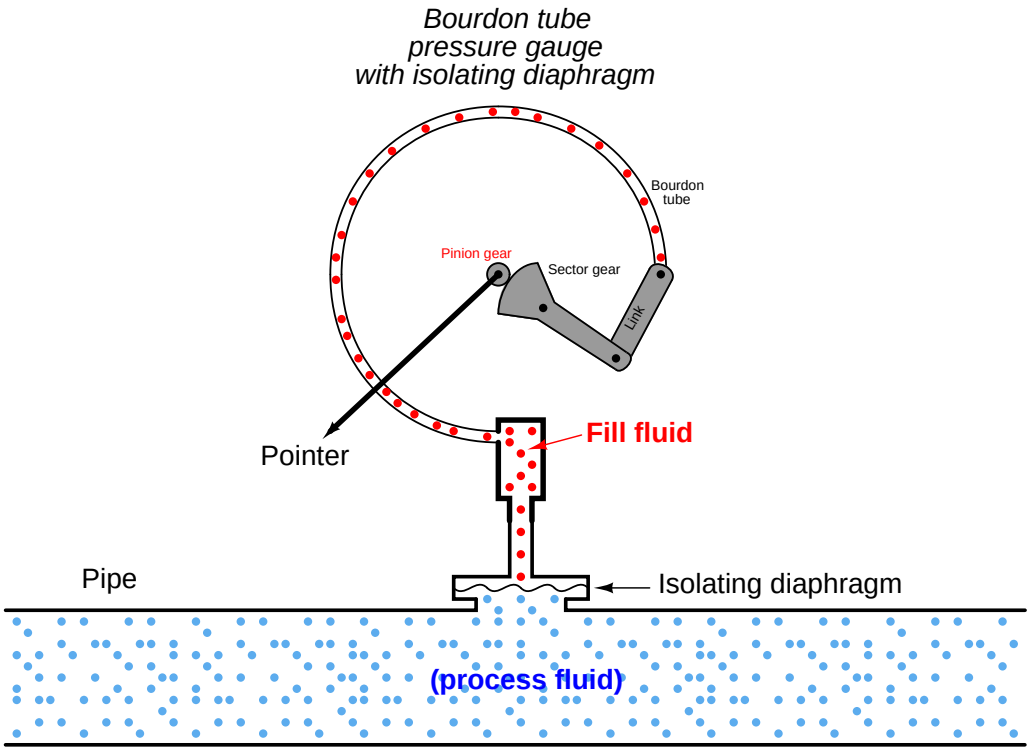
Food processing, pharmaceuticals manufacturing, and biological research processes are naturally sensitive to the presence of micro-organisms such as bacteria, fungi, and algae. It is important in these processes to ensure the absence of harmful micro-organisms, for reasons of both human health and quality control. For this reason, the process piping and vessels in these industries is designed first and foremost to be thoroughly cleaned without the need for disassembly. Regular cleaning and sterilization cycles are planned and executed between production schedules (batches) to ensure no colonies of harmful micro-organisms can grow.

A common *Clean-In-Place* (CIP) protocol consists of flushing all process piping and vessels with alternating acid and caustic solutions, then washing with purified water. For increased sanitization, a *Steam-In-Place* (SIP) cycle may be incorporated as well, flushing all process pipes and vessels with hot steam to ensure the destruction of any micro-organisms.

An important design feature of any sanitary process is the elimination of any "dead ends" (often called *dead legs* in the industry), crevices, or voids where fluid may collect and stagnate. This includes any instruments contacting the process fluids. It would be unsafe, for example, to connect something as simple as a bourdon-tube pressure gauge to a pipe carrying biologically sensitive fluid(s), since the interior volume of the bourdon tube will act as a stagnant refuge for colonies of micro-organisms to grow:



Instead, any pressure gauge must use an isolating diaphragm, where the process fluid pressure is transferred to the gauge mechanism through a sterile "fill fluid" that never contacts the process fluid:



With the isolating diaphragm in place, there are no stagnant places for process fluid to collect and avoid flushing by CIP or SIP cycles.

Standard pipe fittings are problematic in sanitary systems, as tiny voids between the mating threads of male and female pipe fittings may provide refuge for micro-organisms. To avoid this problem, special *sanitary fittings* are used instead. These fittings consist of a matched pair of flanges, held together by an external clamp. An array of sanitary fittings on an instrument test bench appear in the following photograph:

The next photograph shows the installation of a pressure transmitter on an ultra-pure water line using one of these sanitary fittings. The external clamp holding the two flanges together is clearly visible in this photograph:

Sanitary pipe fittings are not limited to instrument connections, either. Here are two photographs of process equipment (a ball valve on the left, and a pump on the right) connected to process pipes using sanitary fittings:

8.2 Tube and tube fittings

Tube, like pipe, is a hollow structure designed to provide an enclosed pathway for fluids to flow. In the case of tubing, it is usually manufactured from rolled or extruded metal (although plastic is a common tube material for many industrial applications). This section discusses some of the more common methods for joining tubes together (and joining tube ends to equipment such as pressure instruments).

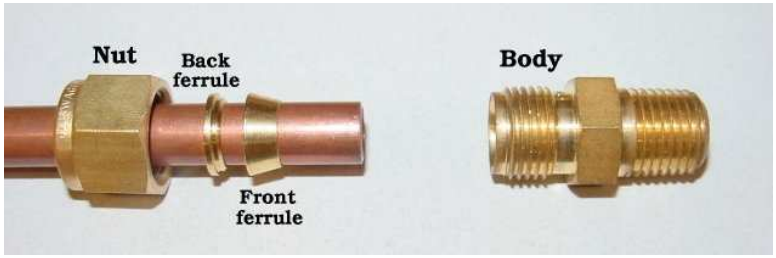
One of the fundamental differences between tube and pipe is that tube is *never* threaded at the end to form a connection. Instead, a device called a *tube fitting* must be used to couple a section of tube to another tube, or to a section of pipe, or to a piece of equipment (such as an instrument). Unlike pipes which are thick-walled by nature, tubes are thin-walled structures. The wall thickness of a typical tube is simply too thin to support threads.

Tubes are generally favored over pipe for small-diameter applications. The ability for skilled workers to readily cut and bend tube with simple hand tools makes it the preferred choice for connecting instruments to process piping. When used as the connecting units between an instrument and a process pipe or vessel, the tube is commonly referred to as an *impulse tube* or *impulse line*[2].

[2]Impulse lines are alternatively called *gauge lines* or *sensing lines*.

8.2.1 Compression tube fittings

By far the most common type of tube fitting for instrument impulse lines is the *compression-style* fitting, which uses a compressible *ferrule* to perform the task of sealing fluid pressure. The essential components of a compression tube fitting are the *body*, the *ferrule*, and the *nut*. The ferrule and body parts have matching conical profiles designed to tightly fit together, forming a pressure-tight metal-to-metal seal. Some compression fitting designs use a two-piece ferrule assembly, such as this tube fitting shown here[3] (prior to full assembly):



Just prior to assembly, we see how the nut will cover the ferrule components and push them into the conical entrance of the fitting body:

After properly tightening the nut, the ferrule(s) will *compress* onto the outside circumference of the tube, slightly crimping the tube in the process and thereby locking the ferrules in place:

[3] This happens to be a Swagelok brass instrument tube fitting being installed on a 3/8 inch copper tube.

When assembling compression-style tube fittings, you should always precisely follow the manufacturer's instructions to ensure correct compression. For Swagelok-brand instrument tube fittings 1 inch in size and smaller, the general procedure is to tighten the nut 1-1/4 turns past finger-tight. Insufficient turning of the nut will fail to properly compress the ferrule around the tube, and excessive turning will over-compress the ferrule, resulting in leakage. Swagelok also provides special gauges which may be used to measure proper ferrule compression during the assembly process.

Parker is another major manufacturer[4] of instrument tube fittings, and their product line uses a single-piece ferrule instead of the two-piece design preferred by Swagelok. Like Swagelok fittings, Parker instrument fitting sized 1/4 inch to 1 inch require 1-1/4 turns past hand tight to properly compress the ferrule around the circumference of the tube. Parker also sells gauges which may be used to precisely determine when the proper amount of ferrule compression is achieved.

Regardless of the brand, compression-style instrument tube fittings are incredibly strong and versatile. Unlike pipe fittings, tube fittings may be disconnected and reconnected with ease. No special procedures are required to "re-make" a disassembled instrument fitting connection: merely tighten the nut "snug" to maintain adequate force holding the ferrule to the fitting body, but not so tight that the ferrule compresses further around the tube than it did during initial assembly.

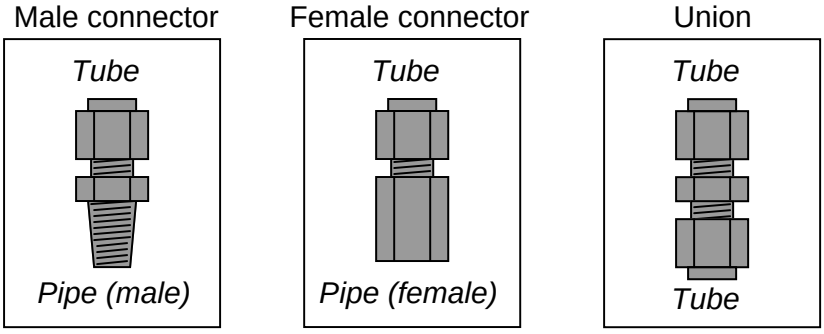
A very graphic illustration of the strength of a typical instrument tube fitting is shown in the following photograph, where a short section of 3/8 inch stainless steel instrument tube was exposed to high liquid pressure until it ruptured. Neither compression fitting on either side of the tube leaked during the test, despite the liquid pressure reaching a peak of 23,000 PSI before rupturing the tube[5]:

[4] So is Gyrolok, Hoke, and a host of others. It is not my intent to advertise for different manufacturers in this textbook, but merely to point out some of the more common brands an industrial instrument technician might encounter on the job.

[5] It should be noted that the fitting nuts became seized onto the tube due to the tube's swelling. The tube fittings may not have leaked during the test, but their constituent components should never be placed into service again!

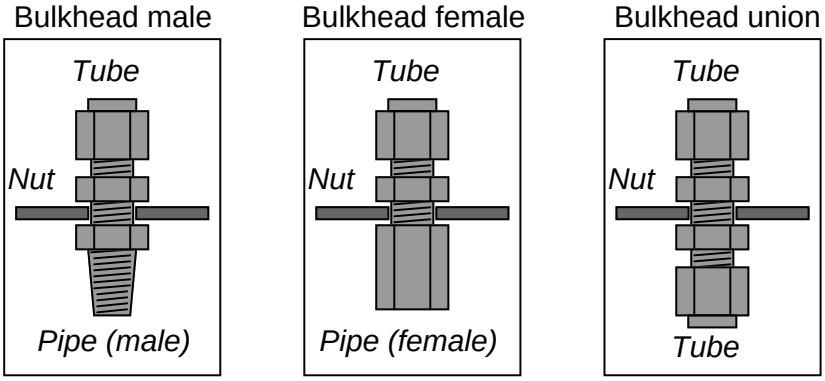
8.2.2 Common tube fitting types and names

Tube fittings designed to connect a tube to pipe threads are called *connectors*. Tube fittings designed to connect one tube to another are called *unions*:

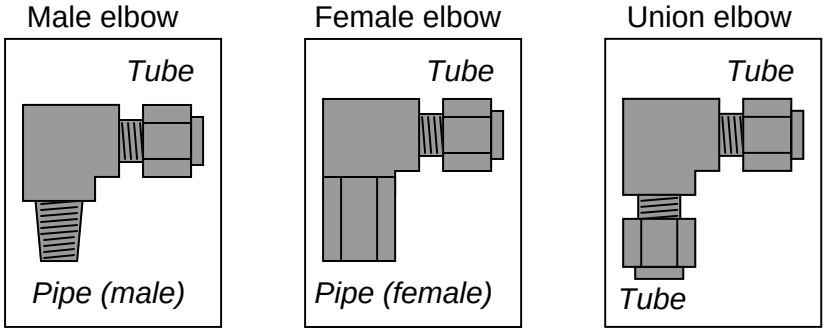


If a tube union joins together different tube sizes rather than tubes of the same size, it is called a *reducing union*.

A variation on the theme of tube connectors and unions is the *bulkhead* fitting. Bulkhead fittings are designed to fit through holes drilled in panels or enclosures to provide a way for a fluid line to pass through the wall of the panel or enclosure. In essence, the only difference between a bulkhead fitting and a normal fitting is the additional length of the fitting "barrel" and a special nut used to lock the fitting into place in the hole. The following illustration shows three types of bulkhead fittings:

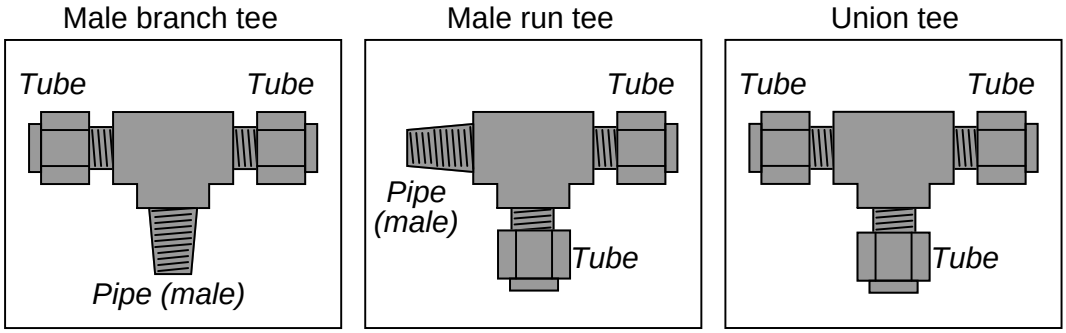


Tubing *elbows* are tube connectors with a bend. These are useful for making turns in tube runs without having to bend the tubing itself. Like standard connectors, they may terminate in male pipe thread, female pipe threads, or in another tube end:

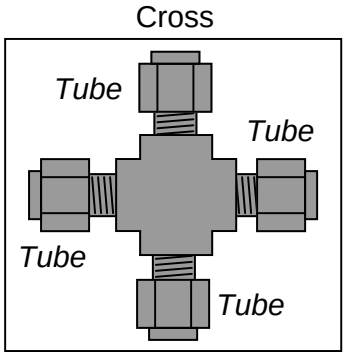


These elbows shown in the above illustration are all 90^o, but this is not the only angle available. 45^o elbows are also common.

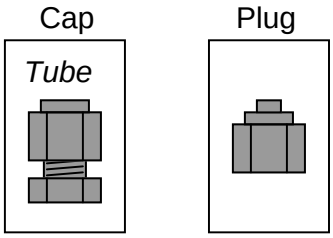
Tee fittings join three fluid lines together. Tees may have one pipe end and two tube ends (*branch* tees and *run* tees), or three tube ends (*union* tees). The only difference between a branch tee and a run tee is the orientation of the pipe end with regard to the two tube ends:



Of course, branch and run tee fittings also come in female pipe thread versions as well. A variation of the theme of union tees is the *cross*, joining four tubes together:



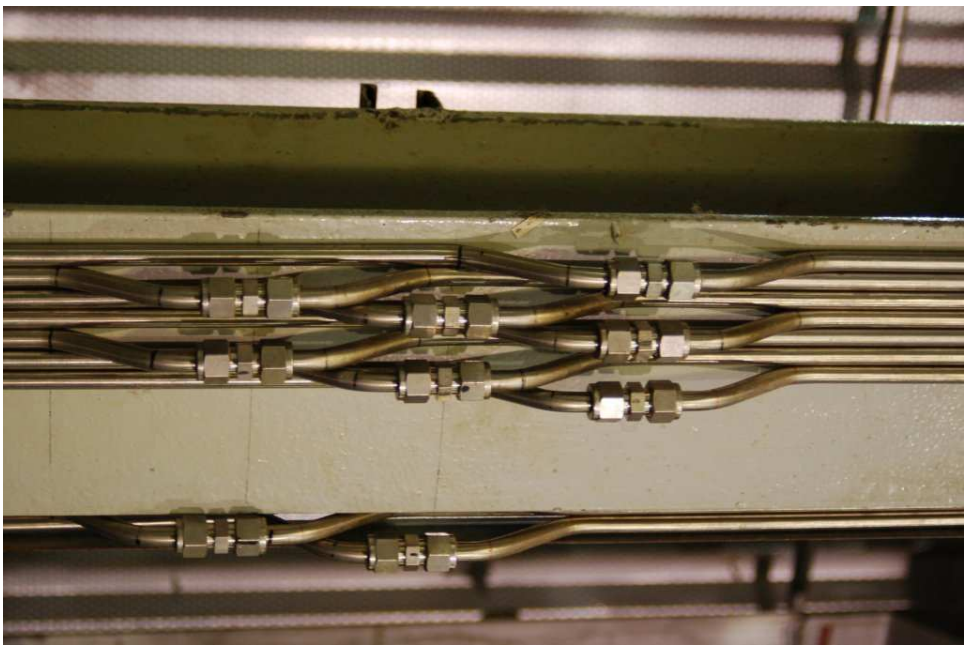
Special tube fittings are made to terminate tube connections, so they are sealed up instead of open. A piece designed to seal off the open end of a tube fitting is called a *plug*, while a piece designed to seal off the end of an open tube is called a *cap*:



8.2.3 Bending instrument tubing

Tube bending is something of an art, especially when done with stainless steel tubing. It is truly magnificent to see a professionally-crafted array of stainless steel instrument tubes, all bends perfectly made, all terminations square, all tubes parallel when laid side by side and perfectly perpendicular when crossing.

If possible, a goal in tube bending is to eliminate as many connections as possible. Connections invite leaks, and leaks are problematic. Long runs of instrument tubing made from standard 20 foot tube sections, however, require junctions be made somewhere, usually in the form of tube *unions*. When multiple tube unions must be placed in parallel tube runs, it is advisable to offset the unions so it is easier to get a wrench around the tube nuts to turn them. The philosophy here, *as always*, is to build the tubing system with future work in mind. A photograph of several tube junctions shows one way to do this:

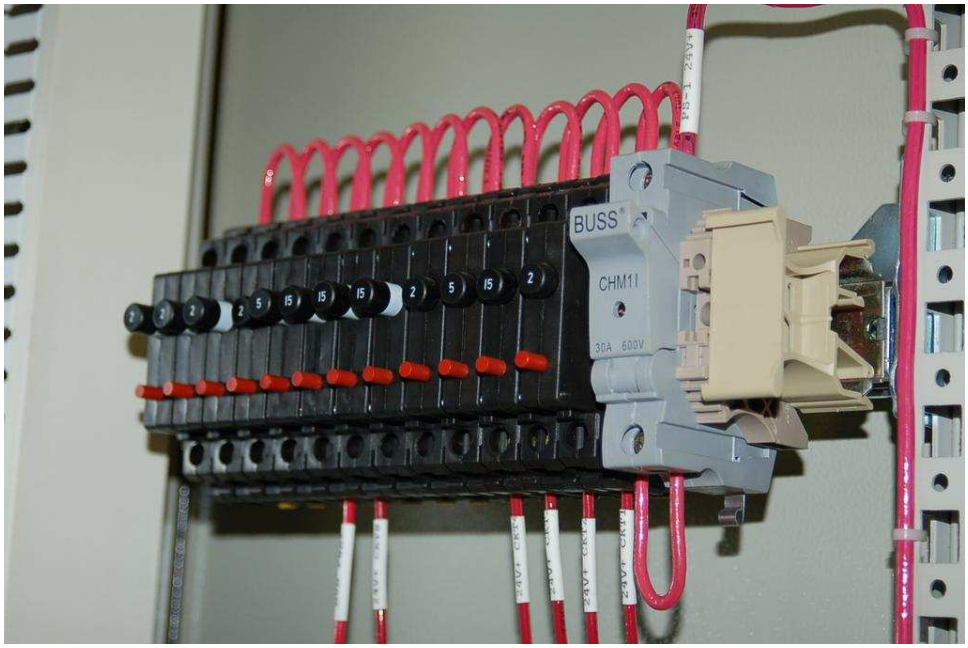


8.3 Electrical signal and control wiring

There is much to be said for neatness of assembly in electrical signal wiring. Even though the electrons don't "care" how neatly the wires are laid in place, human beings who must maintain the system certainly do. Not only are neat installations easier to navigate and troubleshoot, but they tend to inspire a similar standard of neatness when alterations are made[6].

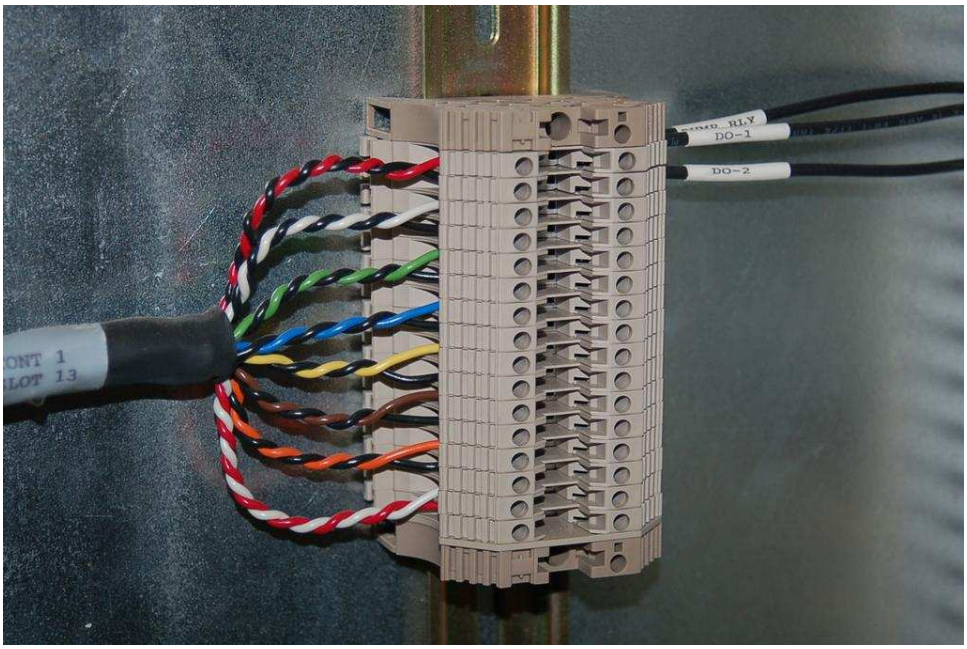
The following photographs illustrate excellent wiring practice. Study them carefully, and strive to emulate the same level of professionalism in your own work!

Here we see 120 volt AC power distribution wiring. Note how the hoop-shaped "jumper" wires are all cut to (nearly) the same length, and how each of the wire labels is oriented such that the printing is easy to read:



[6] No one wants to become known as the person who "messed up" someone else's neat wiring job!

This next photograph shows a great way to terminate multi-conductor signal cable to terminal blocks. Each of the pairs was twisted together using a hand drill set to very slow speed. Note how the end of the cable is wrapped in a short section of heat-shrink tubing for a neat appearance:

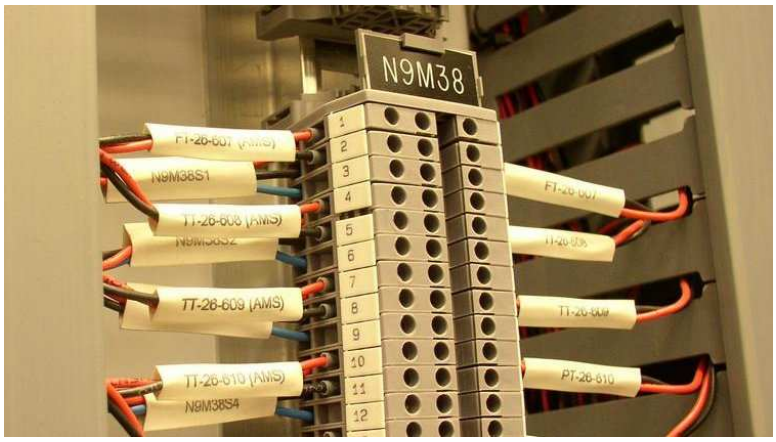


Beyond aesthetic preferences for instrument signal wiring are several practices based on sound electrical theory. The following subsections describe and explain these wiring practices.

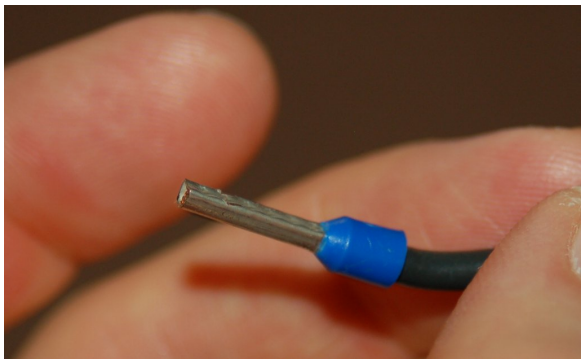
8.3.1 Connections and wire terminations

Many different techniques exist for connecting electrical conductors together: twisting, soldering, crimping (using compression connectors), and clamping (either by the tension of a spring or under the compression of a screw) are popular examples. Most industrial field wiring connections utilize a combination of compression-style crimp "lugs" and screw terminals to attach wires to instruments and to other wires.

The following photograph shows a typical *terminal strip* or *terminal block* array whereby twisted-pair signal cables connect to other twisted-pair signal cables:

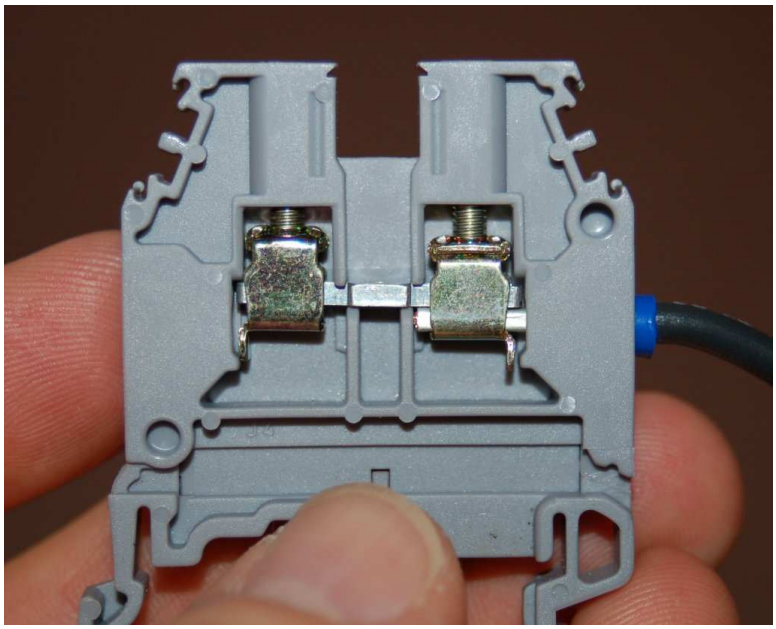


If you look closely at this photograph, you can see the bases of crimp-style compression lugs at the ends of the wires, just where they insert into the terminal block modules. These terminal blocks use screws to apply force which holds the wires in close electrical contact with a metal bar inside each block, but straight lugs have been crimped on the end of each wire to provide a more rugged tip for the terminal block screw to hold to. A close-up view shows what one of these straight compression lugs looks like on the end of a wire:

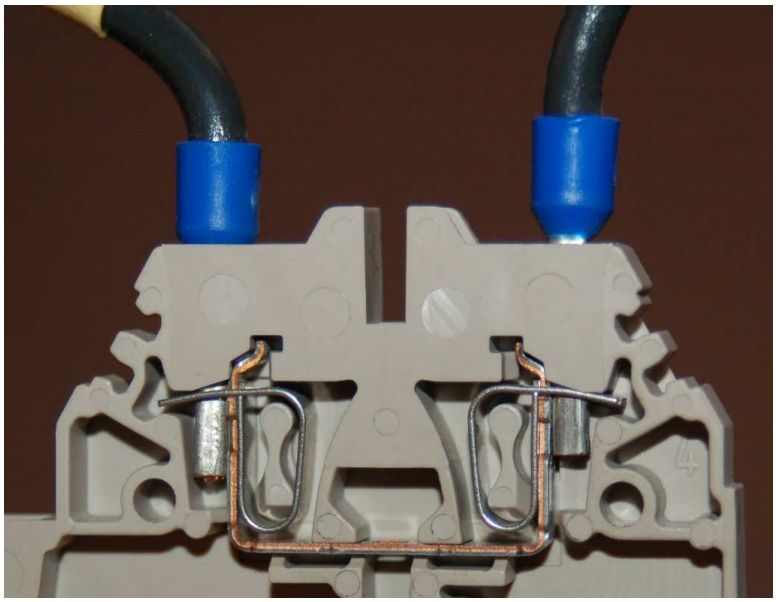


Also evident in this photograph is the dual-level connection points on the left-hand side of each terminal block. Two pairs of twisted signal conductors connect on the left-hand side of each terminal block pair, where only one twisted pair of wires connects on the right-hand side. This also explains why each terminal block section has two screw holes on the left but only one screw hole on the right.

A close-up photograph of a single terminal block module shows how the screw-clamp system works. Into the right-hand side of this block a single wire (tipped with a straight compression lug) is clamped securely. No wire is inserted into the left-hand side:



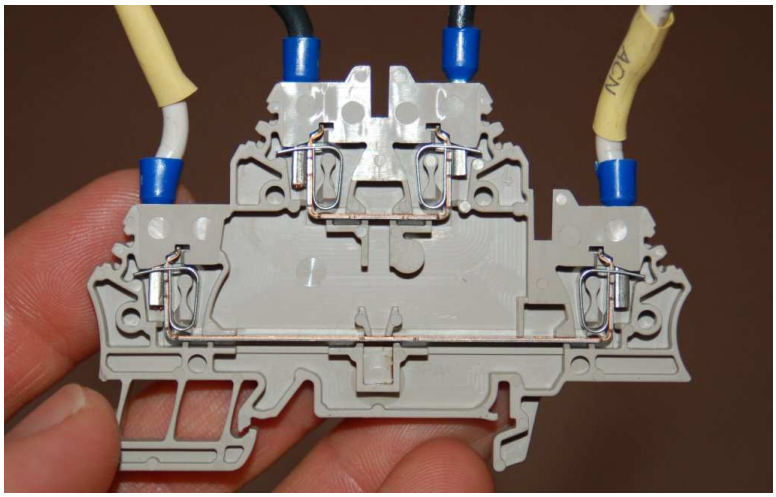
Some terminal blocks are *screwless*, using a spring clip to make firm mechanical and electrical contact with the wire's end:



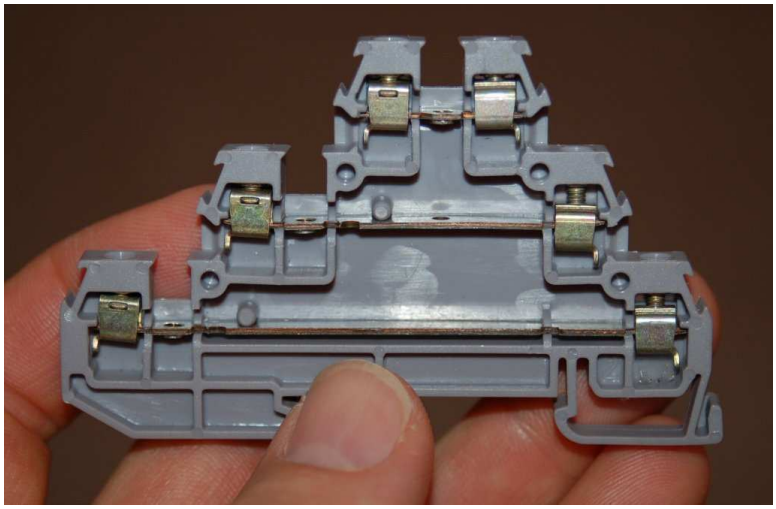
In order to extract or insert a wire end from or two a "screwless" terminal block, you must insert a narrow screwdriver into a hole in the block near the insertion point, then pivot the screwdriver (like a lever) to exert force on the spring clip. Screwless terminal blocks are generally faster to terminate and un-terminate than screw type terminal blocks, and the pushing action of the release tool is gentler on the body[7] than the twisting action required to loosen and tighten screws.

[7] An occupational hazard for technicians performing work on screw terminations is *carpal tunnel syndrome*, where repetitive wrist motion (such as the motions required to loosen and tighten screw terminals) damages portions of the wrist where tendons pass.

Many different styles of modular terminal blocks are manufactured to suit different wiring needs. Some terminal block modules, for example, have multiple "levels" instead of just one. The following photograph shows a two-level terminal block with screwless wire clamps:

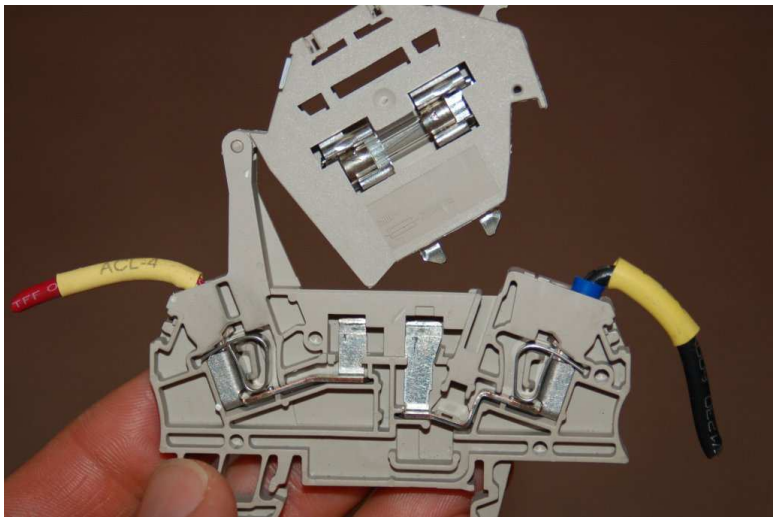


The next photograph shows a three-level terminal block with screw type clamps:

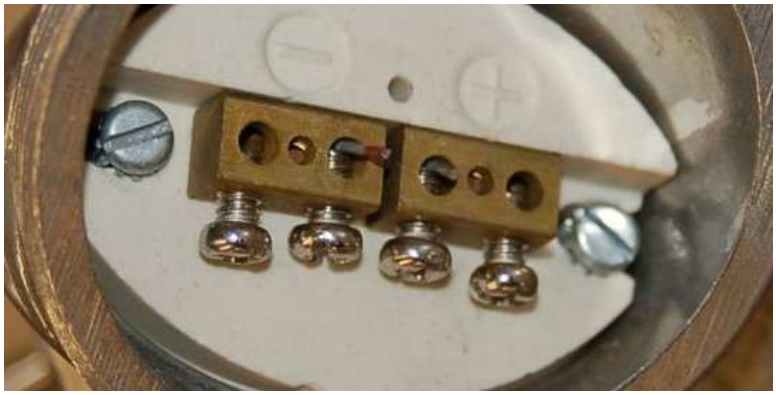


Some multi-level terminal blocks provide the option of *internal jumpers* to connect two or more levels together so they will be electrically common instead of electrically isolated.

Other modular terminal blocks include such features as LED indicator lamps, switches, fuses, and even resettable circuit breakers in their narrow width, allowing the placement of actual circuit components near connection points. The following photograph shows a swing-open fused terminal block module, in the open position:



Modular terminal blocks are useful for making connections with both solid-core and stranded metal wires. The clamping force applied to the wire's tip is direct, with no sliding or other motions involved. Some terminal blocks, however, are less sophisticated in design. This next photograph shows a pair of "isothermal" terminals designed to connect thermocouple wires together. Here you can see how the bare tip of the screw applies pressure to the wire inserted into the block:



The rotary force applied to each wire's tip by these screws necessitates the use of solid wire. Stranded wire would become frayed by this combination of forces.

Many field instruments, however, do not possess "block" style connection points at all. Instead, they are equipped with pan-head machine screws designed to compress the wire tips directly between the heads of the screws and a metal plate below.

Solid wires may be adequately joined to such a screw-head connection point by partially wrapping the bare wire end around the screw's circumference and tightening the head on top of the wire, as is the case with the two short wire stubs terminated on this instrument:

The problem with directly compressing a wire tip beneath the head of a screw is that the tip is subjected to both compressive and shear forces. As a result, the wire's tip tends to become mangled with repeated connections. Furthermore, tension on the wire will tend to turn the screw, potentially loosening it over time.

This termination technique is wholly unsuitable for stranded wire[8], because the shearing forces caused by the screw head's rotation tends to "fray" the individual strands. The best way to attach a stranded wire tip directly to a screw-style connection point is to first crimp a compression-style *terminal* to the wire. The flat metal "lug" portion of the terminal is then inserted underneath the screw head, where it can easily handle the shearing and compressive forces exerted by the head.

[8] An exception is when the screw is equipped with a square washer underneath the head, designed to compress the end of a stranded wire with no shear forces. Many industrial instruments have termination points like this, for the express purpose of convenient termination to either solid or stranded wire ends.

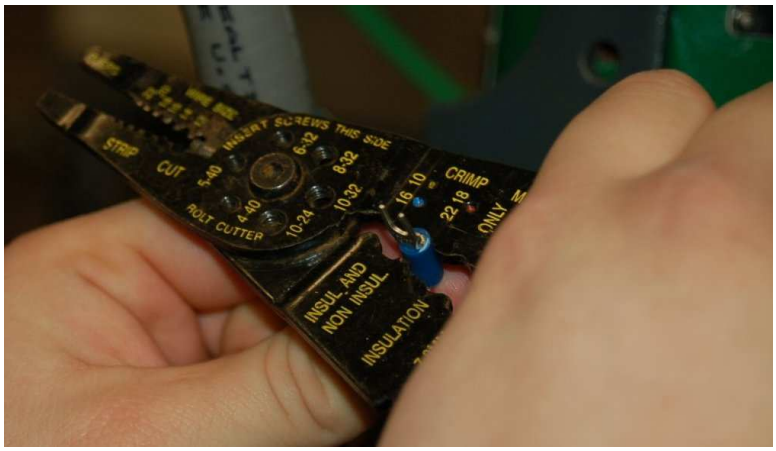
This next photograph shows five such stranded-copper wires connected to screw-style connection points on a field instrument using compression-style terminals:

Compression-style terminals come in two basic varieties: *fork* and *ring*. An illustration of each type is shown here:

Fork terminals are easier to install and remove, since they merely require loosening of the connector screw rather than removal of the screw. Ring terminals are more secure, since they cannot "fall off" the connection point if the screw ever accidently loosens.

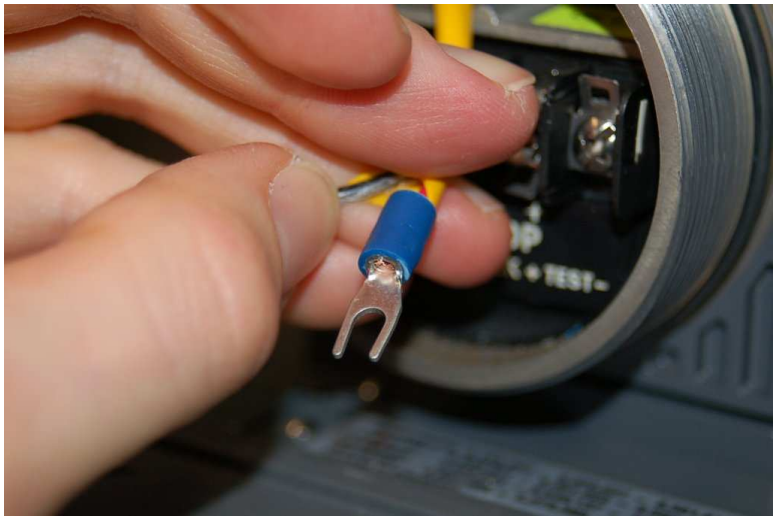
Just as direct termination underneath a screw head is wholly unsuitable for stranded wires, compression-style terminals are wholly unsuitable for solid wire. Although the initial crimp may feel secure, compression terminals lose their tension rapidly on solid wire, especially when there is any motion or vibration stressing the connection. Compression wire terminals should only be crimped to stranded wire!

Properly installing a compression-type terminal on a wire end requires the use of a special *crimping* tool. The next photograph shows one of these tools in use:



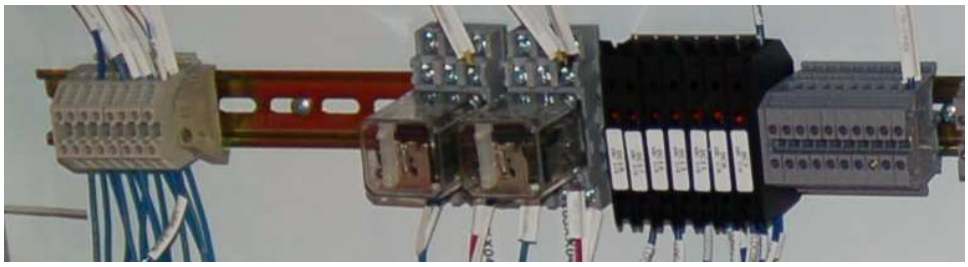
Note the different places on the crimping tool, labeled for different wire sizes (gauges). One location is used for 16 gauge to 10 gauge wire, while the location being used in the photograph is for wire gauges 22 through 18 (the wire inside of the crimped terminal happens to be 18 gauge).

This particular version of a "crimping" tool performs most of the compression on the underside of the terminal barrel, leaving the top portion undisturbed. The final crimped terminal looks like this when viewed from the top:

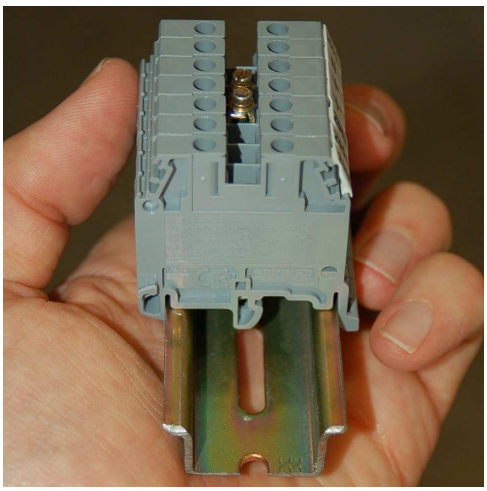


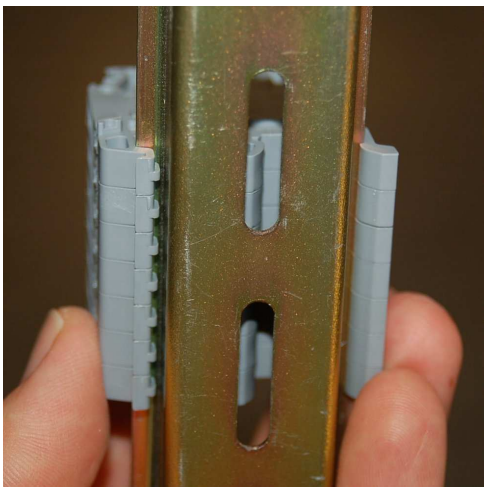
8.3.2 DIN rail

An industry-standard structure for attaching terminal blocks and small electrical components to flat metal panels is something called a *DIN rail*. This is a narrow channel of metal – made of bent sheet steel or extruded aluminum – with edges designed for plastic components to "clip" on. The following photograph shows terminal blocks, relay sockets, fuses, and more terminal blocks mounted to a horizontal length of DIN rail in a control system enclosure:



Two photographs of a terminal block cluster clipped onto a length of DIN rail – one from above and one from below – shows how specially-formed arms on each terminal block module fit the edges of the DIN rail for a secure attachment:





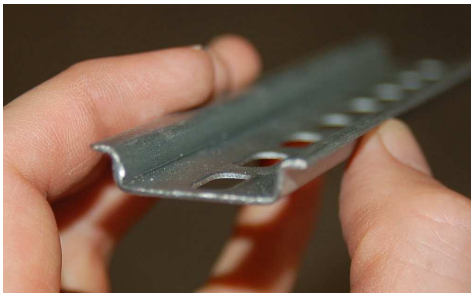
The DIN rail itself mounts on to any flat surface by means of screws inserted through the slots in its base. In most cases, the flat surface in question is the metal subpanel of an electrical enclosure to which all electrical components in that enclosure are attached.

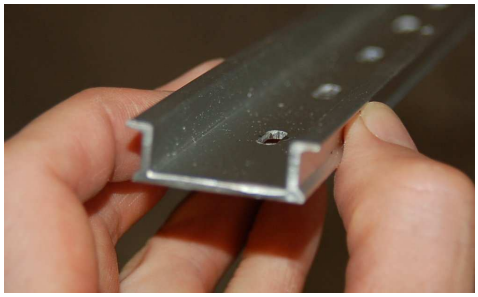
An obvious advantage of using DIN rail to secure electrical components versus individually attaching those components to a subpanel with their own sets of screws is convenience: much less labor is required to mount and unmount a DIN rail-attached component than a component attached with its own set of dedicated screws. This convenience significantly eases the task of altering a panel's configuration. With so many different devices manufactured for DIN rail mounting, it is easy to

upgrade or alter a panel layout simply by unclipping components, sliding them to new locations on the rail, or replacing them with other types or styles of components.

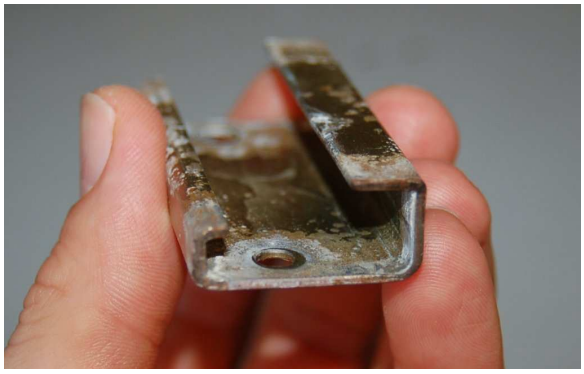
This next photograph shows some of the diversity available in DIN rail mount components. From left to right we see four relays, a power supply, and three HART protocol converters, all clipped to the same extruded aluminum DIN rail:

As previously mentioned, DIN rail is available in both stamped sheet-steel and extruded aluminum forms. A comparison of the two materials is shown here, with sheet steel on the left and aluminum on the right:

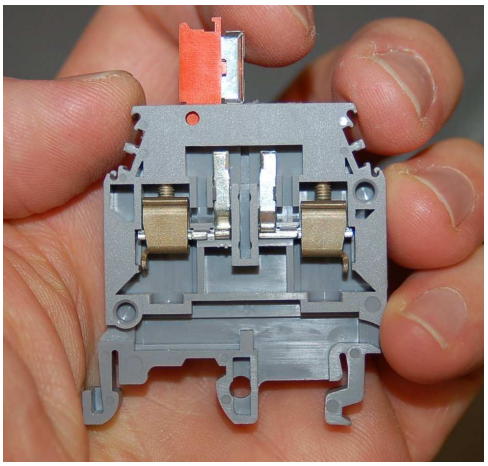


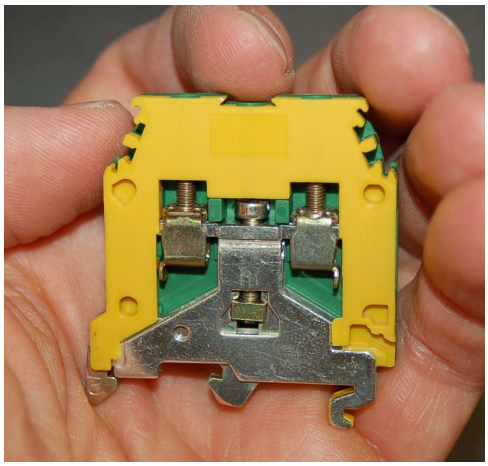


The form of DIN rail shown in all photographs so far is known as "top hat" DIN rail. A variation in DIN rail design is the so-called "G" rail, with a markedly different shape:



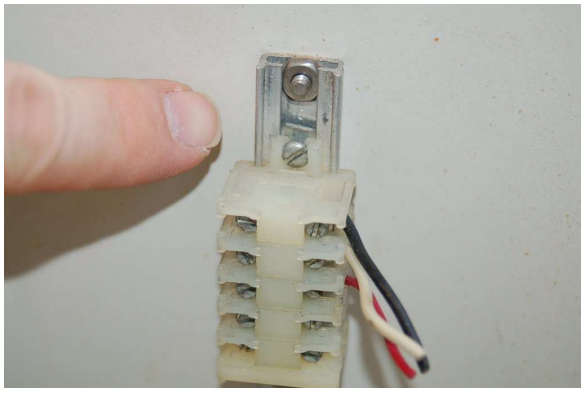
Fortunately, many modular terminal blocks are formed with the ability to clip to either style of DIN rail, such as these two specialty blocks, the left-hand example being a terminal block with a built-in disconnect switch, and the right-hand example being a "grounding" terminal block whose termination points are electrically common to the DIN rail itself:



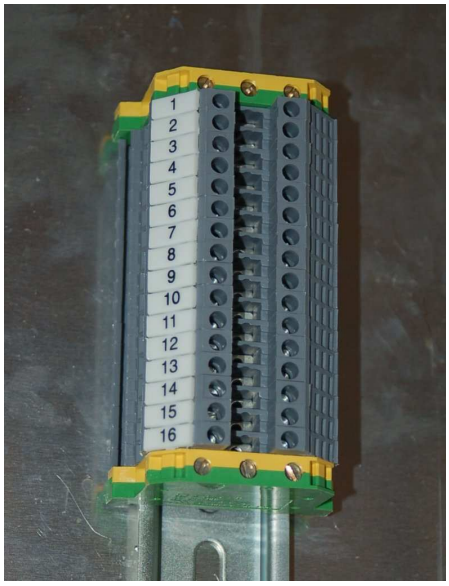


If you examine the bottom structure of each block, you will see formations designed to clip either to the edges of a standard ("top hat") DIN rail or to a "G" shaped DIN rail.

Smaller DIN rail standards also exist, although they are far less common than the standard 35mm size:



A nice feature of many DIN rail type terminal blocks is the ability to attach pre-printed terminal numbers. This makes documentation of wiring much easier, with each terminal connection having its own unique identification number:

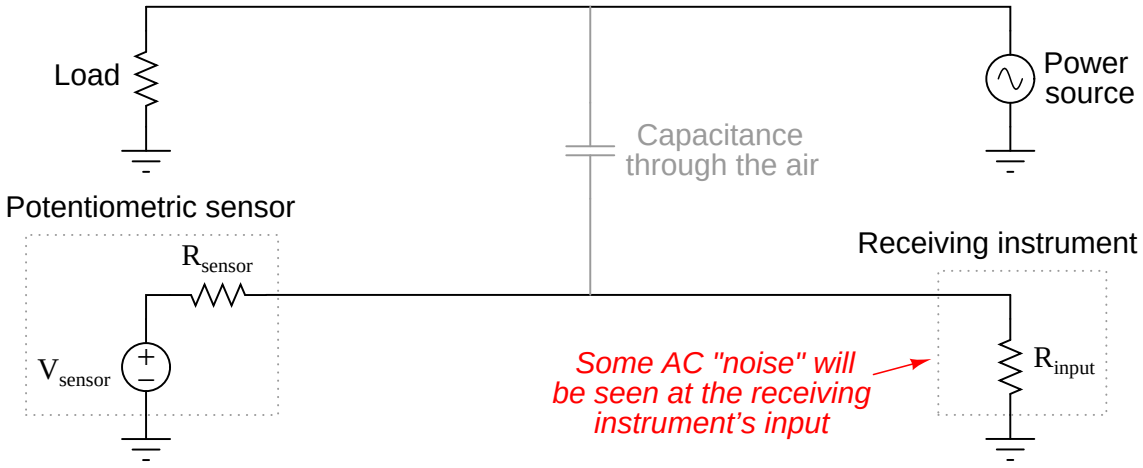


8.3.3 Signal coupling and cable separation

If sets of wires lie too close to one another, electrical signals may "couple" from one wire (or set of wires) to the other(s). This can be especially detrimental to signal integrity when the coupling occurs between AC power conductors and low-level instrument signal wiring such as thermocouple or pH sensor cables.

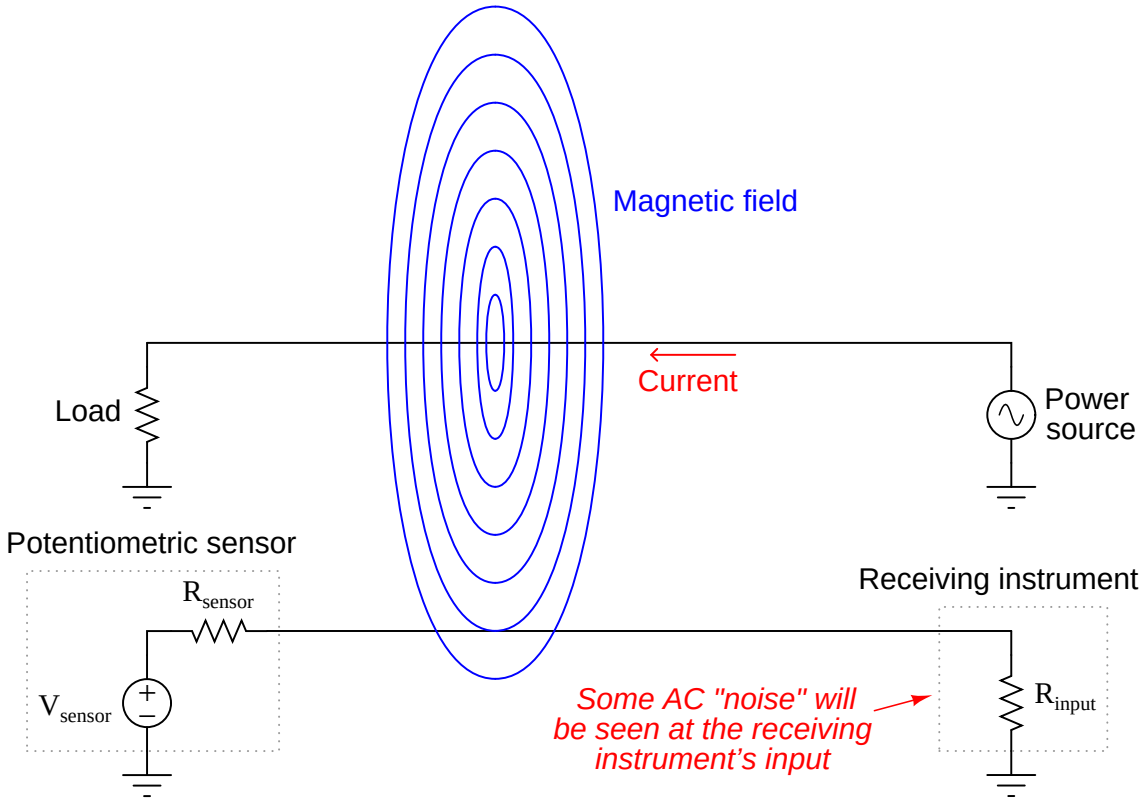
Two mechanisms of electrical "coupling" exist: *capacitive* and *inductive*. Capacitance is a property intrinsic to any pair of conductors separated by a dielectric (an insulating substance), whereby energy is stored in the electric field formed by voltage between the wires. The natural capacitance existing between mutually insulated wires forms a "bridge" for AC signals to cross between those wires, the strength of that "bridge" inversely proportional to the capacitive reactance ($X_C = \frac{1}{2\pi fC}$). Inductance is a property intrinsic to any conductor, whereby energy is stored in the magnetic field formed by current through the wire. Mutual inductance existing between parallel wires forms another "bridge" whereby an AC current through one wire is able to induce an AC voltage along the length of another wire.

Capacitive coupling between an AC power conductor and a DC sensor signal conductor is shown in the following diagram:



If the potentiometric (i.e. the measurement is based on voltage) sensor happens to be a thermocouple and the receiving instrument a temperature indicator, the result of this capacitive coupling will be a "noisy" temperature signal interpreted by the instrument. This noise will be proportional to both the *voltage* and the frequency of the AC power.

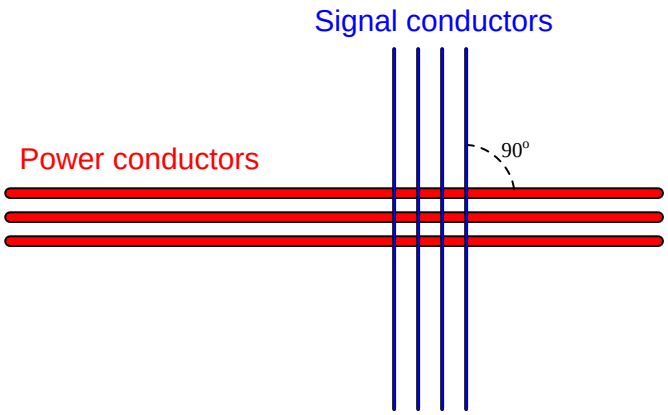
Inductive coupling between an AC power conductor and a DC sensor signal conductor is shown in the following diagram:



If the potentiometric (i.e. the measurement is based on voltage) sensor happens to be a thermocouple and the receiving instrument a temperature indicator, the result of this inductive coupling will be a "noisy" temperature signal interpreted by the instrument. This noise will be proportional to both the *current* and the frequency of the AC power.

A simple way to reduce signal coupling us to simply separate conductors carrying incompatible signals. This is why electrical power conductors and instrument signal cables are almost never found in the same conduit or in the same ductwork together. Separation decreases capacitance between the conductors (recall that $C = \frac{A\epsilon}{d}$ where d is the distance between the conductive surfaces). Separation also decreases the coupling coefficient between inductors, which in turn decreases mutual inductance (recall that $M = k\sqrt{L_1 L_2}$ where k is the coupling coefficient and M is the mutual inductance between two inductances L_1 and L_2). In control panel wiring, it is customary to route AC power wires in such a way that they do not lay parallel to low-level signal wires, so that both forms of coupling may be reduced.

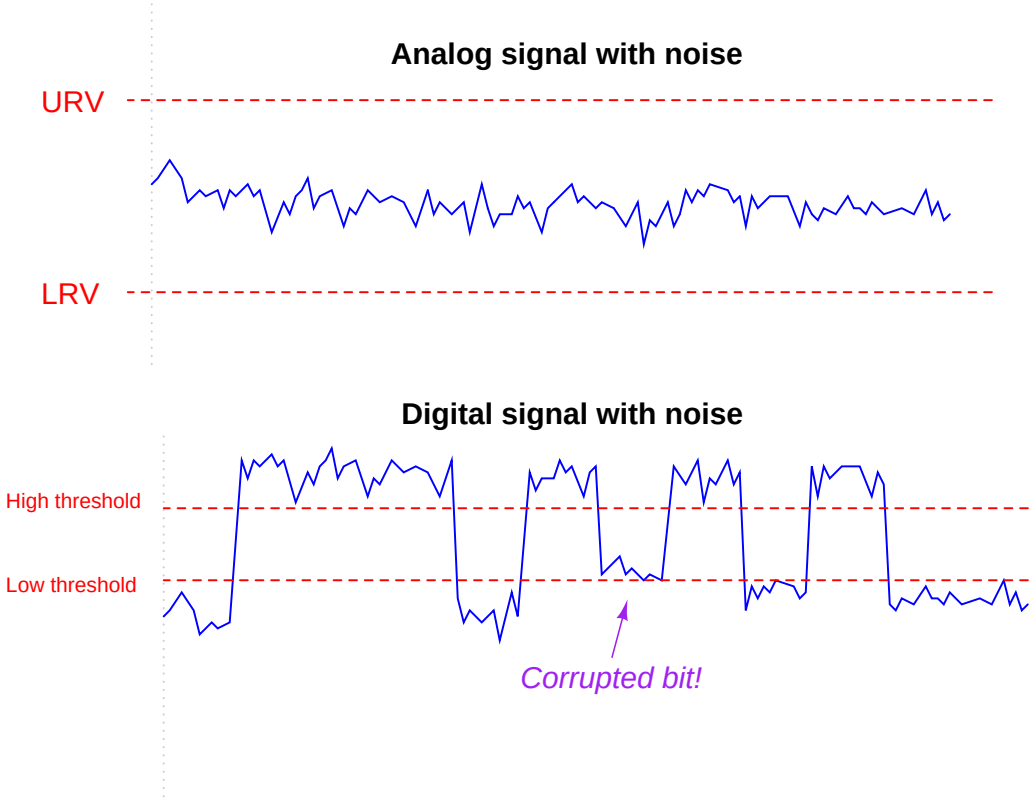
If conductors carrying incompatible signals *must* intersect in a panel, it is advisable to orient them so the crossing is perpendicular rather than parallel, like this:



Parallel conductor orientation reduces both inter-conductor capacitance *and* mutual inductance by two mechanisms. Capacitance between conductors is reduced by means of minimizing overlapping area (A) resulting from the perpendicular crossing. Mutual inductance is reduced by decreasing the coupling coefficient (k) to nearly zero since the magnetic field generated perpendicular to the current-carrying wire will be *parallel* and not perpendicular to the "receiving" wire. Since the vector for induced voltage is perpendicular to the magnetic field (i.e. parallel with the current vector in the "primary" wire) there will be no voltage induced along the length of the "receiving" wire.

The problem of power-to-signal line coupling is most severe when the signal in question is *analog* rather than *digital*. In analog signaling, even the smallest amount of coupled "noise" corrupts the signal. A digital signal, by comparison, will become corrupted only if the coupled noise is so severe that it pushes the signal level above or below a detection threshold is should not cross. This disparity is best described through illustration.

Two signals are shown here, coupled with equal amounts of noise voltage:



The peak-to-peak amplitude of the noise on the analog signal is almost 20% of the entire signal range (the distance between the lower- and upper-range values), representing a substantial degradation of signal integrity. Analog signals have infinite resolution, which means *any* change in signal amplitude has meaning. Any noise whatsoever is degrading to an analog signal because that noise (when interpreted by a receiving circuit) will be interpreted as changes in the quantity that signal is supposed to represent.

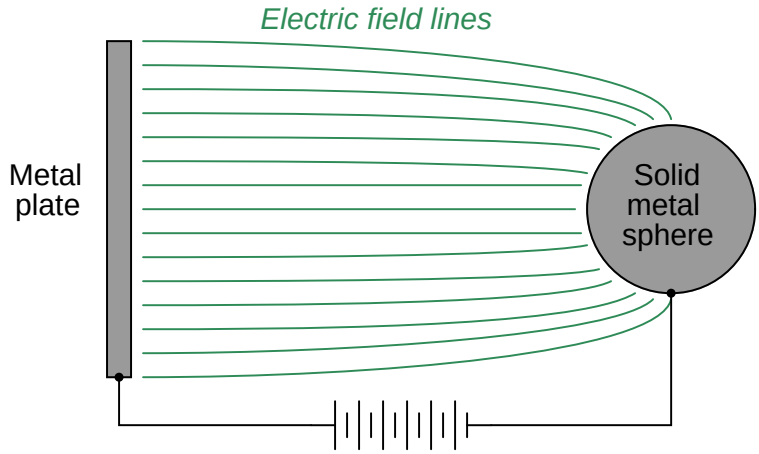
That same amount of noise imposed on a digital signal, however, causes no degradation of the signal except for one point in time where the signal attempts to reach a "low" state but fails to cross the threshold due to the noise. Other than that one incident represented in the pulse waveform, the rest of the signal is completely unaffected by the noise, because digital signals only have meaning above the "high" state threshold and below the "low" state threshold. Changes in signal voltage level caused by induced noise will not affect the meaning of digital data unless and until the amplitude of that noise becomes severe enough to prevent the signal's crossing through a threshold (when it should cross), or causes the signal to cross a threshold (when it should not).

From what we have seen here, digital signals are far more tolerant of induced noise than analog signals, all other factors being equal. If ever you find yourself in a position where you must pull a signal wire through a conduit filled with AC power conductors, and you happen to have the choice

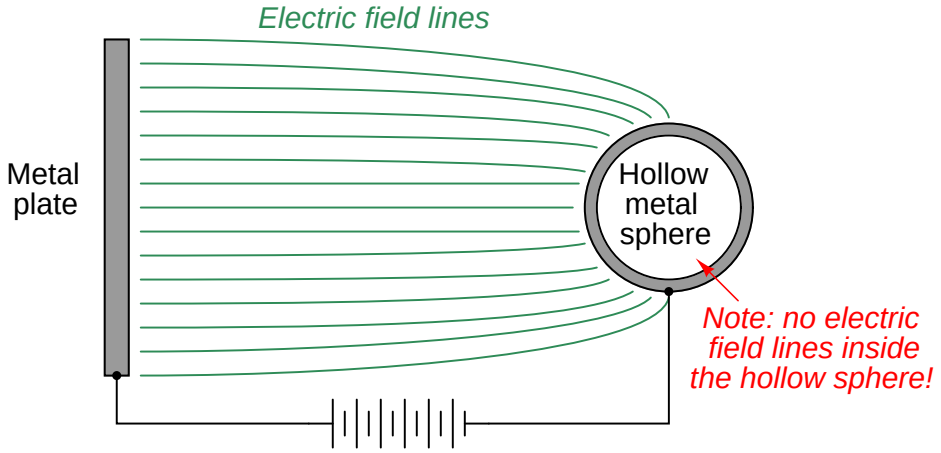
whether it will be an analog signal (e.g. 4-20 mA, 0-10 V) or a digital signal (e.g. EIA/TIA-485, Ethernet) you pull through that power conduit, your best option is to go with the digital signal.

8.3.4 Electric field (capacitive) de-coupling

The fundamental principle invoked in *shielding* signal conductor(s) from external electric fields is that an electric field cannot exist within a solid conductor. Electric fields exist due to imbalances of electric charge. If such an imbalance of charge ever were to exist within a conductor, charge carriers (typically electrons) in that conductor would quickly move to equalize the imbalance, thus eliminating the electric field. Thus, electric flux lines may be found only in the dielectric (insulating media) between conductors:



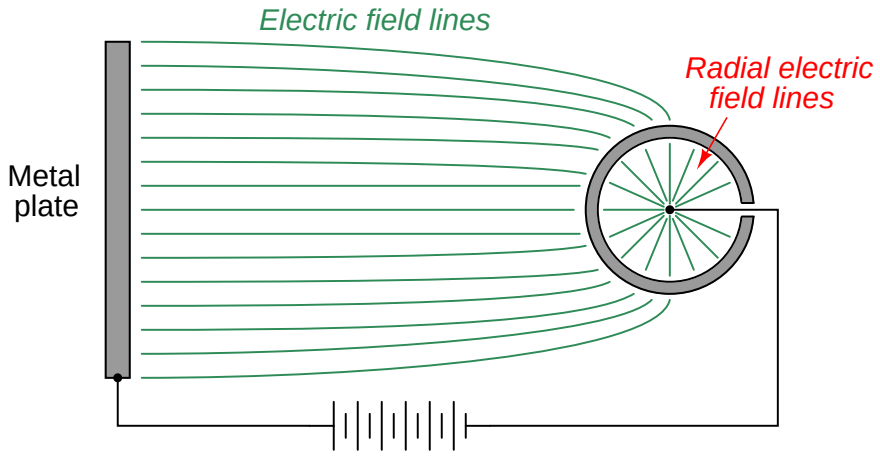
Not only does this mean that static electric fields cannot exist within a conductor, but it also means electric flux lines cannot exist within the confines of a hollow conductor:



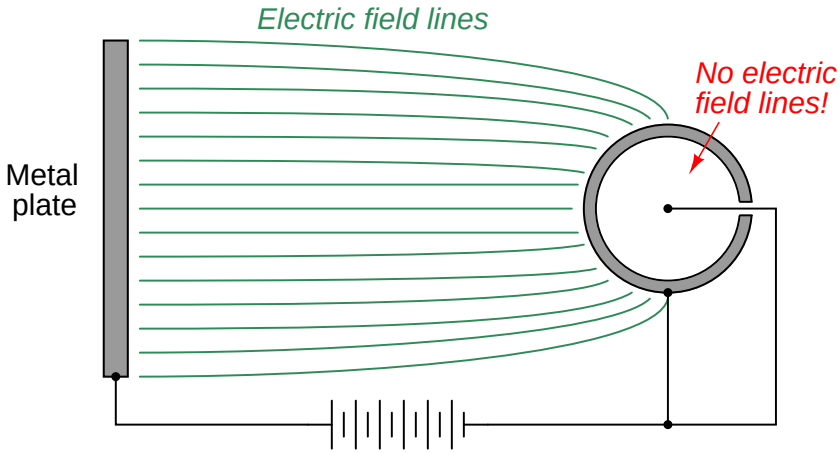
In order for an electric field to span the dielectric space within the hollow conductor, there would have to be an imbalance of electric charge from one side of the conductor to the other, which would be immediately equalized by charge motion within the hollow shell. The interior space of the hollow conductor, therefore, is free from external electric fields imposed by conductors outside the hollow

conductor. To state this differently, the interior of the hollow conductor is *shielded* from external electrostatic interference.

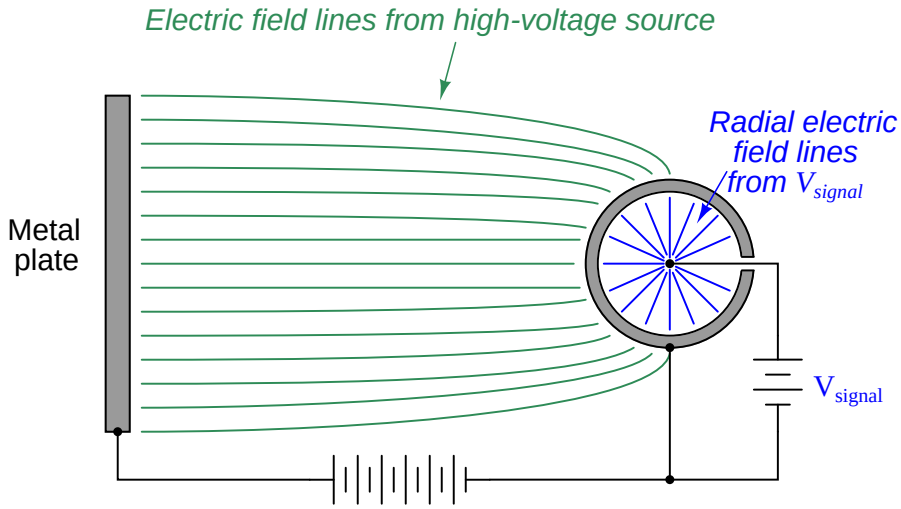
It is possible for an external electric field to penetrate a hollow conductor from the outside, but only if the conductive shell is left "floating" with respect to another conductor placed within the shell. For example:



However, if we make the hollow shell electrically common to the negative side of the high-voltage source, the flux lines inside the sphere vanish, since there is no potential difference between the internal conductor and the conductive shell:



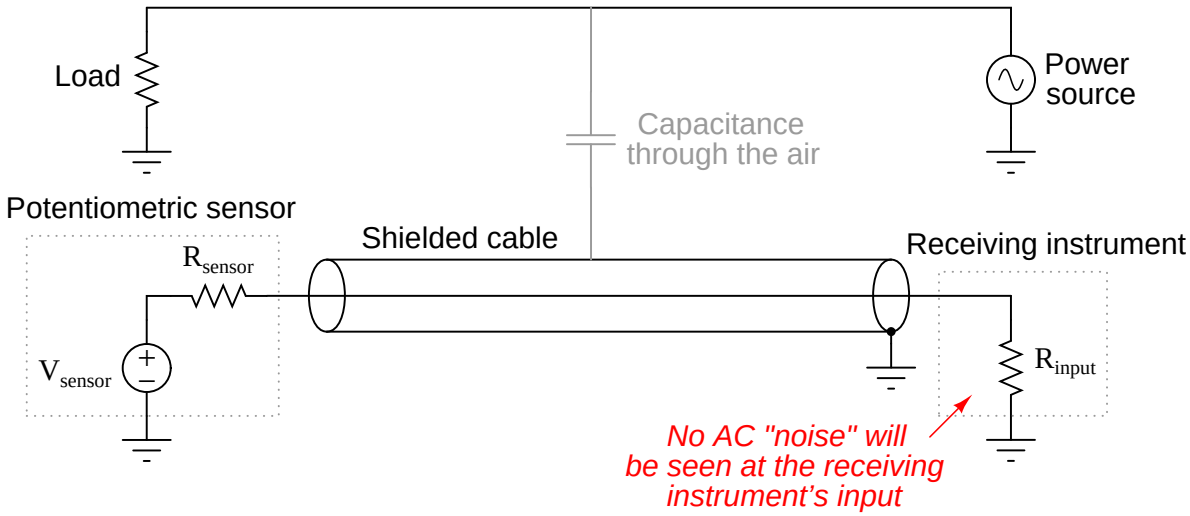
If the conductor within the hollow sphere is elevated to a potential different from that of the high-voltage source's negative terminal, electric flux lines will once again exist inside the sphere, but they will reflect this second potential and not the potential of the original high-voltage source. In other words, an electric field will exist inside the hollow sphere, but it will be completely isolated from the electric field outside the sphere. Once again, the conductor inside is *shielded* from external electrostatic interference:



If conductors located inside the hollow shell are thus shielded from external electric fields, it means there cannot exist any capacitance between external conductors and internal (shielded) conductors. If there is no capacitance between conductors, there will never be capacitive coupling of signals between those conductors, which is what we want for industrial signal cables to protect those signals from external interference[9].

[9] Incidentally, cable shielding likewise guards against strong electric fields *within* the cable from capacitively coupling with conductors outside the cable. This means we may elect to shield "noisy" power cables instead of (or in addition to) shielding low-level signal cables. Either way, good shielding will prevent capacitive coupling between conductors on either side of a shield.

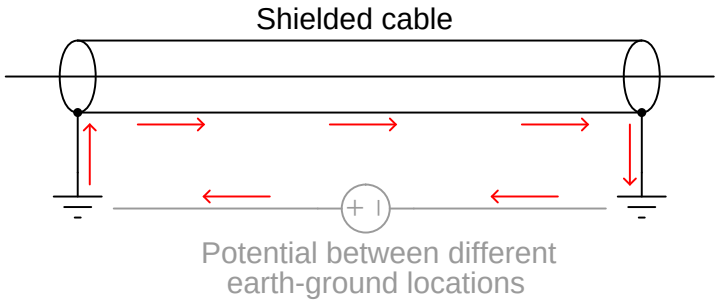
This is how *shielded cables* are manufactured: conductors within the cable are wrapped in a conductive metal foil or conductive metal braid, which may be connected to ground potential (the "common" point between external and internal voltage sources) to prevent capacitive coupling between those external voltage sources and the conductors within the cable:



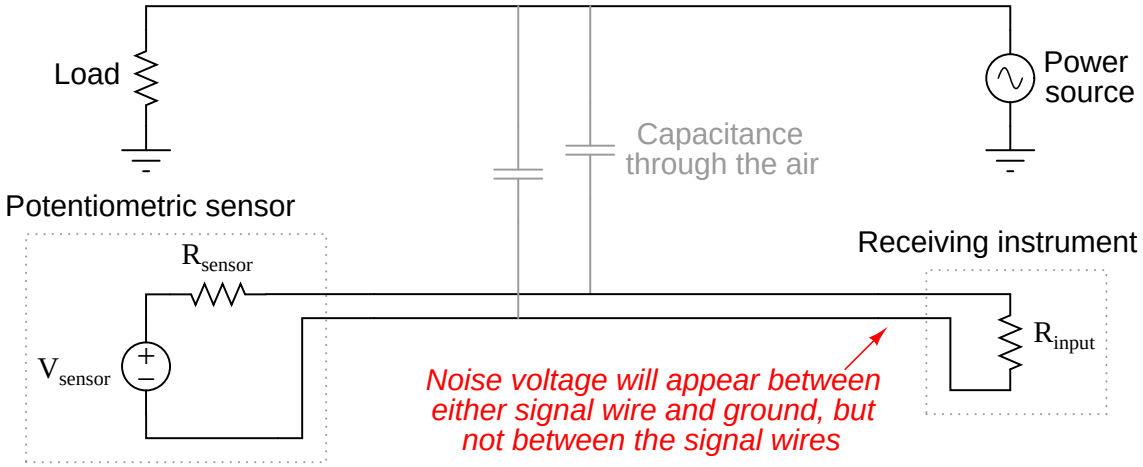
The following photograph shows a set of signal cables with braided shield conductors all connected to a common copper "ground bus." This particular application happens to be in the control panel of a 500 kV circuit breaker, located at a large electrical power substation where strong electric fields abound:

It is very important to ground *only one end* of a cable's shield, or else you will create the possibility for a *ground loop*: a path for current to flow through the cable's shield resulting from differences in Earth potential at the cable ends. Not only can ground loops induce noise in a cable's conductor(s), but in severe cases it can even overheat the cable and thus present a fire hazard!

A ground loop: something to definitely avoid!



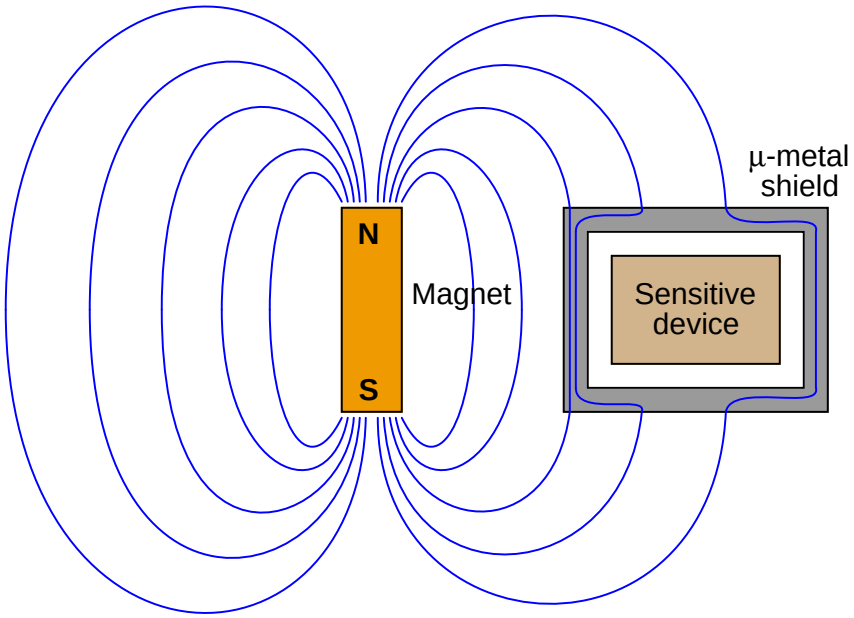
An alternative to shielding electric fields is to use *differential signaling* to help nullify the effects of capacitive coupling. The following schematic diagram illustrates how this works:



The lack of a ground connection in the DC signal circuit prevents capacitive coupling with the AC voltage from corrupting the measurement signal "seen" by the instrument. Noise voltage *will* still appear between either signal wire and ground, but not *between* the two signal wires, which is all the instrument is able to measure.

8.3.5 Magnetic field (inductive) de-coupling

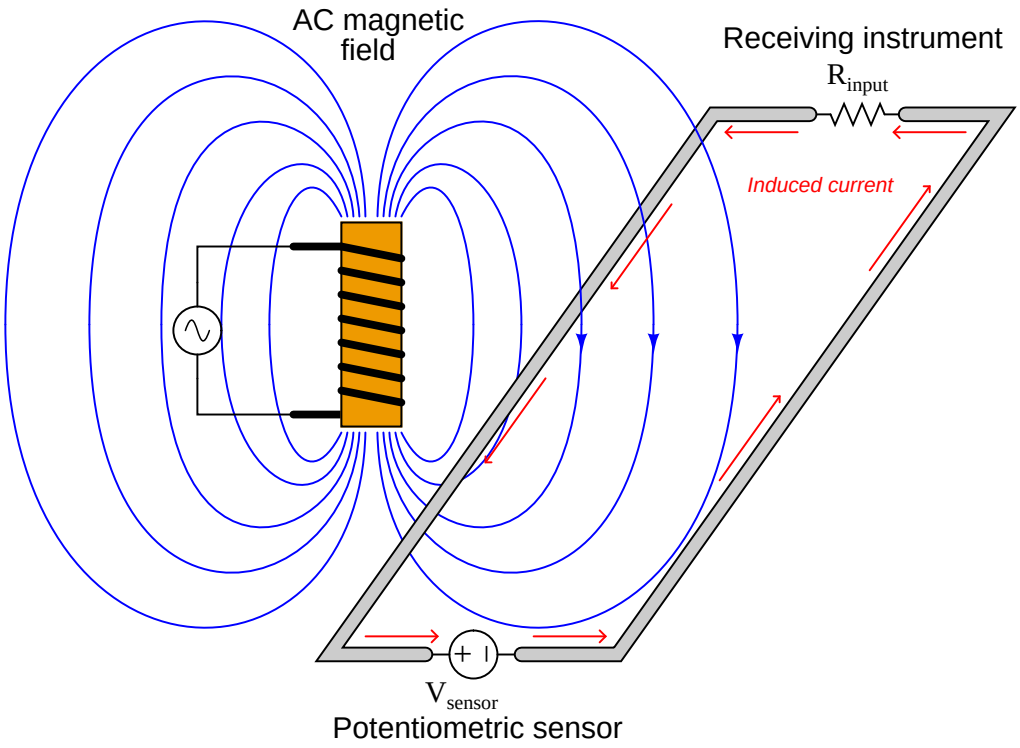
Magnetic fields, unlike electric fields, are exceedingly difficult to completely shield. Magnetic flux lines do not terminate, but rather *loop*. Thus, one cannot "stop" a magnetic field, only re-direct its path. A common method for magnetically shielding a sensitive instrument is to encapsulate it in an enclosure made of some material having an extremely high magnetic permeability (μ): a shell offering much easier passage of magnetic flux lines than air. A material often used for this application is *mu-metal*, or μ-*metal*, so named for its excellent magnetic permeability:



This sort of shielding is impractical for protecting signal cables from inductive coupling, as mu-metal is rather expensive and must be layered relatively thick in order to provide a sufficiently low-reluctance path to re-direct a majority of any imposed magnetic flux lines.

The most practical method of granting magnetic field immunity to a signal cable follows the differential signaling method discussed in the electric field de-coupling section, with a twist (literally). If we *twist* a pair of wires rather than allow them to lie along parallel straight lines, the effects of electromagnetic induction are vastly reduced.

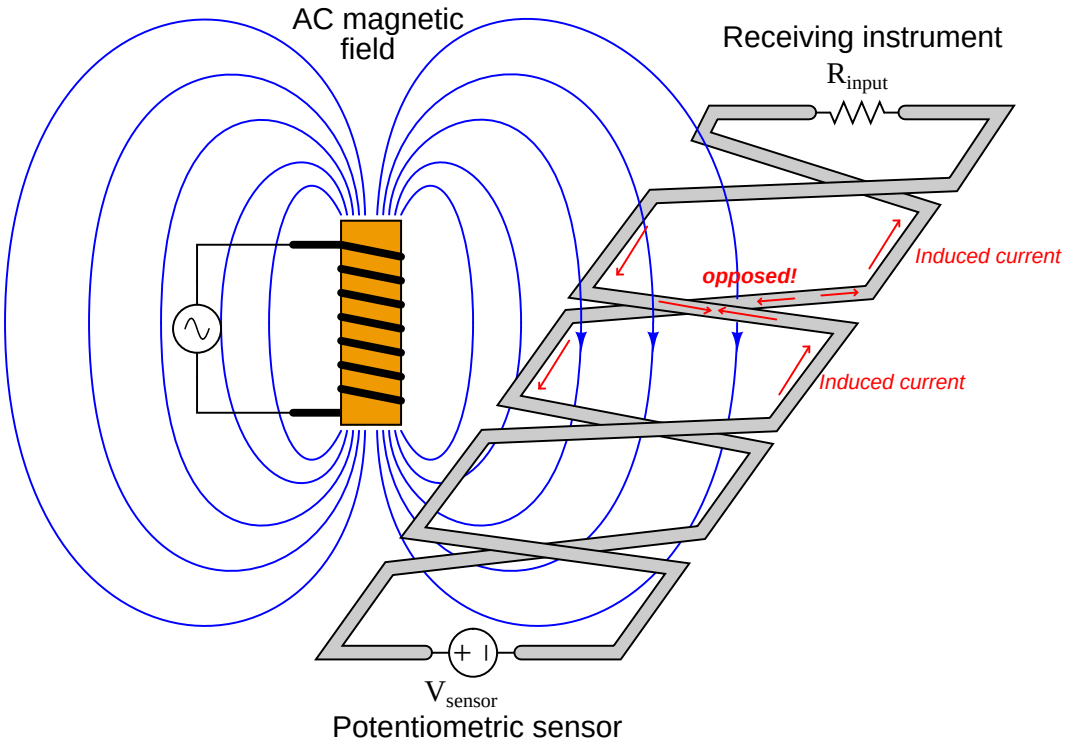
The reason this works is best illustrated by drawing a differential signal circuit with two thick wires, drawn first with no twist at all. Suppose the magnetic field shown here (with three flux lines entering the wire loop) happens to be *increasing* in strength at the moment in time captured by the illustration:



According to Lenz's Law, a current will be induced in the wire loop in such a polarity as to oppose the increase in external field strength. In other words, the induced current tries to "fight" the imposed field to maintain zero net change. According to the right-hand rule of electromagnetism (tracing current in conventional flow notation), the induced current must travel in a counter-clockwise direction as viewed from above the wire loop. This induced current works against the DC current produced by the sensor, detracting from the signal received at the instrument.

When the external magnetic field strength diminishes, then builds in the opposite direction, the induced current will reverse. Thus, as the AC magnetic field oscillates, the induced current will also oscillate in the circuit, causing AC "noise" voltage to appear at the measuring instrument. This is precisely the effect we wish to mitigate.

If we twist the wires so as to create a series of loops instead of one large loop, we will see that the inductive effects of the external magnetic field tend to cancel:



Not all the lines of flux go through the same loop. Each loop represents a reversal of direction for current in the instrument signal circuit, and so the direction of magnetically-induced current in one loop directly opposes the direction of magnetically-induced current in the next. So long as the loops are sufficient in number and spaced close together, the net effect will be complete and total opposition between all induced currents, with the result of no net induced current and therefore no AC "noise" voltage appearing at the instrument.

In order to enjoy the benefits of magnetic *and* electric field rejection, instrument cables are generally manufactured as *twisted, shielded pairs*. The twists guard against magnetic (inductive) interference, while the grounded shield guards against electric (capacitive) interference.

8.3.6 High-frequency signal cables

Electronic signals used in traditional instrumentation circuits are either DC or low-frequency AC in nature. Measurement and control values are represented in *analog* form by these signals, usually by the magnitude of the electronic signal (how many volts, how many milliamps, etc.). Modern electronic instruments, however, often communicate process and control data in *digital* rather than analog form. This digital data takes the form of high-frequency voltage and/or current pulses along the instrument conductors. The most capable *fieldbus* instruments do away with analog signaling entirely, communicating all data in digital form at relatively high speeds.

If the time period of a voltage or current pulse is less than the time required for the signal to travel down the length of the cable (at nearly the speed of light!), very interesting effects may occur. When a pulse propagates down a two-wire cable and reaches the end of that cable, the energy contained by that pulse must be absorbed by the receiving circuit or else be *reflected* back down the cable. To be honest, this happens in all circuits no matter how long or brief the pulses may be, but the effects of a "reflected" pulse only become apparent when the pulse time is short compared to the signal propagation time. In such short-pulse applications, it is customary to refer to the cable as a *transmission line*, and to regard it as a circuit component with its own characteristics (namely, a continuous impedance as "seen" by the traveling pulse). For more detail on this subject, refer to section 5.5 beginning on page 255.

This problem has a familiar analogy: an "echo" in a room. If you step into a large room with hard wall, floor, and ceiling surfaces, you will immediately notice echoes resulting from any sound you make. Holding a conversation in such a room can be quite difficult, as the echoed sounds superimpose upon the most recently-spoken sounds, making it difficult to discern what is being said. The larger the room, the longer the echo delay, and the greater the conversational confusion.

Echoes happen in small rooms, too, but they are generally too short to be of any concern. If the reflected sound(s) return quickly enough after being spoken, the time delay between the spoken (incident) sound and the echo (reflected) sound will be too short to notice, and conversation will proceed unhindered.

We may address the "echo" problem in two entirely different ways. One way is to eliminate the echoes entirely by adding sound-deadening coverings (carpet, acoustic ceiling tiles) and/or objects (sofas, chairs, pillows) to the room. Another way to address the problem of echoes interrupting a conversation is to *slow down the rate of speech*. If the words are spoken slowly enough, the time delay of the echoes will be relatively short compared to the period of each spoken sound, and conversation may proceed without interference (albeit at a reduced speed).

Both the problem of and the solutions for reflected signals in electrical cables follow the same patterns as the problem of and solutions for sonic echoes in a hard-surfaced room. If an electronic circuit receiving pulses sent along a cable receives both the incident pulse and an echo (reflected pulse) with a significant time delay separating those two pulses, the digital "conversation" will be impeded in the same manner that a verbal conversation between two or more people is impeded by echoes in a room. We may address this problem either by eliminating the reflected pulses entirely (by ensuring all the pulse energy is absorbed when it reaches the cable's end) or by slowing down the data transfer rate (i.e. longer pulses, lower frequencies) so that the reflected and incident pulse signals virtually overlap one another at the receiver.

High-speed "fieldbus" instrument networks apply the former solution (eliminate reflections) while the legacy HART instrument signal standard apply the latter (slow data rate). Reflections are eliminated in high-speed data networks by ensuring the two furthest cable ends are both

"terminated" by a resistance value of the proper size (matching the characteristic impedance of the cable). The designers of the HART analog-digital hybrid standard chose to use slow data rates instead, so their instruments would function adequately on legacy signal cables where the characteristic impedance is not standardized.

The potential for reflected pulses in high-speed fieldbus cabling is a cause for concern among instrument technicians, because it represents a new phenomenon capable of creating faults in an instrument system. No longer is it sufficient to have tight connections, clean wire ends, good insulation, and proper shielding for a signal cable to faithfully convey a 4-20 mA DC instrument signal from one device to another. Now the technician must ensure proper termination and the absence of any discontinuities[10] (sharp bends or crimps) along the cable's entire length, in addition to all the traditional criteria, in order to faithfully convey a digital fieldbus signal from one device to another.

Signal reflection problems may be investigated using a diagnostic instrument known as a *time-domain reflectometer*, or *TDR*. These devices are a combination of pulse generator and digital-storage oscilloscope, generating brief electrical pulses and analyzing the returned (echoed) signals at one end of a cable. If a TDR is used to record the pulse "signature" of a newly-installed cable, that data may be compared to future TDR measurements on the same cable to detect cable degradation or wiring changes.

[10] The characteristic, or "surge," impedance of a cable is a function of its conductor geometry (wire diameter and spacing) and dielectric value of the insulation between the conductors. Any time a signal reaches an abrupt change in impedance, some (or all) of its energy is reflected in the reverse direction. This is why reflections happen at the unterminated end of a cable: an "open" is an infinite impedance, which is a huge shift from the finite impedance "seen" by the signal as it travels along the cable. This also means any sudden change in cable geometry such as a crimp, nick, twist, or sharp bend is capable of reflecting part of the signal. Thus, high-speed digital data cables must be installed more carefully than low-frequency or DC analog signal cables.

References

Austin, George T., *Shreve's Chemical Process Industries*, McGraw-Hill Book Company, New York, NY, 1984.

"CPI^{TM} Tube Fittings", catalog 4230, Parker Hannifin Corporation, Cleveland, OH, 2000.

Croft, Terrell and Summers, Wilford I., *American Electrician's Handbook*, Eleventh Edition, McGraw-Hill Book Company, New York, NY, 1987.

"Fitting Installation Manual", Hoke Incorporated, Spartanburg, SC, 1999.

"Gaugeable Tube Fittings and Adapter Fittings", document MS-01-140, revision 7, Swagelok Company, MI, 2004.

Graves, W.V., *The Pipe Fitters Blue Book*, W.V. Graves Publisher, Webster, TX, 1973.

"Industrial Pipe Fittings and Adapters", catalog 4300, Parker Hannifin Corporation, Columbus, OH, 2000.

Morrison, Ralph, *Grounding and Shielding Techniques in Instrumentation*, John Wiley and Sons, Inc., NY, 1967.

"Pipe Fittings", document MS-01-147, revision 3, Swagelok Company, MI, 2002.

"Thread and End Connection Identification Guide", document MS-13-77, revision 3, Swagelok Company, 2005.

Chapter 9

Discrete process measurement

The word "discrete" means *individual* or *distinct*. In engineering, a "discrete" variable or measurement refers to a true-or-false condition. Thus, a discrete sensor is one that is only able to indicate whether the measured variable is above or below a specified setpoint.

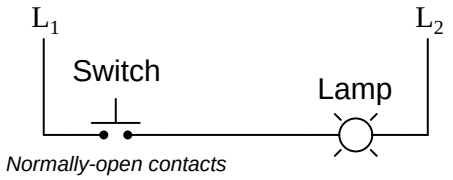
Discrete sensors typically take the form of *switches*, built to "trip" when the measured quantity either exceeds or falls below a specified value. These devices are less sophisticated than so-called *continuous* sensors capable of reporting an analog value, but they are quite useful in industry.

Many different types of discrete sensors exist, detecting variables such as position, fluid pressure, material level, temperature, and fluid flow rate. The output of a discrete sensor is typically electrical in nature, whether it be an active voltage signal or just resistive continuity between two terminals on the device.

9.1 "Normal" status of a switch

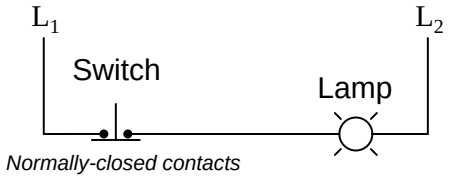
Perhaps the most confusing aspect of discrete sensors is the definition of a sensor's *normal* status. Electrical switch contacts are typically classified as either *normally-open* or *normally-closed*, referring to the open or closed status of the contacts under "normal" conditions. But what exactly defines "normal" for a switch? The answer is not complex, but it is often misunderstood.

The "normal" status for a switch is the status its electrical contacts are in *under a condition of minimum physical stimulus*. For a momentary-contact pushbutton switch, this would be the status of the switch contact when it is *not* being pressed. The "normal" status of any switch is the way it is drawn in an electrical schematic. For instance, the following diagram shows a normally-open pushbutton switch controlling a lamp on a 120 volt AC circuit (the "hot" and "neutral" poles of the AC power source labeled L1 and L2, respectively):



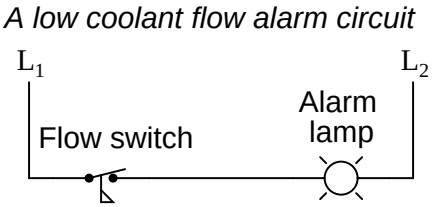
We can tell this switch is a normally-open (NO) switch because it is drawn in an open position. The lamp will energize only if someone presses the switch, holding its normally-open contacts in the "closed" position. Normally-open switch contacts are sometimes referred to in the electrical industry as *form-A* contacts.

If we had used a normally-closed pushbutton switch instead, the behavior would be exactly opposite. The lamp would energize if the switch was left alone, but it would turn off if anyone pressed the switch. Normally-closed switch contacts are sometimes referred to in the electrical industry as *form-B* contacts. :



This seems rather simple, don't you think? What could possibly be confusing about the "normal" status of a switch? The confusion becomes evident, though, when you consider the case of a different kind of discrete sensor such as a flow switch.

A flow switch is built to detect fluid flow through a pipe. In a schematic diagram, the switch symbol appears to be a toggle switch with a "flag" hanging below. The schematic diagram, of course, only shows the circuitry and not the pipe where the switch is physically mounted:



This particular flow switch is used to trigger an alarm light if coolant flow through the pipe ever falls to a dangerously low level, and the contacts are *normally-closed* as evidenced by the closed status in the diagram. Here is where things get confusing: even though this switch is designated as "normally-closed," it will spend most of its lifetime being held in the open status by the presence of adequate coolant flow through the pipe. Only when the flow through the pipe slows down enough will this switch return to its "normal" status (remember, the condition of *minimum stimulus?*) and conduct electrical power to the lamp. In other words, the "normal" status of this switch (closed) is actually an *abnormal* status for the process it is sensing (low flow)!

Students often wonder why process switch contacts are labeled according to this convention of "minimum stimulus" instead of according to the typical status of the process in which the switch is used. The answer to this question is that the manufacturer of the sensor has no idea whatsoever as to your intended use. The manufacturer of the switch does not know and does not care whether you intend to use their flow switch as a low-flow alarm or as a high-flow alarm. In other words, the manufacturer cannot predict what the typical status of *your* process will be, and so the definition of "normal" status for the switch must be founded on some common criterion unrelated to your particular application. That common criterion is the status of minimum stimulus: when the sensor is exposed to the *least* amount of stimulation from the process it senses.

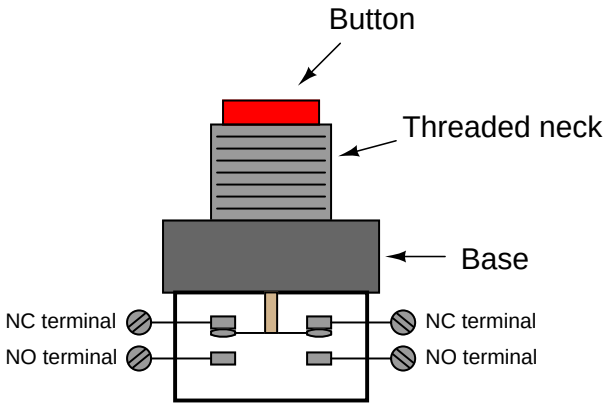
Here is a listing of "normal" definitions for various discrete sensor types:

- **Hand switch**: no one pressing the switch
- **Limit switch**: target not contacting the switch
- **Proximity switch**: target far away
- **Pressure switch**: low pressure (or even a vacuum)
- **Level switch**: low level (empty)
- **Temperature switch**: low temperature (cold)
- **Flow switch**: low flow rate (fluid stopped)

These are the conditions represented by the switch statuses shown in a schematic diagram. These may very well *not* be the statuses of the switches when they are exposed to *typical* operating conditions in the process.

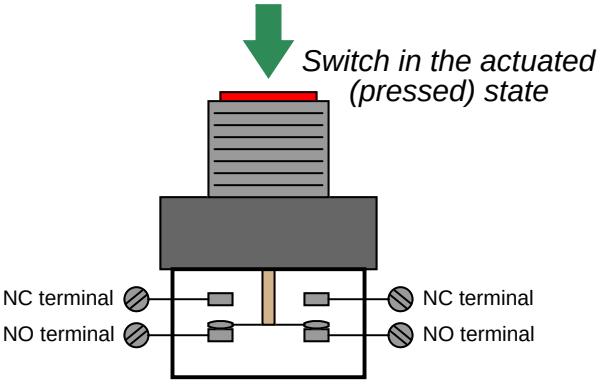
9.2 Hand switches

A *hand switch* is exactly what the name implies: an electrical switch actuated by a person's hand motion. These may take the form of toggle, pushbutton, rotary, pull-chain, etc. A common form of industrial pushbutton switch looks something like this:

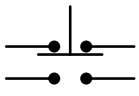


The threaded neck inserts through a hole cut into a metal or plastic panel, with a matching nut to hold it in place. Thus, the button faces the human operator(s) while the switch contacts reside on the other side of the panel.

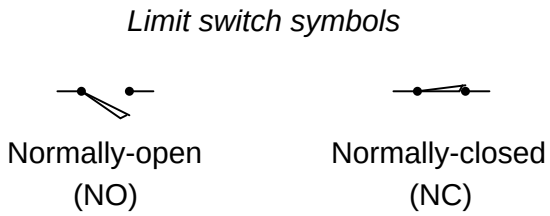
When pressed, the downward motion of the actuator breaks the electrical bridge between the two NC contacts, forming a new bridge between the NO contacts:



The schematic diagram symbol for this type of switch looks much like the real thing, with the normally-closed contact set on top and the normally-open contact set below:



9.3 Limit switches

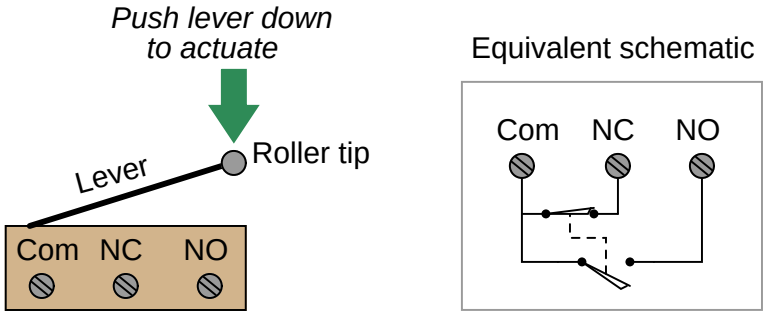


A *limit switch* detects the physical motion of an object by direct contact with that object. An example of a limit switch is the switch detecting the open position of an automobile door, automatically energizing the cabin light when the door opens.

Recall from page 368 that the "normal" status of a switch is the condition of *minimum stimulus*. A limit switch will be in its "normal" status when it is not in contact with anything (i.e. nothing touching the switch actuator mechanism).

Limit switches find many uses in industry, particular in robotic control and CNC (Computer Numerical Control) machine tool systems. In many motion-control systems, the moving elements have "home" positions where the computer assigns a position value of zero. For example, the axis controls on a CNC machine tool such as a lathe or mill all return to their "home" positions upon start-up, so the computer can know with confidence the starting locations of each piece. These home positions are detected by means of limit switches. The computer commands each servo motor to travel fully in one direction until a limit switch on each axis trips. The position counter for each axis resets to zero as soon as the respective limit switch detects that the home position has been reached.

A typical limit switch design uses a roller-tipped lever to make contact with the moving part. Screw terminals on the switch body provide connection points with the NC and NO contacts inside the switch. Most limit switches of this design share a "common" terminal between the NC and NO contacts like this:



This switch contact arrangement is sometimes referred to as a *form-C* contact set, since it incorporates both a form-A contact (normally-open) as well as a form-B contact (normally-closed).

A close-up view of several limit switches (used on a drum sequencer) shows the arrangement of connection terminals for form-C contacts. Each limit switch has its own "NO" (normally-open), "NC" (normally-closed), and "C" (common) screw terminal for wires to attach:

A limit switch assembly attached to the stem of a rotary valve – used to detect the fully-closed and fully-open positions of the valve – is shown in the following photograph:

9.4 Proximity switches

A *proximity switch* is one detecting the proximity (closeness) of some object. By definition, these switches are *non-contact sensors*, using magnetic, electric, or optical means to sense the proximity of objects.

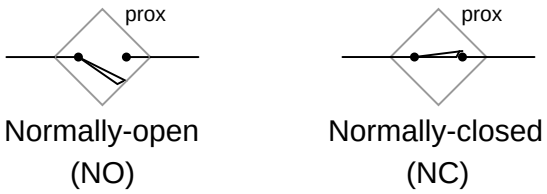
Recall from page 368 that the "normal" status of a switch is the condition of *minimum stimulus*. A proximity switch will be in its "normal" status when it is distant from any actuating object.

Being non-contact in nature, proximity switches are often used instead of direct-contact limit switches for the same purpose of detecting the position of a machine part, with the advantage of never wearing out over time due to repeated physical contact. However, the greater complexity (and cost) of a proximity switch over a mechanical limit switch relegates their use to applications where lack of physical contact yields tangible benefits.

Most proximity switches are *active* in design. That is, they incorporate a powered electronic circuit to sense the proximity of an object. *Inductive* proximity switches sense the presence of metallic objects through the use of a high-frequency magnetic field. *Capacitive* proximity switches sense the presence of non-metallic objects through the use of a high-frequency electric field. Optical switches detect the interruption of a light beam by an object.

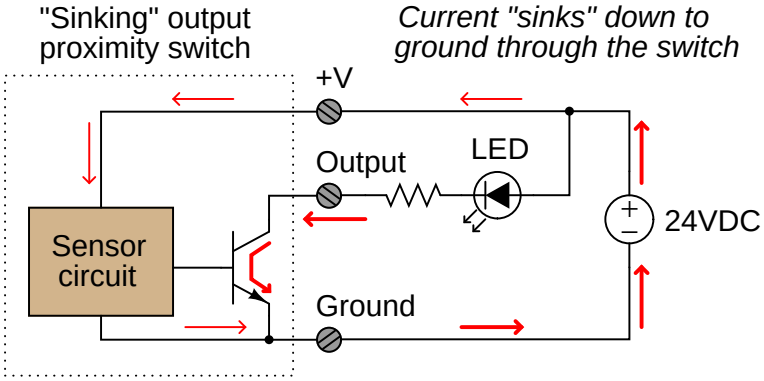
The schematic diagram symbol for a proximity switch with mechanical contacts is the same as for a mechanical limit switch, except the switch symbol is enclosed by a diamond shape, indicating a powered (active) device:

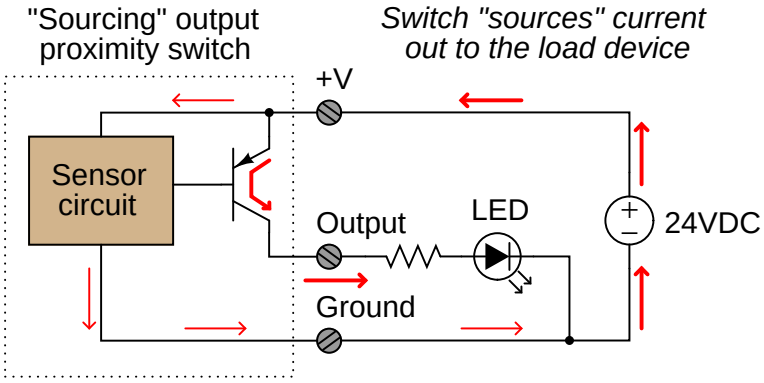
Proximity switch symbols



Many proximity switches, though, do not provide "dry contact" outputs. Instead, their output elements are transistors configured either to *source* current or *sink* current. The terms "sourcing" and "sinking" are best understood by visualizing electric current in the direction of *conventional flow* rather than *electron flow*.

The following schematic diagrams contrast the two modes of switch operation, using red arrows to show the direction of current (conventional flow notation). In both examples, the load being driven by each proximity switch is a light-emitting diode (LED):





These photographs show two different styles of electronic proximity switch:

The next photograph shows a proximity switch detecting the passing of teeth on a chain sprocket, generating a slow square-wave electrical signal as the sprocket rotates. Such a switch may be used as a rotational speed sensor (sprocket speed proportional to signal frequency) or as a broken chain sensor (when sensing the rotation of the driven sprocket instead of the drive sprocket):

9.5 Pressure switches

A *pressure switch* is one detecting the presence of fluid pressure. Pressure switches often use diaphragms or bellows as the pressure-sensing element, the motion of which actuates one or more switch contacts.

Recall from page 368 that the "normal" status of a switch is the condition of *minimum stimulus*. A pressure switch will be in its "normal" status when it senses minimum pressure (e.g. n applied pressure, or in some cases a vacuum condition)[1].

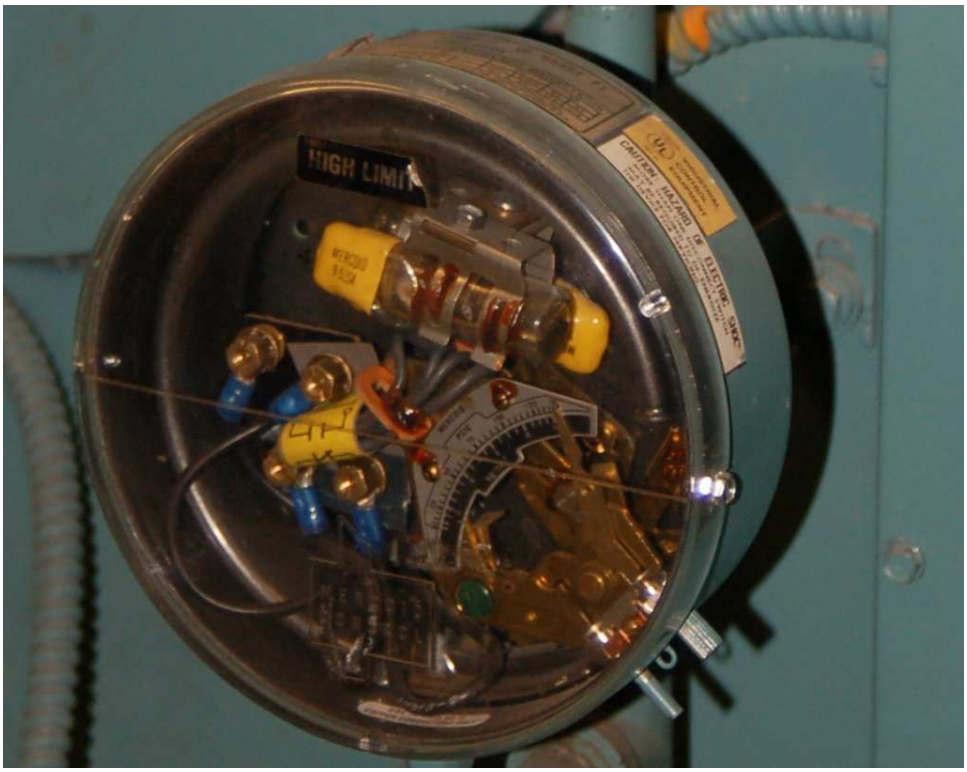
Pressure switch symbols

The following photograph shows two pressure switches sensing the same fluid pressure as an electronic pressure transmitter (the device on the far left):

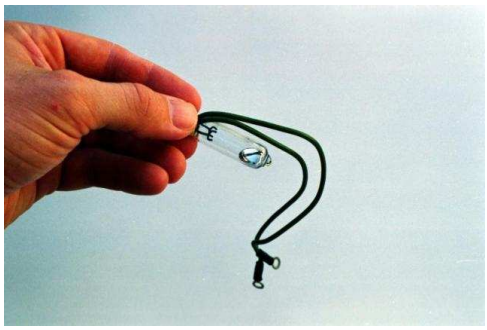
[1] If the trip setting of a pressure switch is below atmospheric pressure, then it will be "actuated" at atmospheric pressure and in its "normal" status only when the pressure falls below that trip point (i.e. a vacuum).

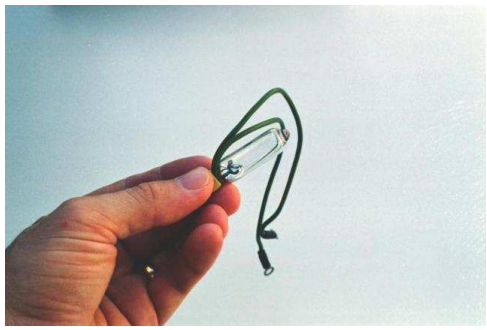
A legacy design of pressure switch uses a bourdon tube as the pressure-sensing element, and a glass bulb partially filled with mercury as the electrical switching element. When applied pressure causes the bourdon tube to flex sufficiently, the glass bulb tilts far enough to cause the mercury to fall against a pair of electrodes, thus completing an electrical circuit. A great many pressure switches of this design were sold under the brand name of "Mercoid," with a few appearing in this photograph of a steam boiler (the round-shaped units with glass covers allowing inspection of the bourdon tube and mercury tilt switch):

A close-up photograph of one of these pressure switches appears here. The bourdon tube is grey in color, and almost as wide in diameter as the circular switch housing. The mercury tilt switch bottles have yellow-colored plastic caps covering up their external electrical contacts:



The next set of photographs show a mercury tilt switch removed from the pressure switch mechanism, so you may see the switch in two different states (contact open on the left, and closed on the right):





Advantages of mercury tilt switches include immunity to switch contact degradation from harmful atmospheres (oil mist, dirt, dust, corrosion) as well as safety in explosive atmospheres (since a spark

contained within a hermetically sealed glass bulb cannot touch off an explosion in the surrounding atmosphere). Disadvantages include the possibility of intermittent electrical contact resulting from mechanical vibration, as well as sensitivity to mounting angle (i.e. you would *not* want to use this kind of switch aboard a moving vehicle!).

A pressure switch manufactured by the Danfoss corporation appears in the next photograph. This particular model of pressure switch has windows on the front cover allowing a technician to see the pressure limit setting inside:

This switch balances the force generated by a pressure-sensing element against a mechanical spring. Tension on the spring may be adjusted by a technician, which means the trip point of this switch is adjustable.

One of the settings on this switch is the *dead-band* or *differential* pressure setting, seen in the lower window. This setting determines the amount of pressure change required to re-set the switch to its normal state after it has tripped. For example, a high-pressure switch with a trip point of 67 PSI (changes state at 67 PSI, increasing) that re-sets back to its normal state at a pressure of 63 PSI decreasing has a "dead-band" or "differential" pressure setting of 4 PSI (67 PSI − 63 PSI = 4 PSI).

The "differential" pressure setting of a gauge pressure switch is not to be confused with a true *differential pressure* switch. In the next photograph, we see a pressure switch truly actuated by *differential* pressure (the difference in fluid pressure sensed between two ports):

The electrical switch element is located underneath the blue cover, while the diaphragm pressure element is located within the grey metal housing. The net force exerted on the diaphragm by the two fluid pressures varies in magnitude and direction with the magnitude of those pressures. If the two fluid pressures are precisely equal, the diaphragm experiences no net force (zero differential pressure).

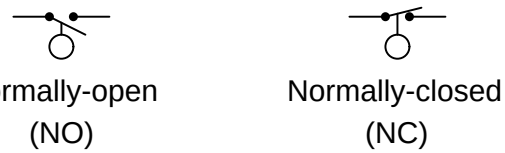
Like the Danfoss gauge pressure switch seen previously, this differential pressure switch has a "trip" or "limit" setting as well as a "dead-band" or "differential" setting. It is important to recognize and clearly distinguish the two meanings of *differential pressure* in the context of this device. It senses differences in pressure between two input ports ("differential pressure" – the difference between two different fluid pressure connections), but being a switch, it also exhibits some dead band in its action ("differential pressure" – a change in pressure required to re-set the switch's state).

9.6 Level switches

A *level switch* is one detecting the level of liquid or solid (granules or powder) in a vessel. Level switches often use floats as the level-sensing element, the motion of which actuates one or more switch contacts.

Recall from page 368 that the "normal" status of a switch is the condition of *minimum stimulus*. A level switch will be in its "normal" status when it senses minimum level (e.g. an empty vessel).

Level switch symbols



Two water level switches appear in this photograph of a steam boiler. The switches sense water level in the steam drum of the boiler. Both water level switches are manufactured by the Magnetrol corporation:

The switch mechanism is a mercury tilt bulb, tilted by a magnet's attraction to a steel rod lifted into position by a float. The float directly senses liquid level, which positions the steel rod either

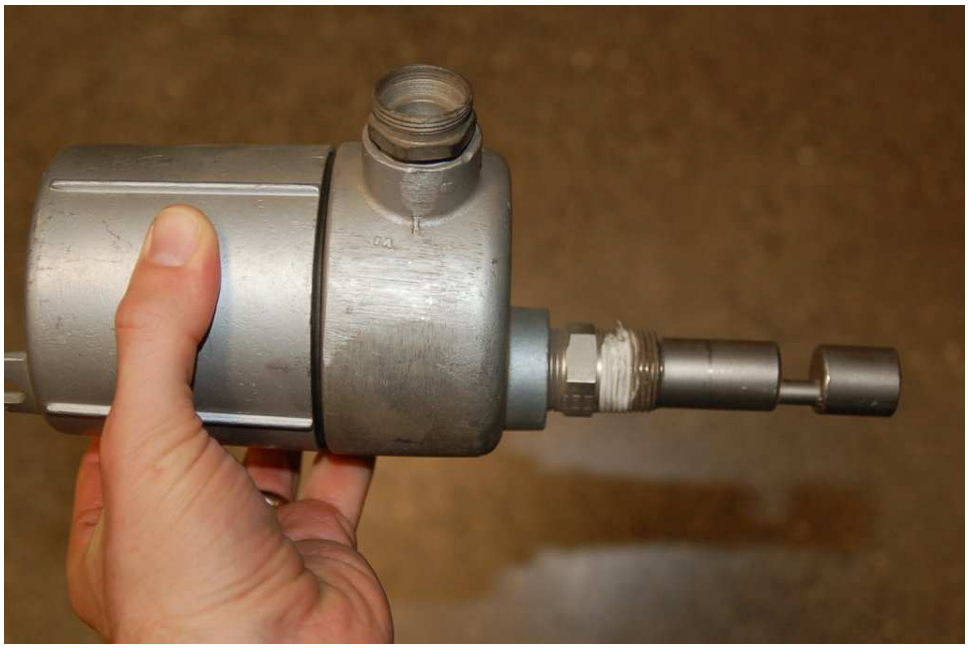
closer to or further away from the magnet. If the rod comes close enough to the magnet, the mercury bottle will tilt and change the switch's electrical status.

This level switch uses a metal *tuning fork* structure to detect the presence of a liquid or solid (powder or granules) in a vessel:

An electronic circuit continuously excites the tuning fork, causing it to mechanically vibrate. When the prongs of the fork contact anything with substantial mass, the resonant frequency of the structure dramatically decreases. The circuit detects this change and indicates the presence of material contacting the fork. The forks' vibrating motion tends to shake off any accumulated material, such that this style of level switch tends to be resistant to fouling.

A more primitive variation on the theme of a "tuning fork" level switch is the *rotating paddle* switch, used to detect the level of powder or granular solid material. This level switch uses an electric motor to slowly rotate a metal paddle inside the process vessel. If solid material rises to the level of the paddle, the material's bulk will place a mechanical load on the paddle. A torque-sensitive switch mechanically linked to the motor actuates when enough torsional effort is detected on the part of the motor. A great many level switches of this design sold in the United States under the trade-name *Bindicator* (so-called because they detected the level of solid material in storage *bins*).

Yet another style of electronic level switch uses ultrasonic sound waves to detect the presence of process material (either solid or liquid) at one point:



Sound waves pass back and forth within the gap of the probe, sent and received by piezoelectric transducers. The presence of any substance other than gas within that gap affects the received audio power, thus signaling to the electronic circuit within the bulkier portion of the device that process level has reached the detection point. The lack of moving parts makes this probe quite reliable, although it may become "fooled" by heavy fouling.

Another electronic liquid level switch technology is *capacitive*: sensing level by changes in electrical capacitance between the switch and the liquid. The following photograph shows a couple of capacitive switches sensing the presence of water in a plastic storage vessel:

9.7 Temperature switches

A *temperature switch* is one detecting the temperature of an object. Temperature switches often use bimetallic strips as the temperature-sensing element, the motion of which actuates one or more switch contacts. An alternative design uses a metal bulb filled with a fluid that expands with temperature, causing the switch mechanism to actuate based on the pressure this fluid exerts against a diaphragm or bellows. This latter temperature switch design is really a pressure switch, whose pressure is a direct function of process temperature by virtue of the physics of the entrapped fluid inside the sensing bulb.

Recall from page 368 that the "normal" status of a switch is the condition of *minimum stimulus*. A temperature switch will be in its "normal" status when it senses minimum temperature (i.e. cold, in some cases a condition colder than ambient)[2].

Temperature switch symbols

Normally-open (NO)

Normally-closed (NC)

[2] If the trip setting of a temperature switch is below ambient temperature, then it will be "actuated" at ambient temperature and in its "normal" status only when the temperature falls below that trip point (i.e. colder than ambient).

The following photograph shows a temperature-actuated switch manufactured by the Ashcroft corporation:

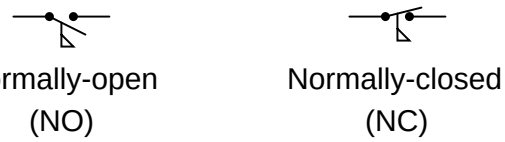
For discrete temperature-sensing applications demanding high accuracy and repeatability, *electronic* temperature switch circuits using thermocouples, RTDs, or thermistors may be used instead of a mechanical (bi-metallic or filled bulb) sensing element. The operation and configuration of discrete electronic temperature switches is very similar to that of continuous electronic temperature transmitters.

9.8 Flow switches

A *flow switch* is one detecting the flow of some fluid through a pipe. Flow switches often use "paddles" as the flow-sensing element, the motion of which actuates one or more switch contacts.

Recall from page 368 that the "normal" status of a switch is the condition of *minimum stimulus*. A flow switch will be in its "normal" status when it senses minimum flow (i.e. no fluid moving through the pipe).

Flow switch symbols



A simple paddle placed in the midst of a fluid stream generates a mechanical force which may be used to actuate a switch mechanism, as shown in the following photograph:

Chapter 10

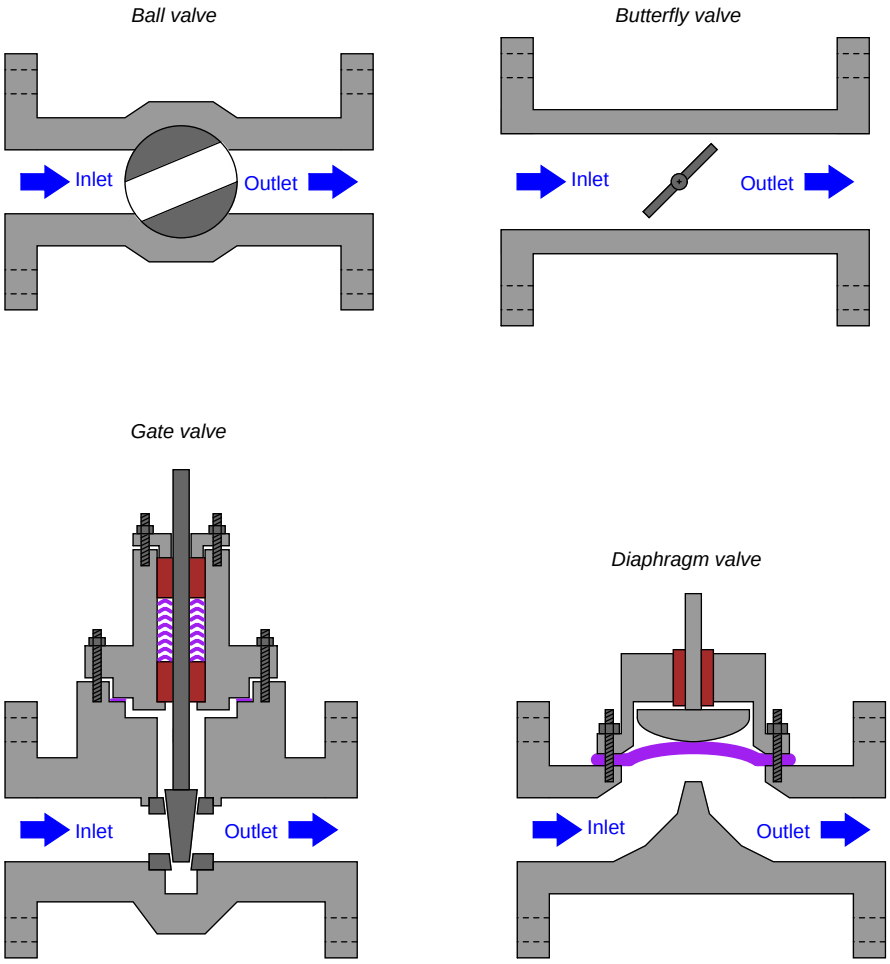
Discrete control elements

The word "discrete" means *individual* or *distinct*. In engineering, a "discrete" variable or measurement refers to a true-or-false condition. Thus, a discrete control element is one that has but a limited number of states (usually two: on and off). In the case of valves, this means a valve designed to operate either in "open" mode or "closed" mode, not in-between.

10.1 On/off valves

An on/off valve is the fluid equivalent of an electrical switch: a device that either allows unimpeded flow or acts to prevent flow altogether. These valves are often used for routing process fluid to different locations, starting and stopping batch processes, and engaging automated safety (shutdown) functions.

Valve styles commonly used for on/off service include ball, plug, butterfly (or disk), gate, and globe. Large on/off valves are generally of such a design that the full-open position provides a nearly unimpeded path for fluid to travel through. Ball, plug[1], and gate valves provide just this characteristic:



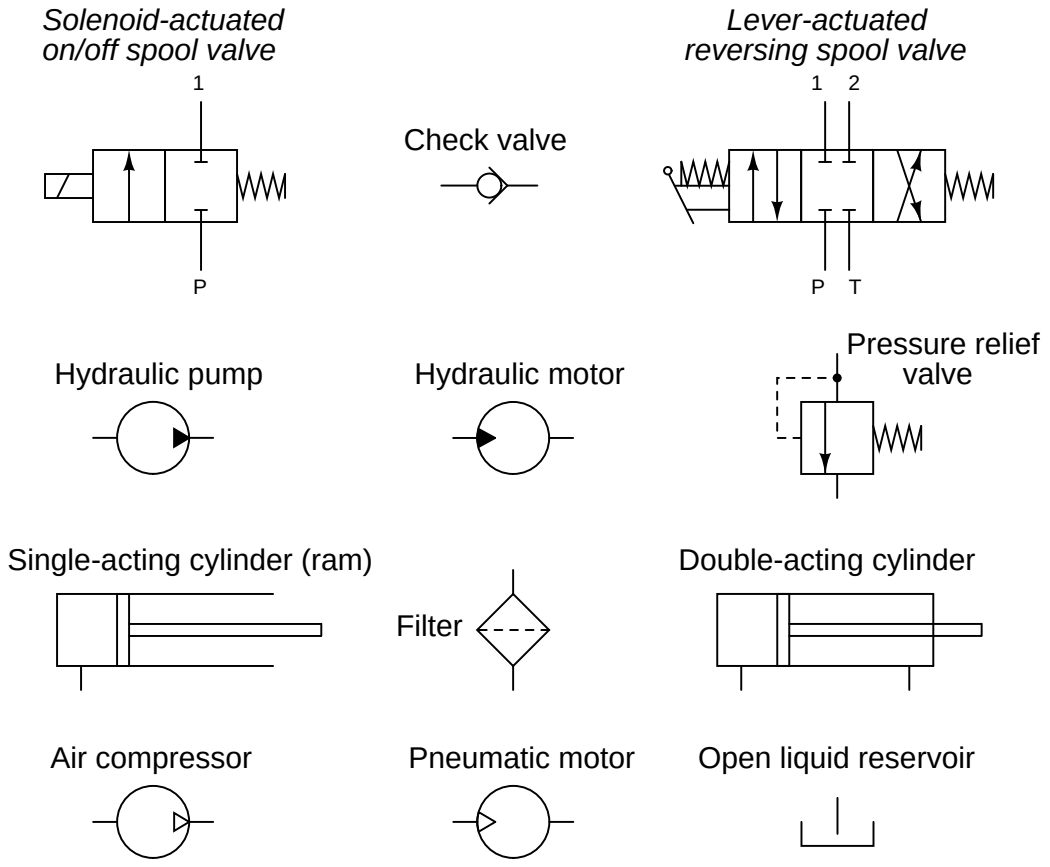
[1] A *plug* valve is very much like a ball valve, the difference being the shape of the rotating element. Rather than a spherical ball, the plug valve uses a truncated cone as the rotary element, a slot cut through the cone serving as the passageway for fluid. The conical shape of a plug valve's rotating element allows it to wedge tightly into the "closed" (shut) position for exceptional sealing.

10.2 Fluid power systems

Given the ability of pressurized fluids to transmit force over long distances, it is not surprising that many practical "fluid power systems" have been built using fluid as a mechanical power-conducting media. Fluid systems may be broadly grouped into *pneumatic* (gas, usually air) and *hydraulic* (liquid, usually oil).

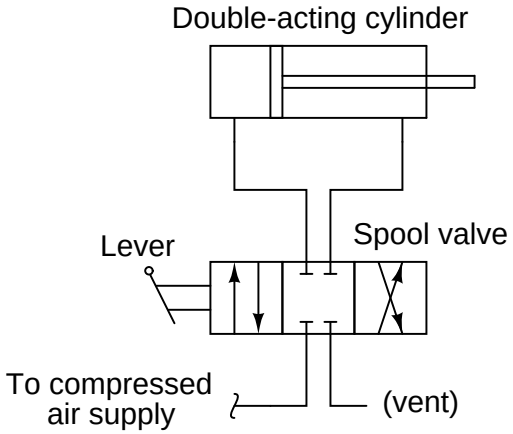
Although there is no particular reason why a fluid power system must be discrete and not continuous, the majority of fluid power systems operate in an on/off control mode rather than throttling, which is why this subject is covered in the "Discrete Control Elements" chapter.

As usual for technical specialties, fluid power has its own unique symbology for describing various components. The following diagram shows some common symbols used in fluid power system diagrams:



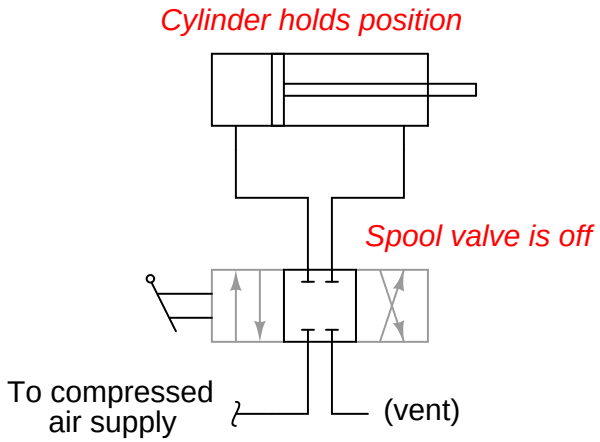
Many of these symbols are self-explanatory, especially the pumps, motors, and cylinders. What seems to cause the most confusion for people new to this symbology are the spool valve symbols. A "spool" valve is a special type of flow-directing valve used in pneumatic and hydraulic systems to

direct the pressurized fluid to different locations. The symbology for a spool valve is a set of boxes, each box containing arrows or other symbols showing the intended direction(s) for the fluid's travel. Take for instance this pneumatic reversing cylinder control system:

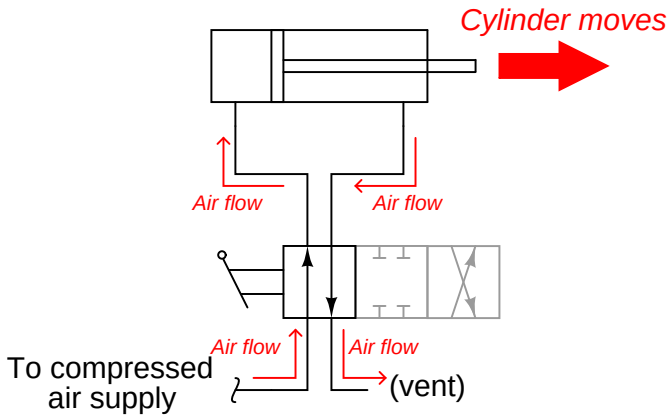


The proper way to interpret a spool valve symbol is to see only one "box" active at any given time. As the actuator (in this case, a hand-actuated lever) is moved one way or the other, the boxes "shift" laterally to redirect the flow of fluid from source to load.

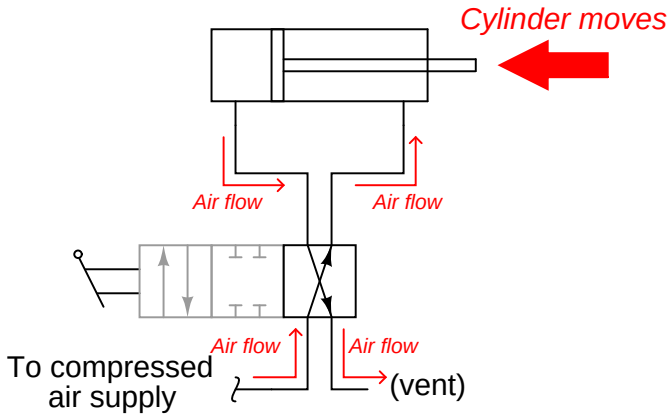
For example, when the spool valve in this reversing control system is in its center position, the outer boxes in the symbol are inactive. This is represented in the following diagram by showing the outer boxes in the color grey. In this position, the spool valve neither admits compressed air to the cylinder nor vents any air from the cylinder. As a result, the cylinder holds its position:



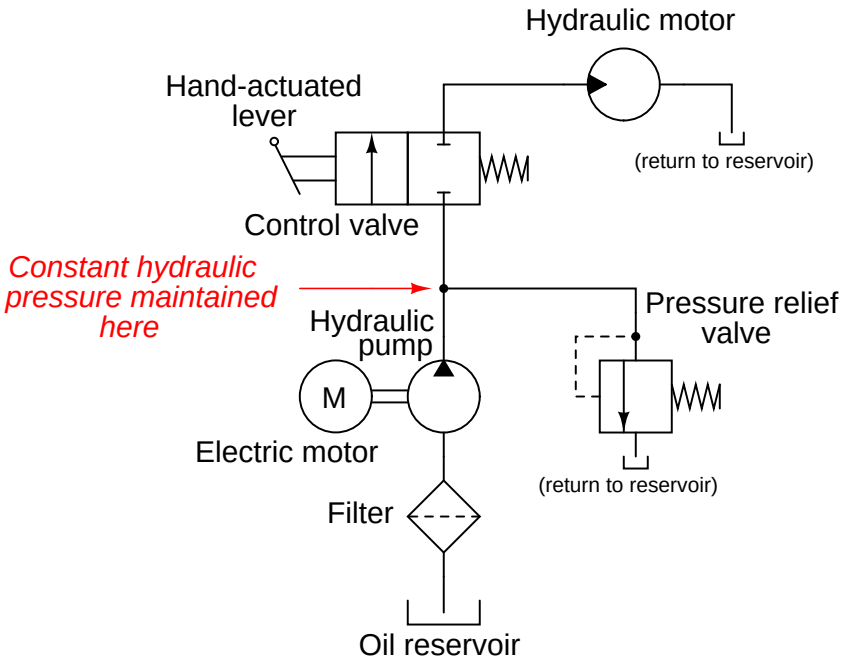
If the spool valve is actuated in one direction, the spool piece inside the valve assembly shifts, directing compressed air to one side of the cylinder while venting air from the other side. This is shown in the following diagram by shifting the boxes to one side, lining up the "active" box with the cylinder and air supply/vent connections:



If the spool valve is actuated in the other direction, the spool piece inside the valve assembly shifts again, switching the directions of air flow to and from the cylinder. Compressed air still flows from the supply to the vent, but the direction within the cylinder is reversed. This causes the cylinder to reverse its mechanical travel:



Hydraulic systems require more components, including filters and pressure regulators, to ensure proper operation. Shown here is a simple uni-directional hydraulic motor control system:

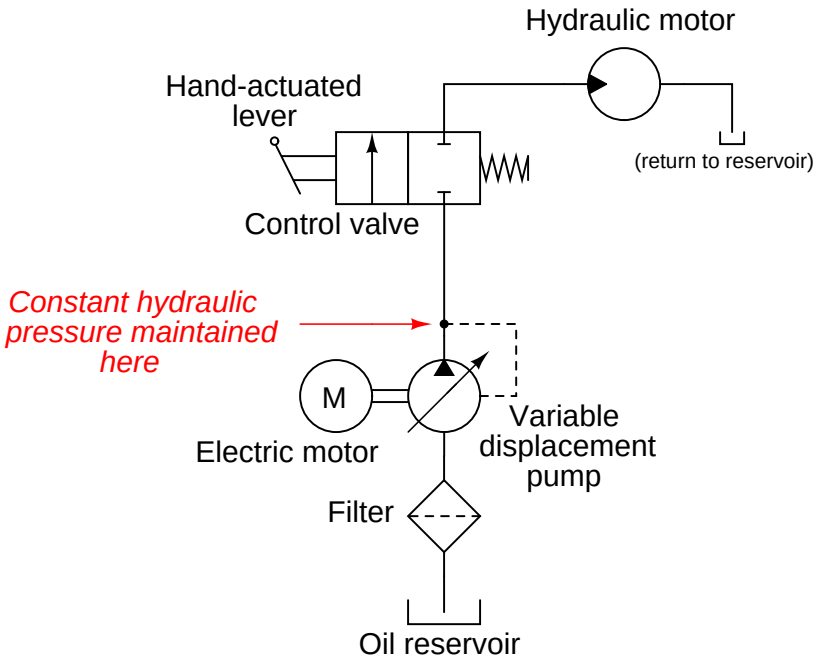


Note the placement of the pressure relief valve: it is a *shunt* regulator, bleeding excess pressure from the discharge of the hydraulic pump back to the reservoir[2]. A "shunt" regulator is necessary because hydraulic pumps are *positive displacement*, meaning they discharge a fixed volume of fluid with every revolution of the shaft. If the discharge of a positive-displacement pump is blocked (as it would be if the spool valve were placed in its default "off" position, with no shunt regulator to bleed pressure back to the reservoir), it will mechanically "lock" and refuse to turn. This would overload the electric motor coupled to the pump, if not for the pressure regulating valve providing an alternative route for oil to flow back to the reservoir. This shunt regulator allows the pump to discharge a fixed rate of oil flow (for a constant electric motor speed) under all hydraulic operating conditions.

[2]Note also how identical reservoir symbols may be placed at different locations of the diagram although they represent the exact same reservoir. This is analogous to "ground" symbols in electronic schematic diagrams, every ground symbol representing a common connection to the same zero-potential point.

An alternative to using a shunt regulating valve in a hydraulic system is to use a *variable-displacement pump*. Variable-displacement pumps still output a certain volume of hydraulic oil per shaft revolution, but that volumetric quantity may be varied by moving a component within the pump. In other words, the pump's per-revolution displacement of oil may be externally adjusted.

If we connect the variable-displacement mechanism of such a hydraulic pump to a pressure-sensing element such as a bellows, in a way where the pump senses its own discharge pressure and adjusts its volumetric output accordingly, we will have a pressure-regulating hydraulic system that not only prevents the pump from "locking" when the spool valve turns off, but also saves energy by not draining pressurized oil back to the reservoir:



Note the placement of a filter at the inlet of the pump in all hydraulic systems. Filtration is an absolute essential for any hydraulic system, given the extremely tight tolerances of hydraulic pumps, motors, valves, and cylinders. Even very small concentrations of particulate impurities in hydraulic oil may drastically shorten the life of these precision components.

Pneumatic fluid power systems require cleanliness as well, since any particulate contamination in the air will likewise cause undue wear in the close-tolerance compressors, motors, valves, and cylinders. Unlike hydraulic oil, compressed air is not a natural lubricant, which means many pneumatic power devices benefit from a small concentration of oil vapor in the air. Pneumatic "oilers" designed to introduce lubricating oil into a flowing air stream are generally located very near the point of use (e.g. the motor or the cylinder) to ensure the oil does not condense and "settle" in the air piping.

Fluid power systems in general tend to be inefficient, requiring much more energy input to

the fluid than what is extracted at the points of use[3]. When large amounts of energy need to be transmitted over long distances, electricity is the a more practical medium for the task. However, fluid power systems enjoy certain advantages over electric power, a few of which are listed here:

- Fluid power motors and cylinders do not overload at low speeds or under locked conditions
- Fluid power systems present little hazard of accidently igniting flammable atmospheres (no sparks produced)
- Fluid power systems present little or no fire hazard
- Fluid power systems present no hazard of electric shock or arc flash
- Fluid power systems are often easier to understand than electric systems
- Fluid power systems may be safely used in submerged (underwater) environments
- Pneumatic systems are relatively easy to equip with back-up energy reserve (e.g. liquefied nitrogen serving as a back-up gas supply in the event of compressor shut-down)
- Pneumatic systems are self-purging (i.e. enclosures housing pneumatic devices will be naturally purged of dusts and vapors by leaking air)

10.3 Solenoid valve actuators

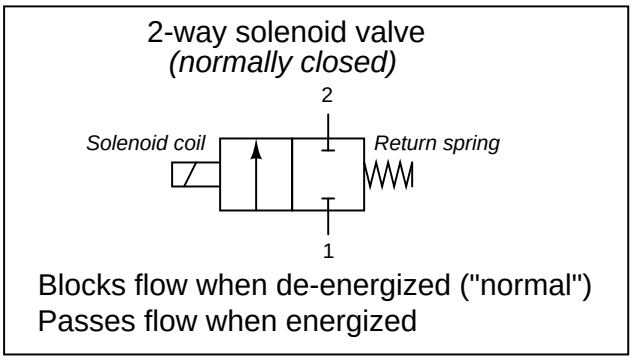
A very common form of on/off valve used for pneumatic and hydraulic systems alike is the *solenoid valve*. A "solenoid" is nothing more than a coil of wire designed to produce a magnetic field when energized. Solenoid actuators work by attracting a movable iron *armature* into the center of the solenoid coil when energized, the force of this attraction working to actuate a small valve mechanism.

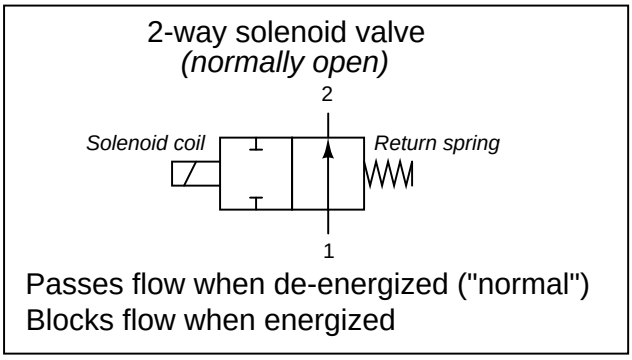
Solenoid-actuated valves are usually classified according to the number of ports ("ways"). A simple on/off solenoid valve controlling flow into one port and out of another port is called a *2-way* valve. Another style of solenoid valve, where flow is directed in one path or to another path – much like a single-pole double-throw (SPDT) electrical switch – is called a *3-way* valve because it has three fluid ports.

[3] Close-coupled hydraulic systems with variable-displacement pumps and/or motors may achieve high efficiency, but they are the exception rather than the rule. One such system I have seen was used to couple a diesel engine to the drive axle of a large commercial truck, using a variable-displacement pump as a continuously-variable transmission to keep the diesel engine in its optimum speed range. The system was so efficient, it did not require a cooler for the hydraulic oil!

10.3.1 2-way solenoid valves

Solenoid valve symbols often appear identical to fluid power valve symbols, with "boxes" representing flow paths and directions between ports in each of the valve's states. Like electrical switches, these valve symbols are always drawn in their "normal" (de-energized) state, where the return spring's action determines the valve position:





Unlike electrical switches, of course, the terms *open* and *closed* have opposite meanings for valves. An "open" electrical switch constitutes a break in the circuit, ensuring no current; an "open" valve, by contrast, freely allows fluid flow through it. A "closed" electrical switch has continuity, allowing current through it; a "closed" valve, on the other hand, shuts off fluid flow.

The arrow inside a solenoid valve symbol actually denotes a preferred direction of flow. Most solenoid valves use a "globe" or "poppet" style of valve element, where a metal plug covers up a hole (called the "seat"). Process fluid pressure should be applied to the valve in such a way that the pressure difference tends to hold the solenoid valve in its "normal" position (the same position as driven by the return spring). Otherwise[4], enough fluid pressure might override the return spring's

[4] One could argue that enough fluid pressure could override the solenoid's energized state as well, so why choose to have the fluid pressure act in the direction of helping the return spring? The answer to this (very good) question is that the solenoid's energized force exceeds that of the return spring. This is immediately obvious on first inspection, as the solenoid *must* be stronger than the return spring or else the solenoid valve would never actuate! Realizing this, now, we see that the spring is the weaker of the two forces, and thus it makes perfect sense why we should use the

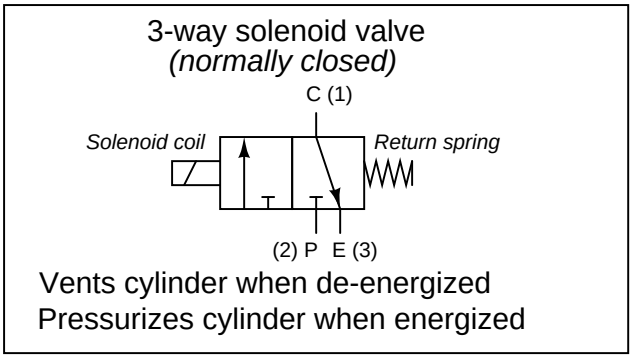
action, preventing the valve from achieving its "normal" state when de-energized. Thus, we see that the label "2-way" does not refer to two directions of flow as one might assume, but rather two *ports* on the valve for fluid to travel through.

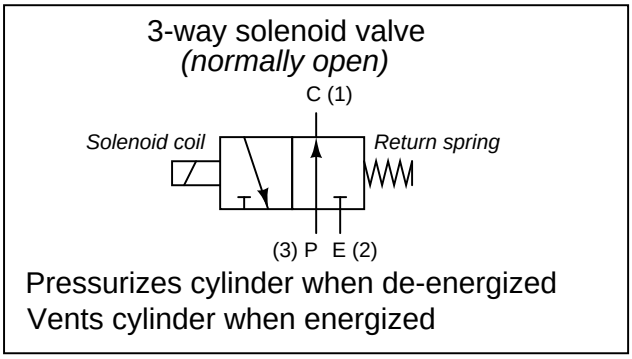
Some solenoid valves are designed in such a way that the direction of fluid flow through them is irrelevant. In such valves, the arrow symbols will be double-headed (one head at each end, pointing in opposite directions) to show the possibility of flow in either direction.

valve in such a way that the process pressure helps the spring: the solenoid's force has the best chance of overcoming the force on the plug produced by process pressure, so those two forces should be placed in opposition, while the return spring's force should work *with* (not against) the process pressure.

10.3.2 3-way solenoid valves

3-way solenoid valves have three ports for fluid, with two positions customarily referred to as *normally-open* and *normally-closed*. Ports on a 3-way valve are commonly labeled with the letters "P," "E," and "C," representing *Pressure* (compressed air supply), *Exhaust* (vent to atmosphere), and *Cylinder* (the actuating mechanism), respectively. Alternatively, you may see the cylinder port labeled "A" (for *actuator*) instead of "E".





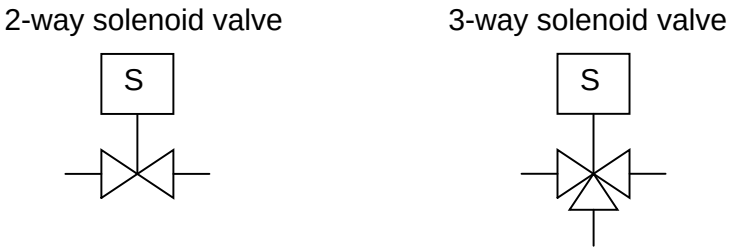
Alternatively, the numbers 1, 2, and 3 may be used to label the same ports. However, the numbers do not consistently refer to pressure source (P) and exhaust (E) ports, but rather to the 3-way valve's "normal" versus "actuated" statuses. A 3-way valve will pass fluid between ports 1 and 3 in its "normal" (de-energized) state, and pass fluid between ports 1 and 2 in its energized state. The following table shows the correspondence between port numbers and port letters for both styles of 3-way solenoid valve:

Valve type	Pressure (P) port	Exhaust (E) port	Cylinder (C) port
Normally-closed	2	3	1
Normally-open	3	2	1

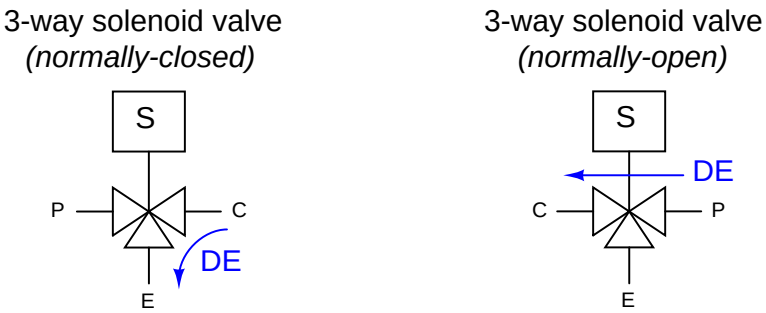
As with 2-way solenoid valves, the arrows denote preferred direction of fluid flow. Bidirectional

3-way valves will be drawn with double-headed arrows (pointing both directions).

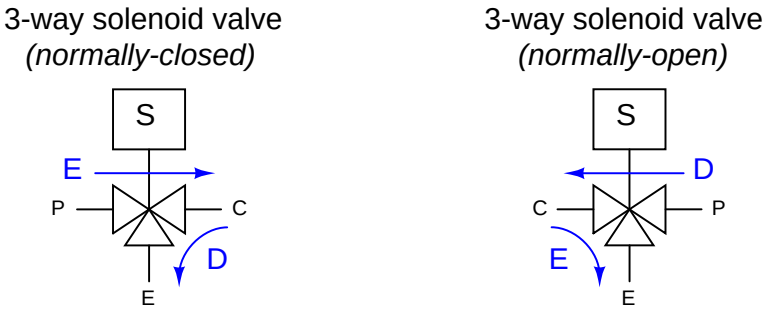
A different symbology is used in loop diagrams and P&IDs than that found in fluid power diagrams – one more resembling general valve symbols in instrumentation:



Unfortunately, these symbols are not nearly as descriptive as those used in fluid power diagrams. In order to show directions of flow (especially for 3-way valves), one must add arrows showing "normal" (de-energized, DE) flow directions:



Alternatively, a *pair* of arrows shows the directions of flow in both energized (E) and de-energized (D) states:

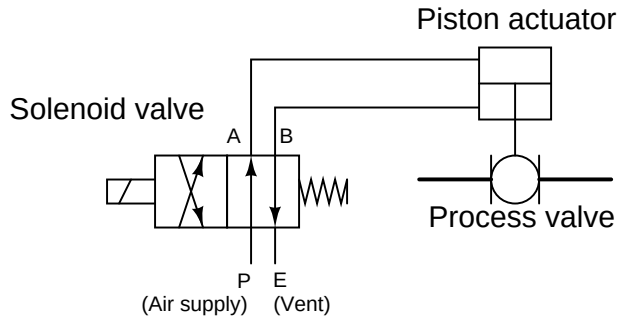


Photographs of an actual 3-way solenoid valve (this one manufactured by ASCO) appear here:

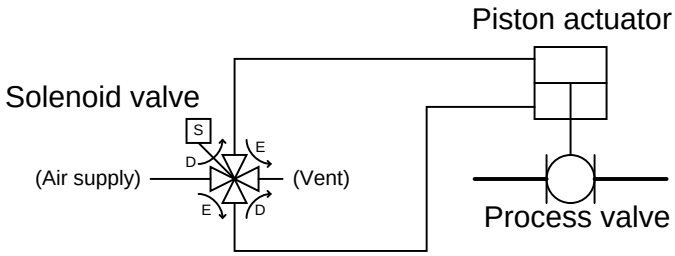
A view of the nameplate for this particular solenoid valve reveals some of its ratings and characteristics:

10.3.3 4-way solenoid valves

When a pneumatic actuator requires air pressure applied to two different ports in order to move two different directions (such as the case for cylinders lacking a return spring), the solenoid valve supplying air to that actuator must have four ports: one for air supply (P), one for exhaust (E), and two for the cylinder ports (typically labeled A and B). The following diagram shows a 4-way solenoid valve connected to the piston actuator of a larger (process) ball valve:



The same diagram could be drawn using the "triangle" solenoid valve symbols rather than the "block" symbols more common to fluid power diagrams:

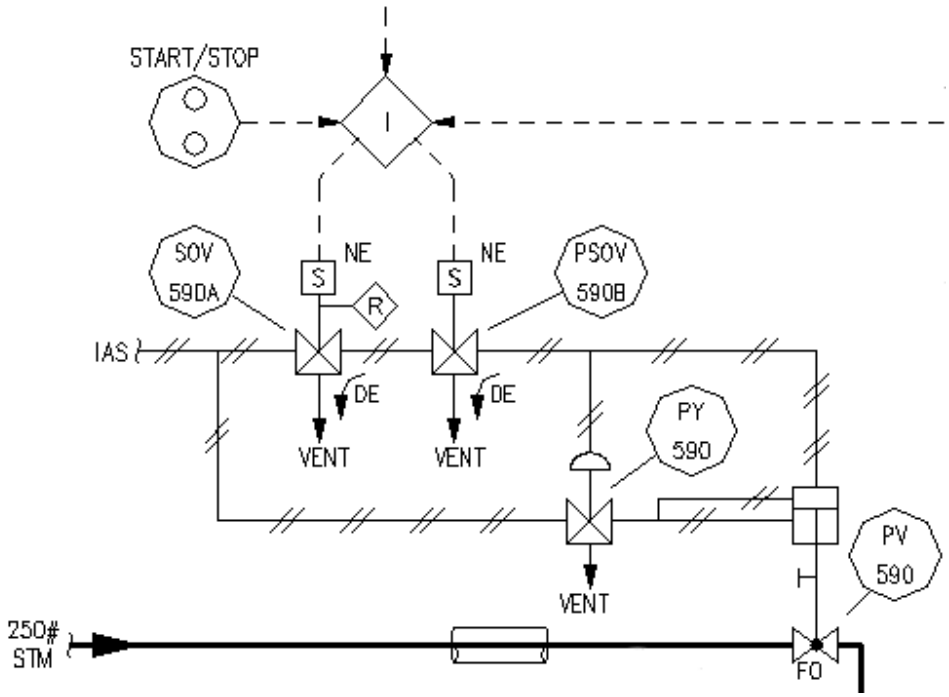


Here, the letters "D" and "E" specify which directions air is allowed to flow when the solenoid is de-energized and energized, respectively.

10.3.4 Normal energization states

Solenoid valves may be used in such a way that they spend most of their time de-energized, energizing only for brief periods of time when some special function is required. Alternatively, solenoids may be maintained in an energized state, and de-energized to perform their design function. The choice to use a solenoid's energized or de-energized state to perform a specific function is left to the system designer, but nevertheless it is important for all maintenance personnel to know in order to perform work on a solenoid-controlled system.

Take the following segment of a P&ID for a steam turbine-driven pump control system for example, where a pair of 3-way solenoid valves control instrument air pressure to a piston-actuated steam valve to start the turbine in the event that an electric motor-driven pump happens to fail:



If *either* of the two solenoid valves de-energizes, instrument air pressure will vent from the top of the piston actuator to atmosphere, causing the steam valve to "fail" to the full-open position and send steam to the turbine. This much is evident from the curved arrows showing air flowing to the "Vent" ports in a de-energized (DE) condition. An additional valve (PY-590) ensures the piston actuator's motion by simultaneously applying air pressure to the bottom of the actuator if ever air is vented from the top. As an additional feature, the left-hand solenoid valve (SOV-590A) has a manual "Reset" lever on it, symbolized by the letter "R" inside a diamond outline.

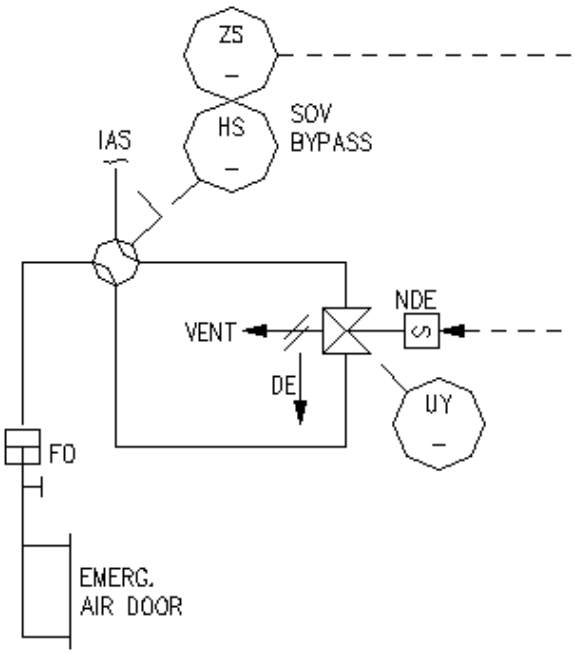
The only indication of the solenoids' typical status (energized or de-energized) comes from the letters "NE" next to each solenoid coil. In this case, "NE" stands for *normally energized*. Therefore, this steam turbine control system, which serves as a back-up to an electric motor-driven pump, relies on either (or both) of the solenoid valves *de-energizing* to make the turbine start up. Under

"normal" conditions, where the turbine is not needed, the solenoids remain energized and the steam valve remains shut.

Unfortunately, this use of the word "normal" is altogether different from the use of the word "normal" when describing a solenoid valve's open/close characteristics. Recall that a *normally open* solenoid valve allows fluid to pass through when it is de-energized. A *normally closed* solenoid valve, by contrast, shuts off fluid flow when de-energized. In this context, the word "normally" refers to the *unpowered* state of the solenoid valve. This is quite similar to how the word "normally' is used to describe switch contact status: a normally-open (NO) electrical switch is open when unactuated; a normally-closed (NC) electrical switch is closed when unactuated. In both cases, with solenoid valves and with electrical switches, the word "normally" refers to the *condition of minimum stimulus*.

However, when an engineer designs a solenoid control system and declares a solenoid to be "normally energized," that engineer is describing the *typical* status of the solenoid valve *as it is intended to function in the process*. This may or may not correspond to the *manufacturer's* definition of "normally," since the solenoid manufacturer cannot possibly know which state any of their customers intends to have their solenoid valve typically operate in. To illustrate using the previous steam turbine control system P&ID, those two solenoid valves would be considered *normally closed* by the manufacturer, since their de-energized states block air flow from the "P" port to the "C" port and vent air pressure from the "C" port to the "E" (vent) port. However, the engineer who designed this system wanted both solenoids to be energized whenever the turbine was not needed to run (the "normal" state of the process), and so the engineer labeled both solenoid coils as *normally energized*, which means both solenoids will be actuated to pass air pressure from their "P" ports to their "C" ports (and close off the vent ports) under typical conditions. Here, we see the manufacturer's definition of "normal" is precisely opposite that of the process engineer's definition of "normal" for this application.

The manufacturer's and process engineer's definitions of "normally" are not always in conflict. Take for example this P&ID segment, showing the solenoid control of an air vent door on a large furnace, designed to open up if the forced-draft fan (blowing combustion air into the furnace) happens to stop for any reason:



Here we have a *normally open* solenoid valve, designed by the manufacturer to pass instrument air pressure from the pressure ("P") port to the cylinder ("C") port when de-energized. The straight arrow with the "DE" label next to it reveals this to be the case. Instrument air pressure sent to the air door actuator holds the door shut, meaning the air door will swing open if ever instrument air pressure is vented by the solenoid. For this particular solenoid, this would require an *energized* condition.

The process engineer designing this emergency Air Door control system choose to operate the solenoid in its de-energized state under typical operating conditions (when the furnace air door should be shut), a fact revealed by the letters "NDE" (normally de-energized) next to the solenoid coil symbol. Therefore, the "normal" process operating condition for this solenoid happens to be de-energized, which makes the manufacturer's definition of "normal" match the engineer's definition of "normal." The solenoid valve should be open (passing air to the door's actuating cylinder) under "normal" operating conditions.

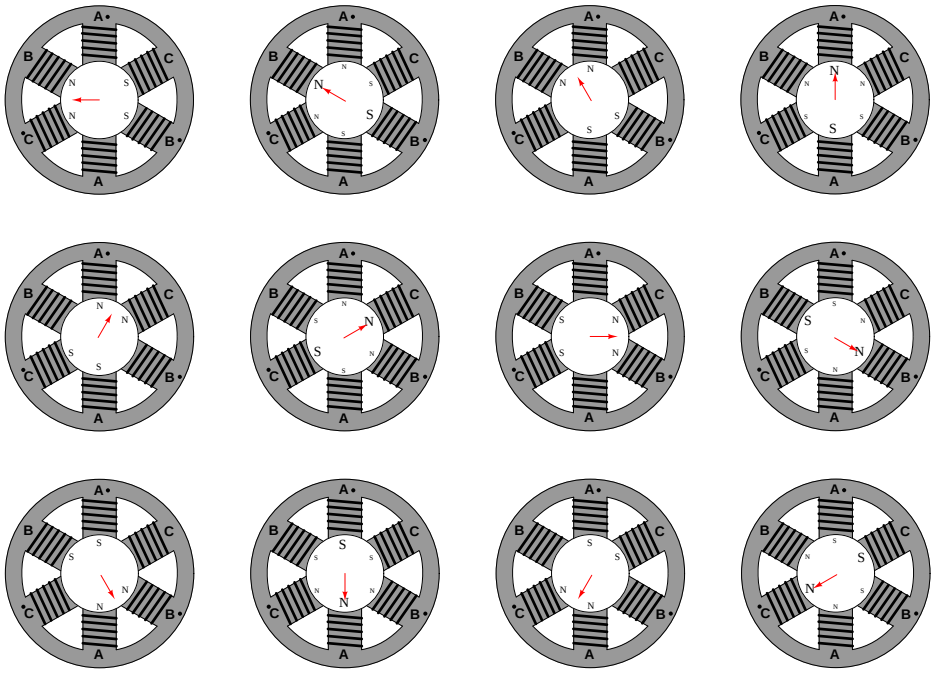
10.4 On/off electric motor control circuits

An electric motor is often used as a discrete control element in a control system if driving a pump, conveyor belt, or other machine for the transportation of a process substance. As such, it is important to understand the functioning of motor control circuits.

Of all the available electric motor types, the most common found in industrial applications (by far) is the three-phase AC induction motor. For this reason, this section of the book will focus exclusively on this type of motor as a final control element.

10.4.1 AC induction motors

The basic principle of an AC induction motor is that one or more out-of-phase AC (sinusoidal) currents energize sets of electromagnet coils (called *stator* coils or windings) arranged around the circumference of a circle. As these currents alternately energize the coils, a magnetic field is produced which "appears" to rotate around the circle. This *rotating magnetic field* is not unlike the appearance of motion produced by a linear array of light bulbs blinking on and off in sequence: although the bulbs themselves are stationary, the out-of-phase sequence of their on-and-off blinking makes it appear as though a pattern of light "moves" along the length of the array. Likewise, the superposition of magnetic fields created by the out-of-phase coils resembles a magnetic field of constant intensity revolving around the circle. The following twelve images show how the magnetic field vector (the red arrow) is generated by a superposition of magnetic poles through one complete cycle (1 revolution), viewing the images from left to right, top to bottom (the same order as you would read words in an English sentence):



If an electrically conductive object is placed within the circle on a shaft so that it is free to rotate, the relative motion between the rotating magnetic field and the conductive object induces electric currents in the conductive object, which produce magnetic fields of their own. Lenz's Law tells us that the effect of these induced magnetic fields will be to try to oppose change: in other words, the induced fields react against the rotating magnetic field of the stator coils in such a way as to minimize the relative motion. This means the conductive object will try to rotate in sync with the stator's rotating magnetic field. In a typical *squirrel-cage* induction motor design, the rotor is made up of aluminum bars joining two aluminum "shorting rings," one at either end of the rotor. Iron fills the spaces between the rotor bars to provide a lower-reluctance magnetic "circuit" between stator poles than would be otherwise formed if the rotor were simply made of aluminum.

A photograph of a small, disassembled three-phase AC induction "squirrel-cage" motor is shown here, revealing the construction of the stator coils and the rotor:

Given the simple design of AC induction motors, they tend to be quite reliable machines. So long as the stator coil insulation is not damaged by excessive moisture, heat, or chemical exposure, they will just about run forever. The only "wearing" components are the bearings supporting the rotor shaft, and those tend to be easy to replace.

Starting a three-phase induction motor is as simple as applying full power to the stator windings. When this happens, the motor will draw a large amount of current (as much as ten times its normal running current) called the *inrush* current, causing the rotor to produce a large mechanical torque. As the rotor gains speed, the current reduces to a normal level, with the speed approaching the "synchronous" speed of the rotating magnetic field[5].

Reversing the rotational direction of a three-phase motor is as simple as swapping any two out of three power conductor connections. This has the effect of reversing the *phase sequence* of the power "seen" by the motor[6].

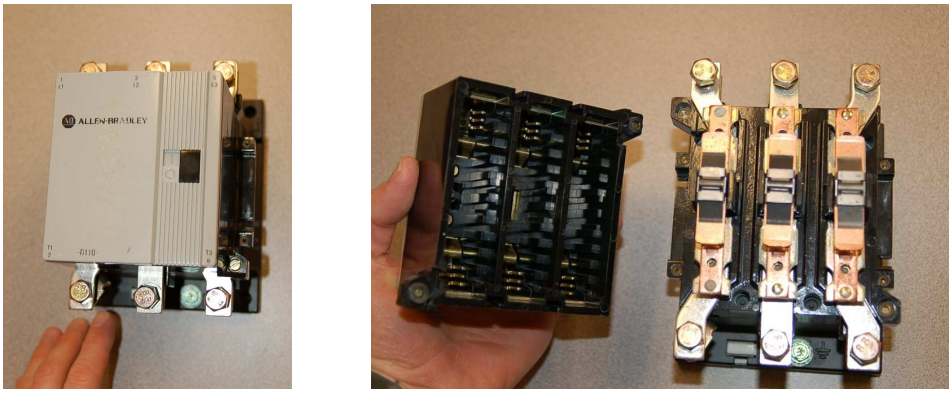
[5] The rotor can never fully achieve the same rotational speed as the magnetic field produced by the stator windings, because if it did there would be zero relative motion between the rotating magnetic field and the rotating rotor, and thus no induction of currents in the rotor bars to create the induced magnetic fields necessary to produce a reaction torque. Thus, the rotor must "slip" behind the speed of the rotating magnetic field in order to produce a torque, which is why the full-load speed of an induction motor is always just a bit slower than the synchronous speed of the rotating magnetic field (e.g. a 4-pole motor with a synchronous speed of 1800 RPM will rotate at approximately 1750 RPM).

[6] This principle is not difficult to visualize if you consider the phase sequence as a repeating pattern of letters, such as `ABCABCABC`. Obviously, the reverse of this sequence would be `CBACBACBA`, which is nothing more than the original sequence with letters A and C transposed. However, you will find that transposing *any* two letters of the original sequence transforms it into the opposite order: for example, transposing letters A and B turns the sequence `ABCABCABC` into `BACBACBAC`, which is the same *order* as the sequence `CBACBACBA`.

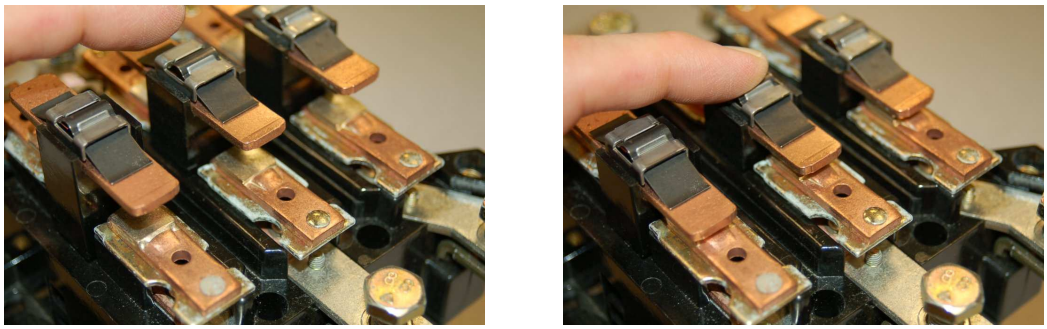
10.4.2 Starter (contactor) relays

Due to the large "inrush" currents at start-up time for a three-phase induction motor, a special form of electromechanical relay with high overcurrent capacity is used to make and break the three-phase line connections to the motor's stator winding terminals. These special motor-starting relays are commonly referred to as *starters* or *contactors*.

A photograph of a motor starter, or contactor, rated at 75 horsepower (assuming 480 volt AC 3-phase power) is shown here, both assembled and with the top cover removed to reveal the three sets of high-current electrical contacts:



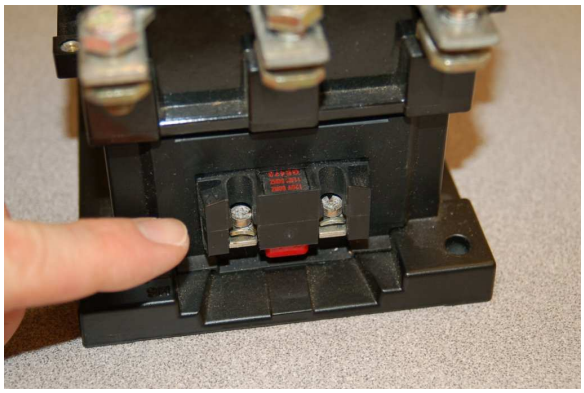
Each contact is actually a series pair of contacts that make and break simultaneously with the actuation of an iron armature attracted by an electromagnet coil in the base of the contactor assembly. The operation of the three contact sets may be seen in this pair of photographs, the left-hand image showing the contacts in their normal (open) state, and the right-hand image showing the contacts closed by the force of my finger:



Of course, it would be highly unsafe to touch or manually actuate the contacts of a motor starting relay with the cover removed as shown. Not only would there be an electric shock hazard from touching any one of the bare copper contacts with your finger, but the arc blast produced by closing and opening such contacts would pose a burn and blast hazard. This is why all modern motor contactors are equipped with arc shield covers.

The actual contact pads are not made of pure copper, but rather silver (or a silver alloy) designed to survive the repeated arcing and blasting action of large AC currents being initiated and interrupted.

Below the main power (line) connection terminals on this starter hide two small screw terminals providing connection points to the electromagnet coil actuating the starter:

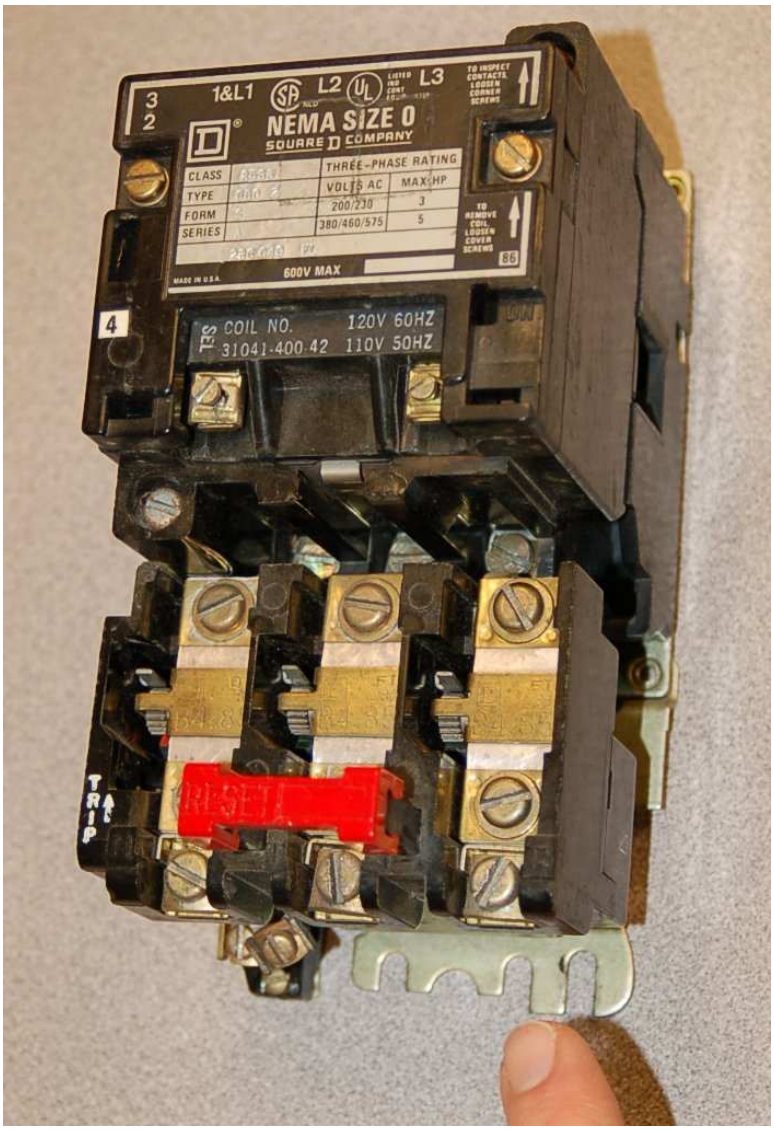


Like nearly every other three-phase contactor in existence, this one's coil is rated for 120 volts AC. Although the electric motor may operate on three-phase, 480 volt AC power, the contactor coil and the rest of the control circuitry operates on a lower voltage for reasons of safety.

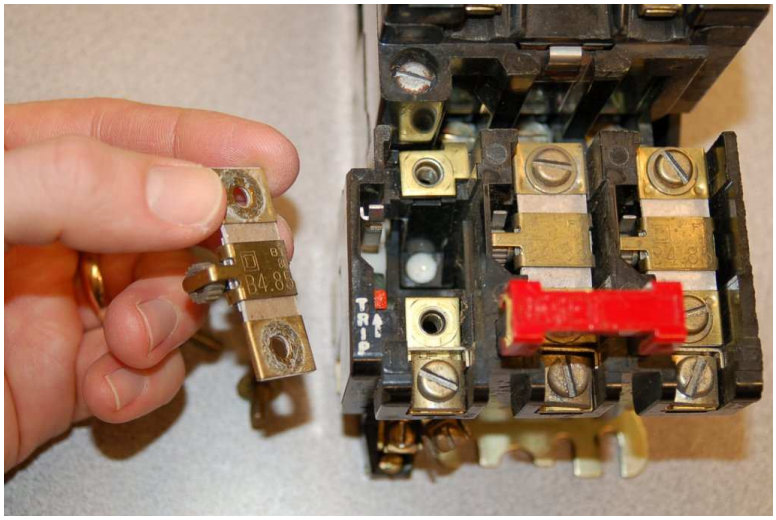
10.4.3 Motor overload protective devices

An essential component of any motor control circuit is some device to detect a condition of excessive *overload* and interrupt power to the motor before thermal damage will occur to it. A very simple and common overload protective device is known as an *overload heater*, consisting of resistive elements connected in series with the three lines of a 3-phase AC motor, designed to heat and to cool at rates modeling the thermal characteristics of the motor itself.

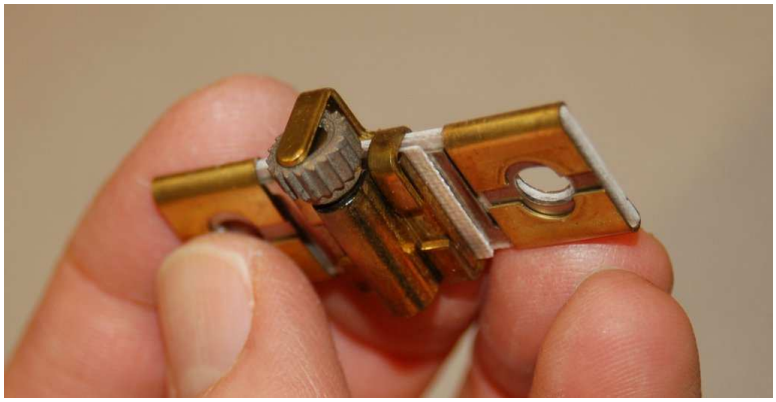
The following photograph shows a three-phase starter (contactor) relay joined together with a set of three "overload heaters" through which all of the motor's current flows. The overload heaters appear as three brass-colored metal strips near a red push-bar labeled "Reset:"



Removing one of the heater elements reveals its mechanical nature: a small toothed wheel on one side engages with a lever when it is bolted into place in the overload assembly. That lever connects to a spring-loaded mechanism charged by the manual actuation of the red "Reset" push-bar, which in turn actuates a small set of electrical switch contacts:



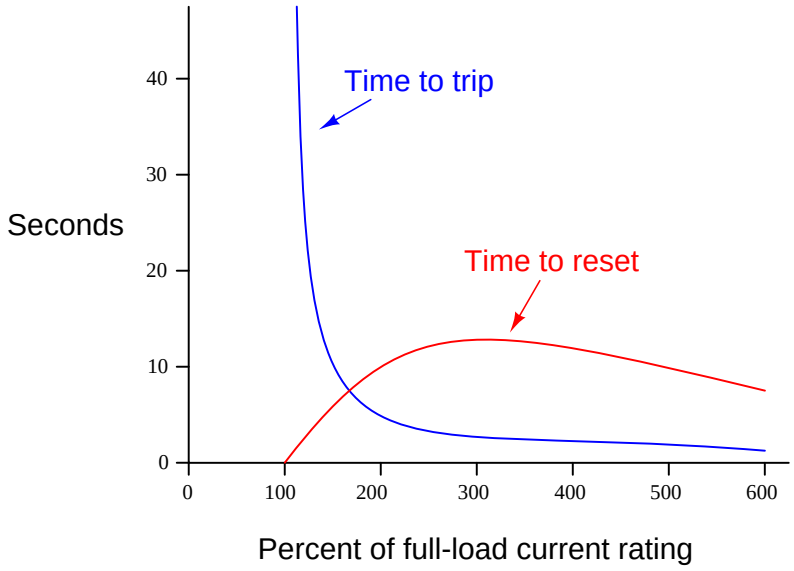
The purpose of the overload heater is to heat up as the motor draws excessive current. The small toothed wheel is held in place by a rod immersed in a solidified mass of solder, encased in a brass cylinder underneath the heater strip. The next photograph shows the underside of the heater element, with the toothed wheel and brass cylinder plainly visible:



If the heater element becomes too hot (due to excessive motor current), the solder inside the brass cylinder will melt, allowing the toothed wheel to spin. This will release spring tension in the overload mechanism, allowing the small electrical switch to spring to an open state. This "overload contact" then interrupts current to the motor starter's electromagnet coil, causing the starter to de-energize and the motor to stop.

Manually pressing the "Reset" push-bar will re-set the spring mechanism and re-close the overload contact, allowing the starter to energize once more, but only once the overload heater element has cooled down enough for the solder inside the brass cylinder to re-solidify. Thus, this simple mechanism prevents the overloaded motor from being immediately re-started after a thermal overload "trip" event, giving it time to cool down as well.

A typical "trip curve" for a thermal overload unit is shown here, with time plotted against the severity of the overcurrent level:

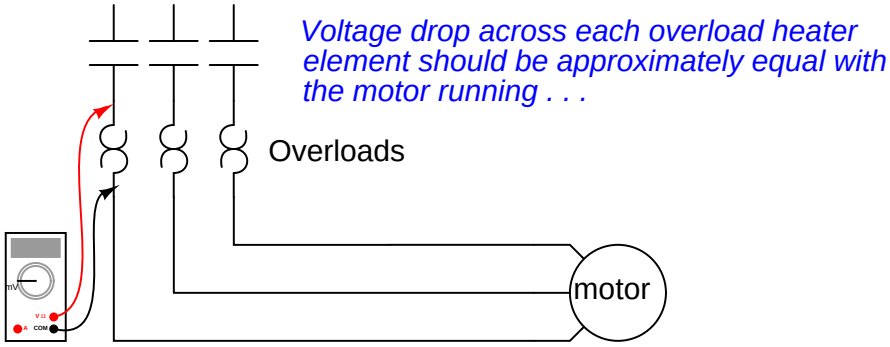


In contrast to a circuit breaker or fuse – which is sized to protect the power wiring from overcurrent heating – the overload heater elements are sized specifically to protect the *motor*. As such, they act as thermal models of the motor itself, heating to the "trip" point just as fast as the motor itself will heat to the point of maximum rated temperature, and taking just as long to cool to a safe temperature as the motor will. Another difference between overload heaters and breakers/fuses is that the heaters are not designed to directly interrupt current by opening[7], as fuses or breakers do. Rather, each overload heater serves the simple purpose of *warming* proportionately to the magnitude and time duration of motor overcurrent, causing a different electrical contact to open, which in turn triggers the starter relay to open.

Of course, overload heaters only work to protect the motor from thermal overload if they experience similar ambient temperature conditions. If the motor is situated in a very hot area of the industrial process unit, whereas the overload elements are located in a climate-controlled "motor control center" (MCC) room, they may fail to protect the motor as designed. Conversely, if the overload heaters are located in a hot room while the motor is located in a freezing-cold environment (e.g. the MCC room lacks air conditioning while the motor is located in a freezer), they may "trip" the motor prematurely.

[7] This is not to say overload heaters cannot fail open, because they can and will under extraordinary circumstances. However, opening like a fuse is not the design function of an overload heater.

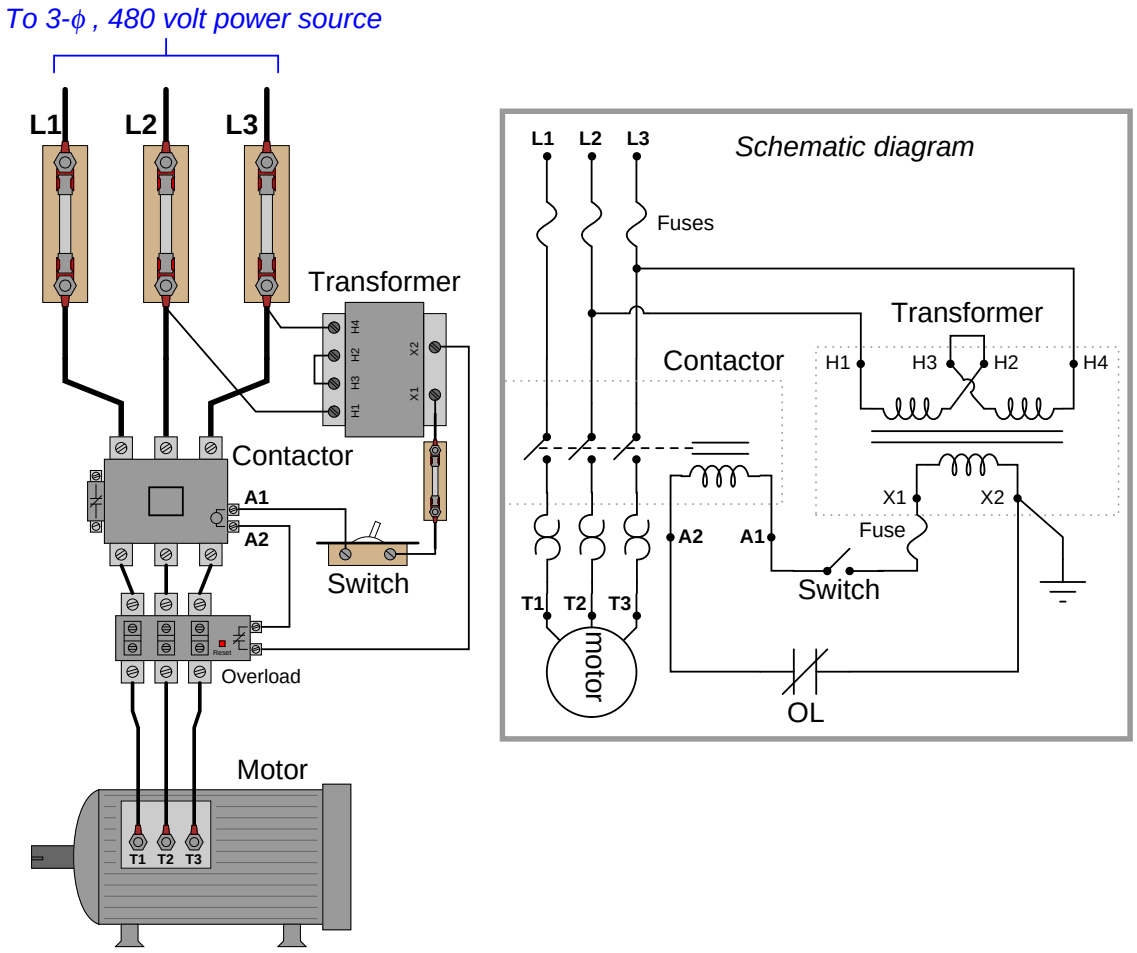
An interesting "trick" to keep in mind for motor control circuit diagnosis is that overload heaters are nothing more than low-value resistors. As such, they will drop small amounts of voltage (usually quite a bit less than 1 volt AC) under full load current. This voltage drop may be used as a simple, qualitative measure of motor phase current. By measuring the voltage dropped across each overload heater (with the motor running), one may ascertain whether or not all phases are carrying equal currents. Of course, overload heaters are not precise enough in their resistance to serve as true current-measuring "shunts," but they are more than adequate as qualitative indicators of relative phase current, to aid you in determining (for instance) if the motor suffers from an open or high-resistance phase winding:



As useful as thermal overload "heaters" are for motor protection, there are more effective technologies available. An alternative way to detect overloading conditions is to monitor the temperature of the stator windings directly, using thermocouples or (more commonly) RTDs, which report winding temperatures to an electronic "trip" unit with the same control responsibilities as an overload heater assembly. This sophisticated approach is used on large (thousands of horsepower) electric motors, and/or in critical process applications where motor reliability is paramount. Machine vibration equipment used to monitor and protect against excessive vibration in rotary machines is often equipped with such temperature-sensing "trip" modules just for this purpose. Not only can motor winding temperatures be monitored, but also bearing temperatures and other temperature-sensitive machine components so that the protective function extends beyond the health of the electric motor.

10.4.4 Motor control circuit wiring

A simple three-phase, 480 volt AC motor-control circuit is shown here, both in pictorial and schematic form:

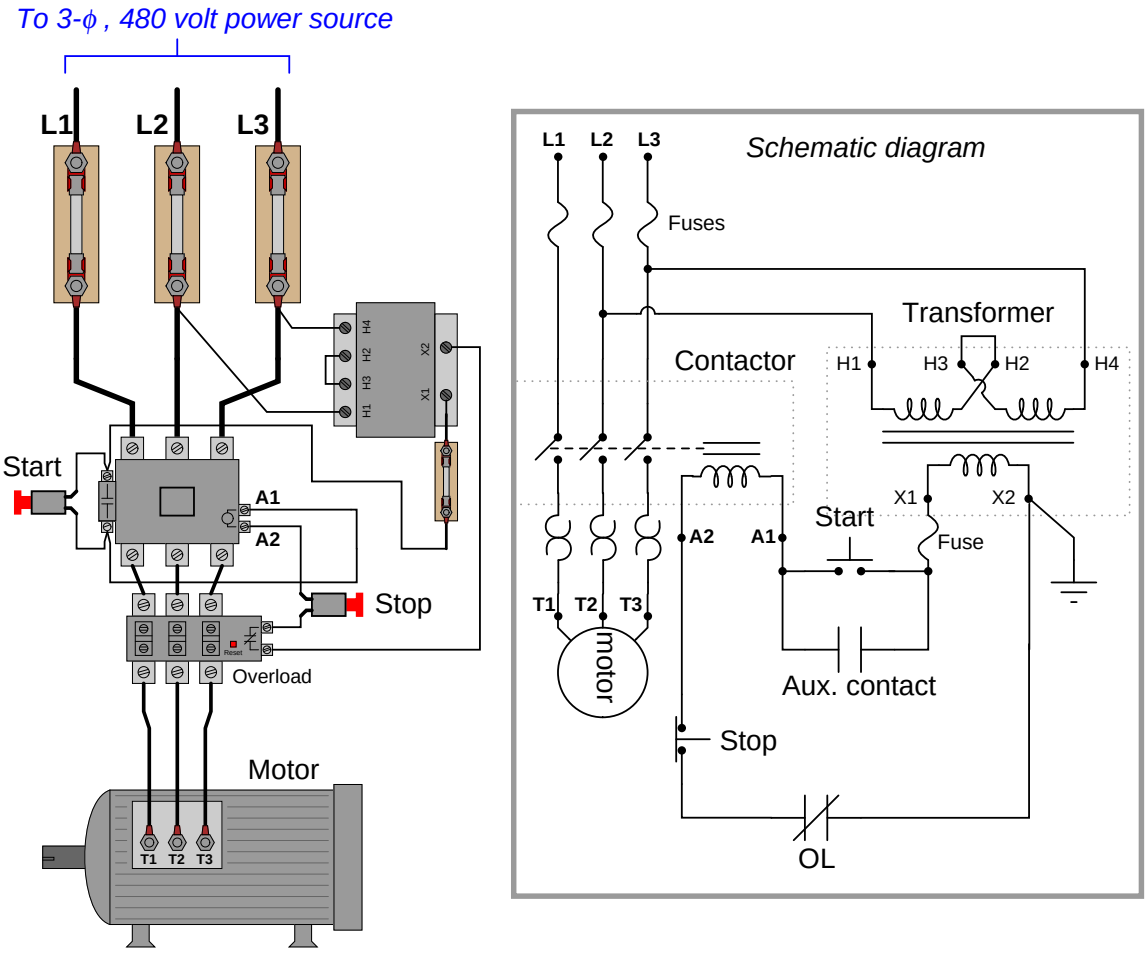


Note how a *control power transformer* steps down the 480 volt AC to provide 120 volt AC power for the starter coil to operate on. Furthermore, note how the overload ("OL") contact is wired in series with the starter's coil so that a thermal overload event forces the starter to de-energize and thus interrupt power to the motor even if the control switch is still in the "on" position. The overload heaters appear in the schematic diagram as pairs of back-to-back "hook" shapes, connected in series with the three "T" lines of the motor. Remember that these "OL" heater elements do not directly interrupt power to the motor in the event of an overload, but rather signal the "OL" contact to open up and de-energize the starter (contactor).

In an automatic control system, the toggle switch would be replaced by another relay contact (that relay controlled by the status of a process), a process switch, or perhaps the discrete output

channel of a programmable logic controller (PLC).

It should be noted that a toggling-style of switch is necessary in order for the motor to continue to run after a human operator actuates the switch. The motor runs with the switch is in the closed state, and stops when the switch opens. An alternative to this design is to build a *latching* circuit allowing the use of momentary contact switches (one to start, and one to stop). A simple latching motor control circuit is shown here:

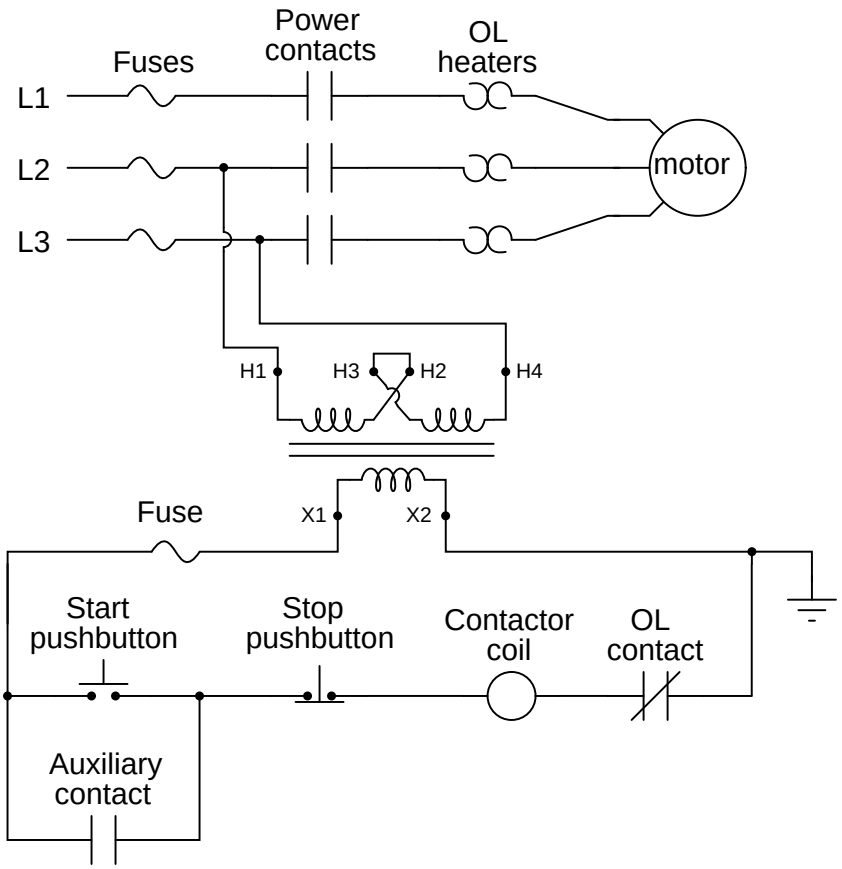


In this circuit, an *auxiliary contact* actuated by the motor starter (contactor) is wired in parallel with the "Start" pushbutton switch, so that the motor contactor continues to receive power after the operator releases the switch. This parallel contact – sometimes called a *seal-in contact* – latches the motor in an "on" state after a momentary closure of the "Start" pushbutton switch.

A normally-closed "Stop" switch provides a means to "un-latch" the motor circuit. Pressing this pushbutton switch stops current in the coil of the contactor, causing it to de-energize, which then

opens the three motor power contacts as well as the auxiliary contact that used to maintain the contactor's energized state.

A simple *ladder diagram* showing the interconnections of all components in this motor control circuit makes this system easier to understand:



Most on/off motor control circuits in the United States are some variation on this wiring theme, if not identical to it. Once again, this system could be automated by replacing the "Start" and "Stop" pushbutton switches with process switches (e.g. pressure switches for an air compressor control system) to make a system that starts and stops automatically. A programmable logic controller (PLC) may also be used to provide the latching function rather than an auxiliary contact on the motor starter. Once a PLC is included in the motor control circuit, a great many automatic control features may be added to enhance the system's capabilities. Examples include timing functions, motor cycle count functions, and even remote start/stop capability via a digital network connecting to operator interface displays or other computers.

References

Fitzgerald, A.E. and Higginbotham, David E., *Basic Electrical Engineering*, Second Edition, McGraw-Hill Book Company, New York, NY, 1957.

"General Service Solenoid Valves – 3/2 Series 8300/8315", document 8300R1, ASCO Valve Inc.

"ASCO Nuclear Catalog – Nuclear Products Qualified to IEEE Specifications", ASCO Valve Inc.

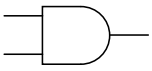
Croft, Terrell and Summers, Wilford I., *American Electrician's Handbook*, Eleventh Edition, McGraw-Hill Book Company, New York, NY, 1987.

Chapter 11

Relay control systems

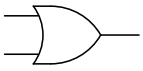
The word "discrete" means *individual* or *distinct*. In engineering, a "discrete" variable or measurement refers to a true-or-false condition. Thus, a discrete control system is one designed to operate on Boolean ("on" or "off") signals supplied by discrete sensors such as process switches.

A form of discrete control taught in every introductory course on digital electronics involves the use of circuits called *logic gates*. These circuits input one or more Boolean signals, and output a Boolean signal according to a simple rule such as "AND" or "OR":



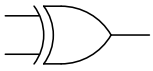
AND

A	B	Output
0	0	0
0	1	0
1	0	0
1	1	1



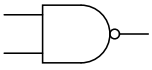
OR

A	B	Output
0	0	0
0	1	1
1	0	1
1	1	1



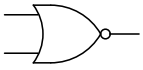
XOR

A	B	Output
0	0	0
0	1	1
1	0	1
1	1	0



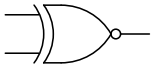
NAND

A	B	Output
0	0	1
0	1	1
1	0	1
1	1	0



NOR

A	B	Output
0	0	1
0	1	0
1	0	0
1	1	0



XNOR

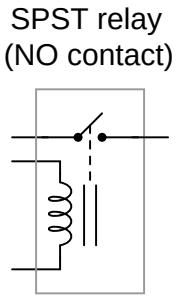
A	B	Output
0	0	1
0	1	0
1	0	0
1	1	1

Industrial control systems rarely utilize logic gates in a direct fashion for discrete control systems, although the fundamental *concepts* of "AND," "OR," and other gate types are universally applied. Instead, control functions are either implemented using electromechanical relays and/or

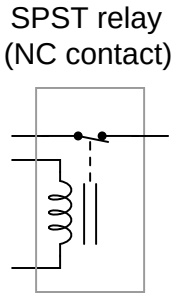
with programmable digital devices such as PLCs (Programmable Logic Controllers). This chapter focuses on the practical use of both technologies for industrial discrete control.

11.1 Control relays

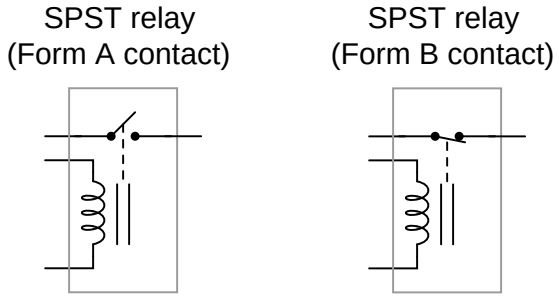
An *electromechanical relay* is an electrical switch actuated by an electromagnet coil. As switching devices, they exhibit simple "on" and "off" behavior with no intermediate states. The electronic schematic symbol for a simple single-pole, single-throw (SPST) relay is shown here:



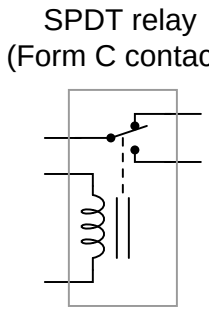
A coil of wire wrapped around a laminated iron core provides the magnetic field necessary to actuate the switch mechanism. This particular relay is equipped with *normally open* (NO) switch contacts, which means the switch will be in the open (off) state when the relay coil is de-energized. Recall from page 368 that the "normal" status of a switch is the condition of *minimum stimulus*. A relay switch contact will be in its "normal" status when its coil is not energized. A single-pole, single-throw relay with a normally-closed (NC) switch contact would be represented in an electronic schematic like this:



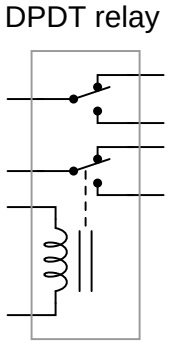
In the electrical control world, the labels "Form-A" and "Form-B" are synonymous with "normally open" and "normally closed" contact status. Thus, we could have labeled the SPST relay contacts as "Form-A" and "Form-B," respectively:



An extension of this theme is the single-pole, double-throw (SPDT) relay contact, otherwise known as a "Form-C" contact. This design of switch provides both a normally-open and normally-closed contact set in one unit, actuated by the electromagnet coil:



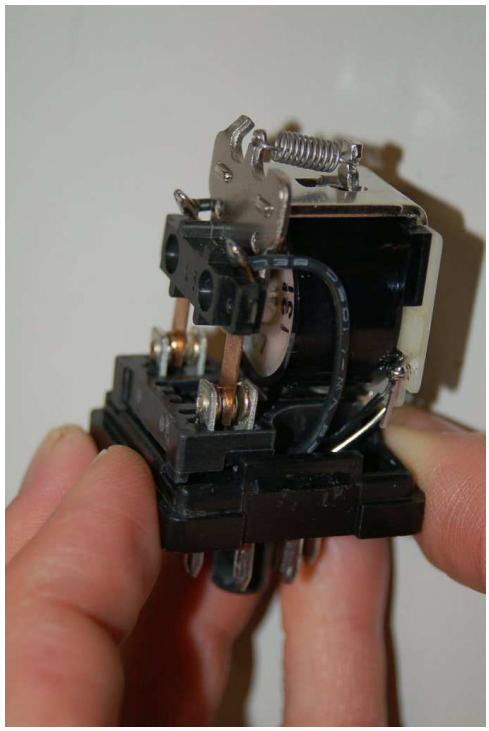
A further extension of this theme is the double-pole, double-throw (DPDT) relay contact. This design of switch provides two sets of Form-C contacts in one unit, simultaneously actuated by the electromagnet coil:



DPDT relays are some of the most common found in industry, due to their versatility. Each Form-C contact set offers a choice of either normally-open or normally-closed contacts, and the two

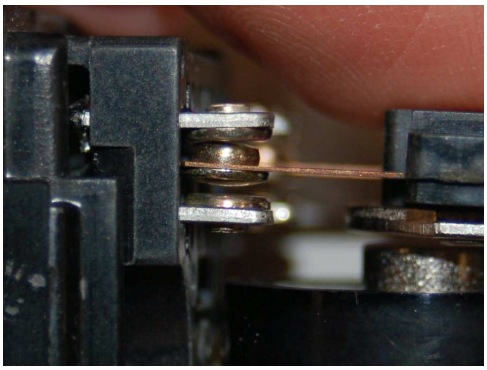
sets (two "poles") are electrically isolated from each other so they may be used in different circuits.

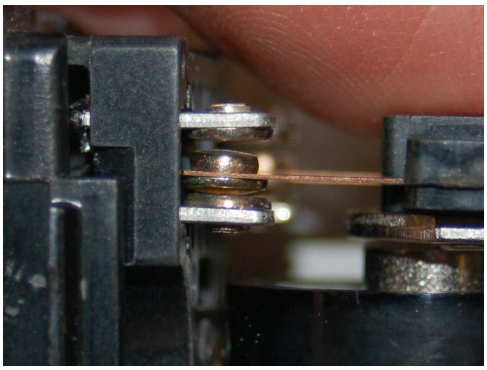
A common package for industrial relays is the so-called *ice cube relay*, named for its clear plastic case allowing inspection of the working elements. These relays plug into multi-pin base sockets for easy removal and replacement in case of failure. A DPDT "ice cube" relay is shown in the following photographs, ready to be plugged into its base (left) and with the plastic cover removed to expose both sets of Form-C contacts (right):



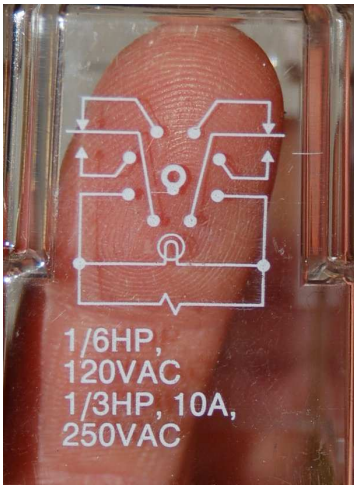
These relays connect to the socket with eight pins: three for each of the two Form-C contact set, plus two more pins for the coil connections. Due to the pin count (8), this style of relay base is often referred to as an *octal* base.

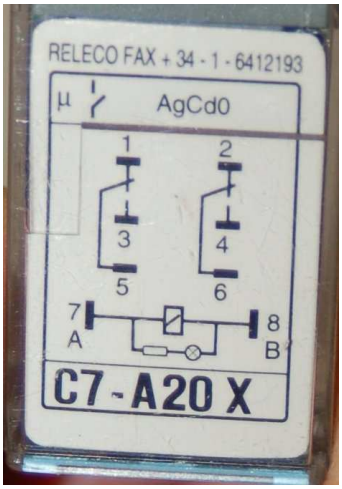
A closer view of one Form-C contact shows how the moving metal "leaf" contacts one of two stationary points, the actual point of contact being made by a silver-coated "button" at the end of the leaf. The following photographs show one Form-C contact in both positions:





Industrial control relays usually have connection diagrams drawn somewhere on the outer shell to indicate which pins connect to which elements inside the relay. The style of these diagrams may vary somewhat, even between relays of identical function. Take for instance the diagrams shown here, photographed on three different brands of DPDT relay:





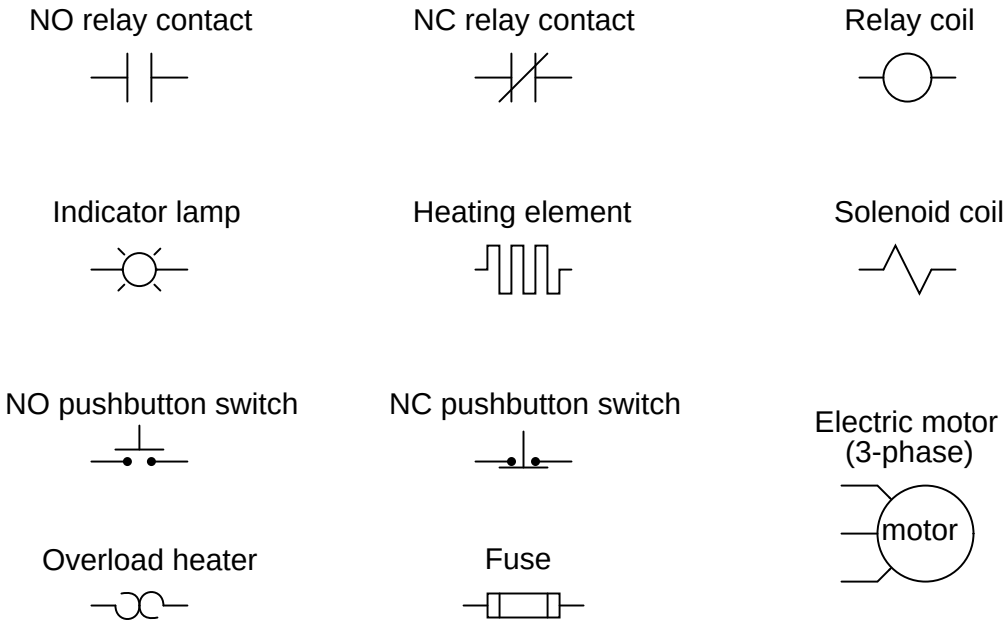
Bear in mind that these three relays are *identical* in their essential function (DPDT switching), despite differences in physical size and contact ratings (voltage and current capacities). Only two of the three diagrams shown use the same symbols to represent contacts, and all three use unique symbols to represent the coil.

11.2 Relay circuits

Electromechanical relays may be connected together to perform logic and control functions, acting as logic elements much like digital gates (AND, OR, etc.). A very common form of schematic diagram showing the interconnection of relays to perform these functions is called a *ladder diagram*. In a "ladder" diagram, the two poles of the power source are drawn as vertical rails of a ladder, with horizontal "rungs" showing the switch contacts, relay contacts, relay coils, and final control elements (lamps, solenoid coils, motors) drawn in between the power rails.

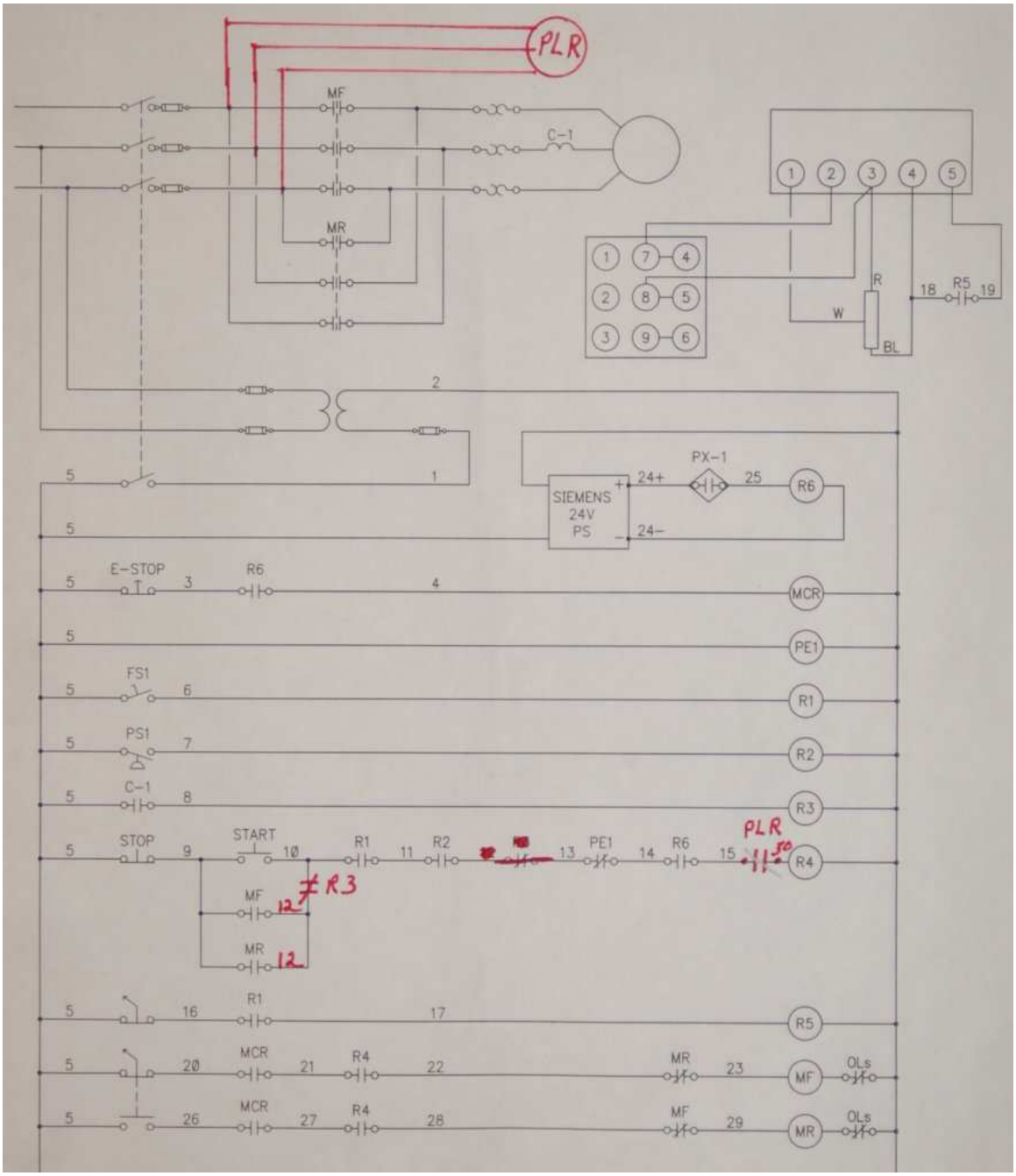
Ladder diagrams differ from regular schematic diagrams of the sort common to electronics technicians primarily in the strict orientation of the wiring: vertical power "rails" and horizontal control "rungs." Symbols also differ a bit from common electronics notation: relay coils are drawn as circles, with relay contacts drawn in a way that resembles capacitors:

Ladder diagram symbols



Another notable convention in relay circuits and their ladder diagrams is that each and every wire in the circuit is labeled with a number corresponding to common connection points. That is, wires connected together always bear the same number: the common number designates a condition of electrical commonality (all points bearing the same number are *equipotential* to each other). Wire numbers only change when the connection passes through a switch or other device capable of dropping voltage.

An actual ladder diagram of a relay-based motor control system is shown here, complete with *red-line* edits showing modifications to the circuit made by an industrial electrician:



References

Summers, Wilford I. and Croft, Terrell, *American Electrician's Handbook*, Eleventh Edition, McGraw-Hill Book Company, New York, NY, 1987.

Chapter 12

Programmable Logic Controllers

Electromechanical relays are versatile and reliable devices, and it is possible to perform virtually any control function given enough relays and wiring. However, they are subject to wear due to their moving parts, and re-configuring a relay-based control system entails disconnecting and re-connecting wires: a tedious task at best. Solid-state logic gates may replace electromechanical relays to avoid problems of wear (as well as vastly increase speed of response), but once again re-configuration would require disconnecting and re-connecting many wires between the gates.

A very practical alternative to relay- or gate-based control logic circuits is to use a *programmable* computer to perform the same functions. A computer equipped with the necessary input and output peripheral circuits, configured to be easily programmed by technical personnel for virtually any logical task, is called a *programmable logic controller* or *PLC*.

12.1 PLC examples

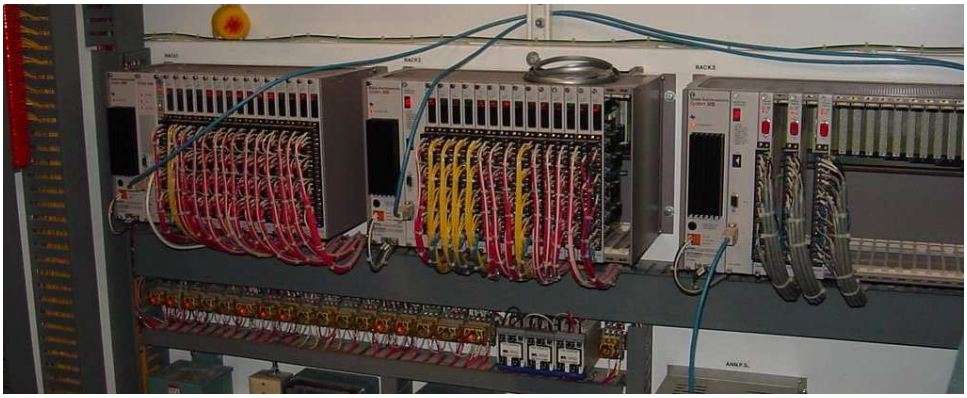
Programmable logic controllers are essentially nothing more than special-purpose, industrial computers. As such, they are built far more ruggedly than an ordinary personal computer (PC), and designed to run extremely reliable operating system software[1]. PLCs as a rule do not contain hard disk drives, cooling fans, or any other components with moving parts. This is an intentional design decision, intended to maximize the reliability of the hardware in harsh industrial environments where the PLC chassis may be subjected to temperature extremes, vibration, humidity, and airborne particulates (dust, fibers, and/or fumes).

Large PLC systems consist of a rack into which circuit "cards" are plugged. These cards include processors, input and output (I/O) points, communications ports, and other functions necessary to the operation of a complete PLC system. Such "modular" PLCs may be configured differently according to the specific needs of the application. Individual card failures are also easier to repair in a modular system, since only the failed card need be replaced, not all the cards or the whole card rack.

Small PLC systems consist of a monolithic "brick" containing all processor, I/O, and communication functions. These PLCs are typically far less expensive than their modular cousins, but are also more limited in I/O capability and must be replaced as a whole in the event of failure.

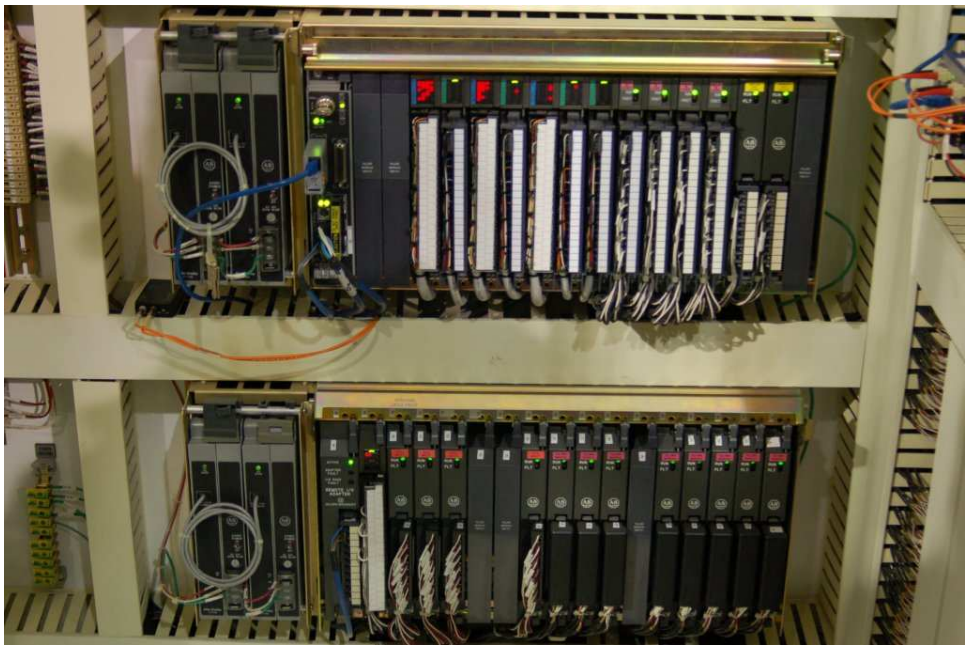
The following photographs show several examples of real PLC systems, some modular and some monolithic. These selections are not comprehensive by any means, as there are many more manufacturers and models of PLC than those I have photographed. They do, however, represent some of the more common brands and models in current (2009) industrial use.

The first photograph is of a Siemens (Texas Instruments) 505 series PLC, installed in a control panel of a municipal wastewater treatment plant. This is an example of a modular PLC, with individual processor, I/O, and communication cards plugged into a rack. Three racks appear in this photograph (two completely filled with cards, and the third only partially filled):



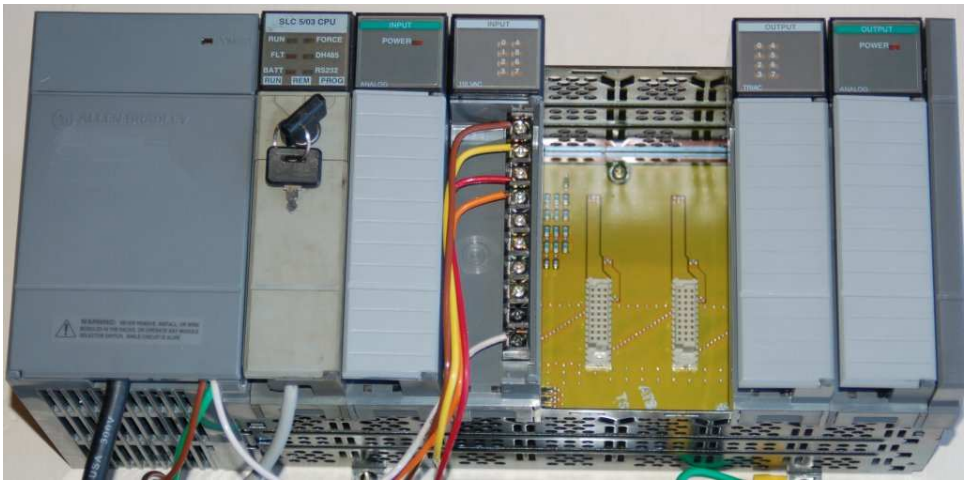
[1] There are such things as *soft PLCs*, which consist of special-purpose software running on an ordinary PC with some common operating system. Soft PLCs enjoy the high speed and immense memory capacity of modern personal computers, but do not possess the same ruggedness either in hardware construction or in operating system design. Their applications should be limited to non-critical controls where neither main process production nor safety would be jeopardized by a control system failure.

The next photograph shows an Allen-Bradley (Rockwell) PLC-5 system, used to monitor and control the operation of a large natural gas compressor. Two racks appear in this first photograph, with different types of I/O cards plugged into each rack:



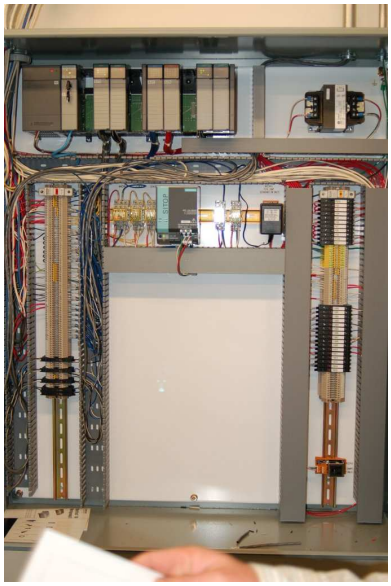
Both the Siemens (formerly Texas Instruments) 505 and Allen-Bradley (Rockwell) PLC-5 systems are considered "legacy" PLC systems by modern standards, the two systems in the previous photographs being about 20 years old each. It is not uncommon to find "obsolete" PLCs still in operation, though. Given their extremely rugged construction and reliable design, these control systems may continue to operate without significant trouble for decades.

A newer model of PLC manufactured by Allen-Bradley is the SLC 500 series (often verbally referred to as the "Slick 500" series), which is also modular in design like the older PLC-5 system, although the racks and modules of the SLC 500 design are quite compact by comparison. The SLC 500 rack shown in the next photograph has 7 "slots" for processor and I/O cards to plug in to:

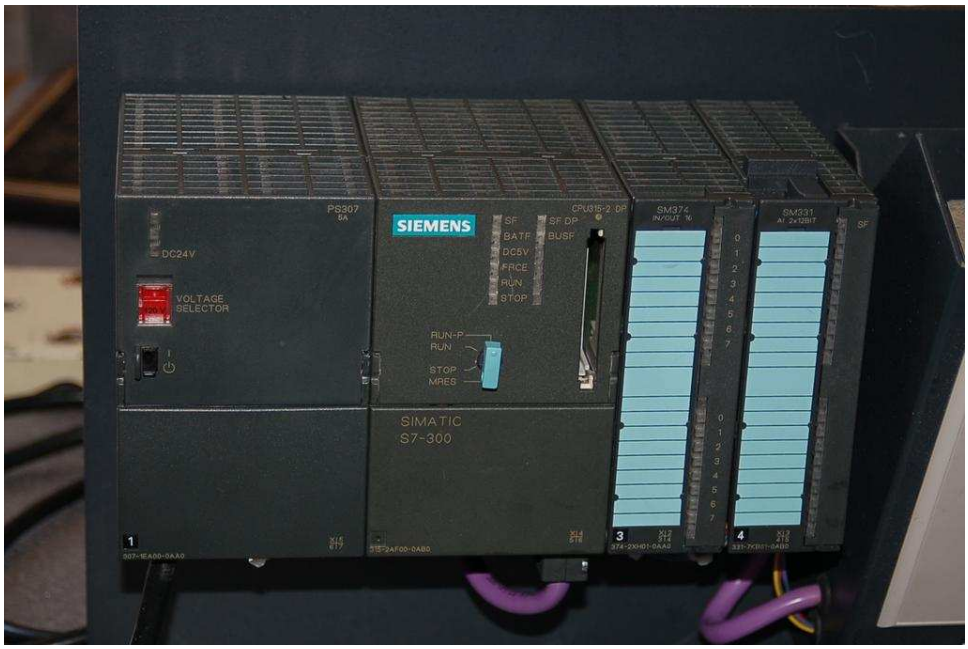


The first three slots of this SLC 500 rack acre occupied by the processor card, an analog input card, and a discrete input card, from left to right. The next two slots are empty (revealing the circuit board and connectors for accepting new cards). The last two slots hold discrete output and analog output cards, respectively.

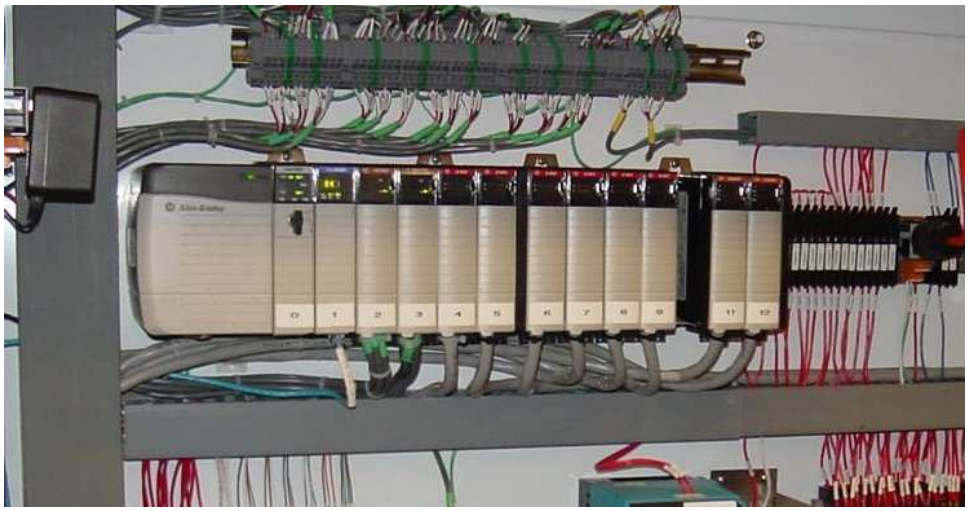
A nine-slot SLC 500 system is shown in the next photograph, controlling a high-purity water treatment system for a biopharmaceuticals manufacturing facility:



A modern PLC manufactured by Siemens appears in this next photograph, an S7-300, which is a different design of modular PLC. Instead of individual cards plugging into a rack, this modular PLC design uses individual modules plugging into each other on their sides to form a wider unit:



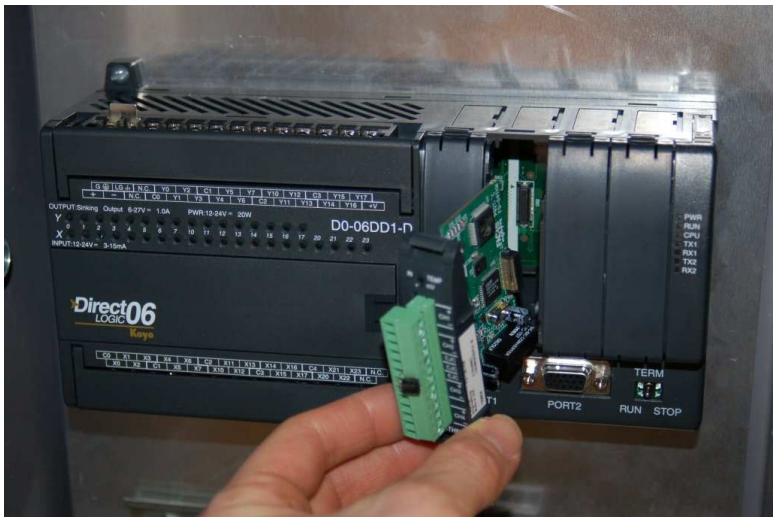
A modern PLC manufactured by Allen-Bradley (Rockwell) is this ControlLogix 5000 system, shown in this photograph used to control a cereal manufacturing process. The modular design of the ControlLogix 5000 system follows the more traditional scheme of individual cards plugged into a rack of fixed size:



While the Siemens S7 and Rockwell ControlLogix PLC platforms represent large-scale, modular PLC systems, there exist much smaller PLCs available for a fraction of the cost. Perhaps the least expensive PLC on the market at this time of writing is the Koyo "Click" PLC series, the processor module (with eight discrete input and six discrete output channels built in) shown in my hand (sold for less than 80 US dollars in the year 2009, and with free programming software!):

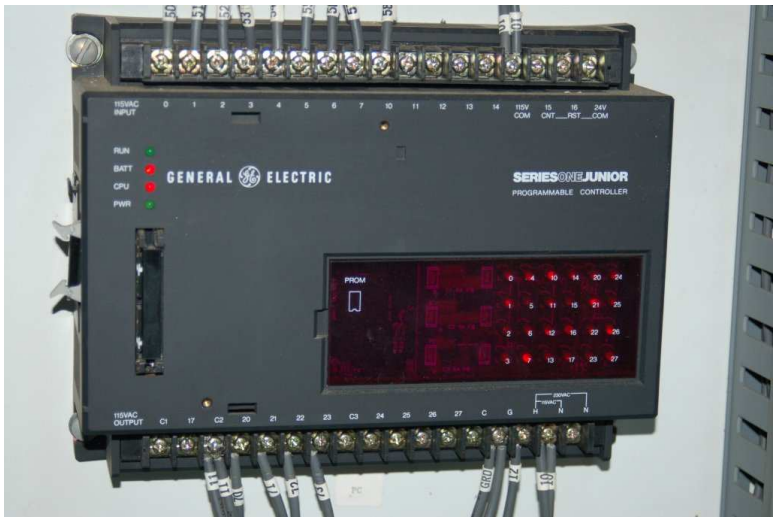
This is a semi-modular PLC design, with a minimum of input/output (I/O) channels built into the processor module, but having the capacity to accept multiple I/O modules plugged in to the side, much like the Siemens S7-300 PLC.

Other semi-modular PLCs expand using I/O cards that plug in to the base unit not unlike traditional rack-based PLC systems. The Koyo DirectLogic DL06 is a good example of this type of semi-modular PLC, the following photograph showing a model DL06 accepting a thermocouple input card in one of its four available card slots:



This photograph shows the PLC base unit with 20 discrete input channels and 16 discrete output channels, accepting an analog input card (this particular card is designed to input signals from thermocouples to measure up to four channels of temperature).

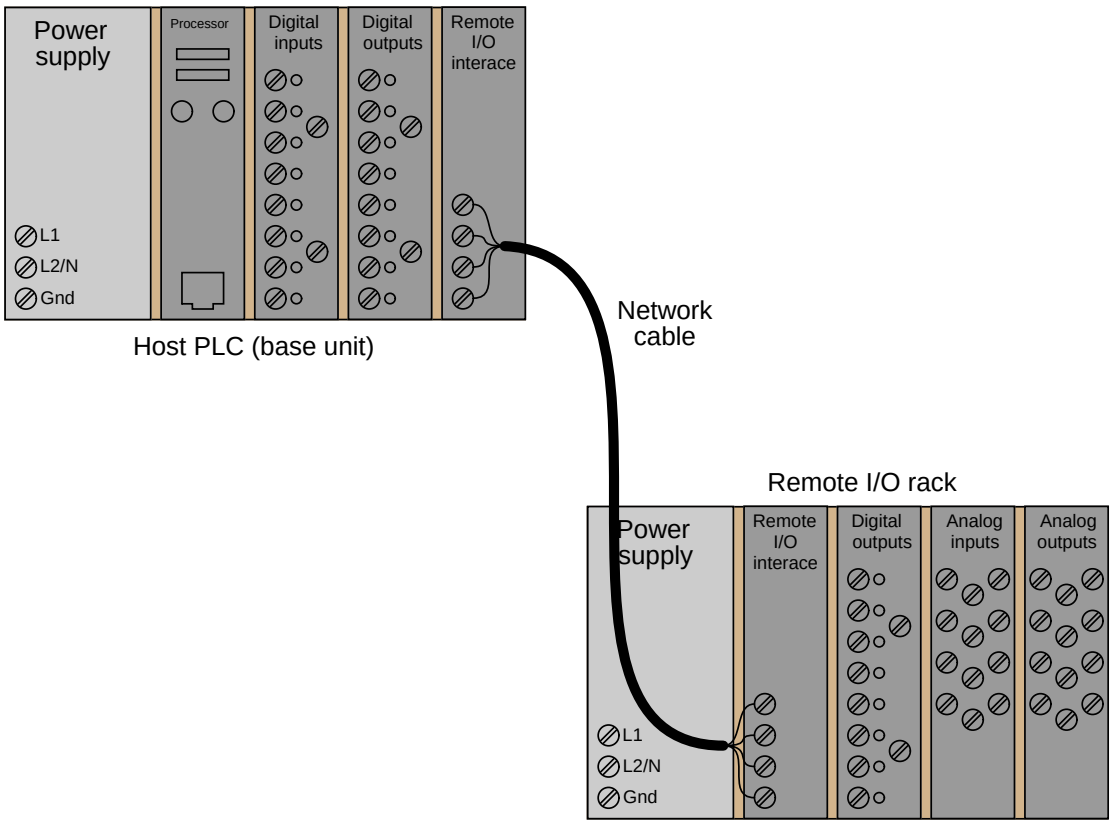
Some low-end PLCs are strictly monolithic, with no ability to accept additional I/O modules. This General Electric Series One PLC (used to monitor a small-scale hydroelectric power generating station) is an example of a purely monolithic design, having no "expansion" slots to accept I/O cards:



12.2 Input/Output (I/O) capabilities

Every programmable logic controller must have some means of receiving and interpreting signals from real-world sensors such as switches, and encoders, and also be able to effect control over real-world control elements such as solenoids, valves, and motors. This is generally known as *input/output*, or *I/O*, capability. Monolithic ("brick") PLCs have a fixed amount of I/O capability built into the unit, while modular ("rack") PLCs use individual circuit board "cards" to provide customized I/O capability.

Some PLCs have the ability to connect to processor-less remote racks filled with additional I/O cards or modules, thus providing a way to increase the number of I/O channels beyond the capacity of the base unit. The connection from host PLC to remote I/O racks usually takes the form of a special digital network, which may span a great physical distance:



Input/output capability for programmable logic controllers comes in three basic varieties: *discrete*, *analog*, and *network*; each type discussed in a following subsection.

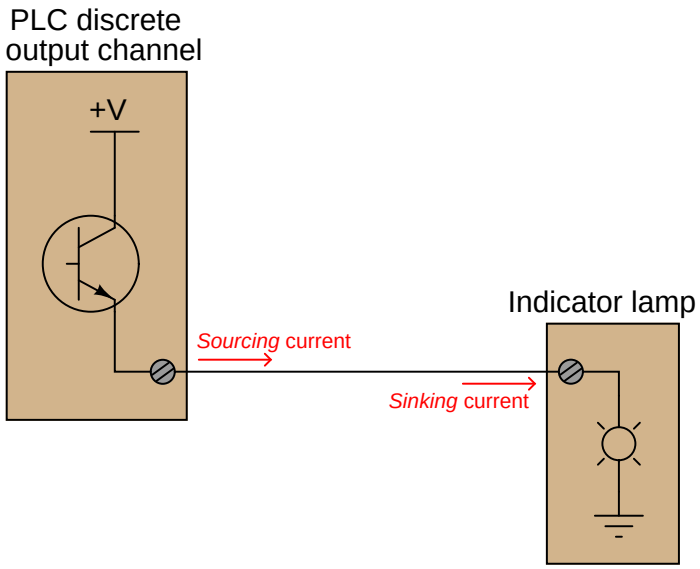
12.2.1 Discrete I/O

A "discrete" data point is one with only two states *on* and *off*. Process switches, pushbutton switches, limit switches, and proximity switches are all examples of discrete sensing devices. In order for a PLC to be aware of a discrete sensor's state, it must receive a signal from the sensor through a *discrete input* channel. Inside the discrete input module is (typically) a light-emitting diode (LED) which will be energized when the corresponding sensing device turns on. Light from this LED shines on a photo-sensitive device such as a phototransistor inside the module, which in turn activates a *bit* (a single element of digital data) inside the PLC's memory. This opto-coupled arrangement makes each input channel of a PLC rather rugged, capable of isolating the sensitive computer circuitry of the PLC from transient voltage "spikes" and other electrical phenomena capable of causing damage.

Indicator lamps, solenoid valves, and motor contactors (starters) are all examples of discrete control devices. In a manner similar to discrete inputs, a PLC connects to any number of different discrete final control devices through a *discrete output channel*. Discrete output modules typically use the same form of opto-isolation to allow the PLC's computer circuitry to send electrical power to loads: the internal PLC circuitry driving an LED which then activates some form of photosensitive switching device. Alternatively, small electromechanical relays may be used to interface the PLC's output bits to real-world electrical control devices.

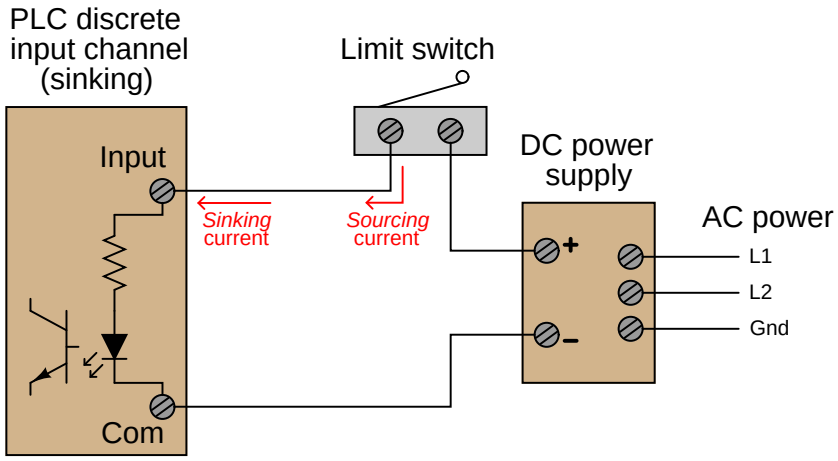
An important concept to master when working with DC discrete I/O is the distinction between *current-sourcing* and *current-sinking* devices. The terms "sourcing" and "sinking" refer to the direction of current (as denoted by conventional flow notation) into or out of a device's control wire. A device sending (conventional flow) current out of its control terminal to some other device(s) is said to be *sourcing* current, while a device accepting (conventional flow) current into its control terminal is said to be *sinking* current.

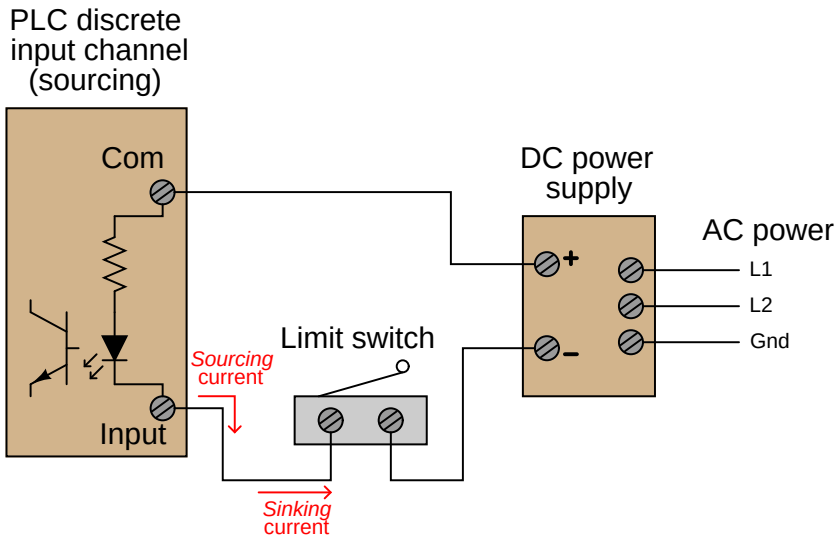
To illustrate, the following illustration shows a PLC output channel is *sourcing* current to an indicator lamp, which is *sinking* current to ground:



These terms really only make sense when electric current is viewed from the perspective of conventional flow, where the positive terminal of the DC power supply is envisioned to be the "source" of the current, with current finding its way "down" to ground (the negative terminal of the DC power supply). In every circuit formed by the output channel of a PLC driving a discrete control device, or by a discrete sensing device driving an input channel on a PLC, one element in the circuit must be sourcing current while the other is sinking current.

If the discrete device connecting to the PLC is not polarity-sensitive, either type of PLC I/O module will suffice. For example, the following diagrams show a mechanical limit switch connecting to a sinking PLC input and to a sourcing PLC input:

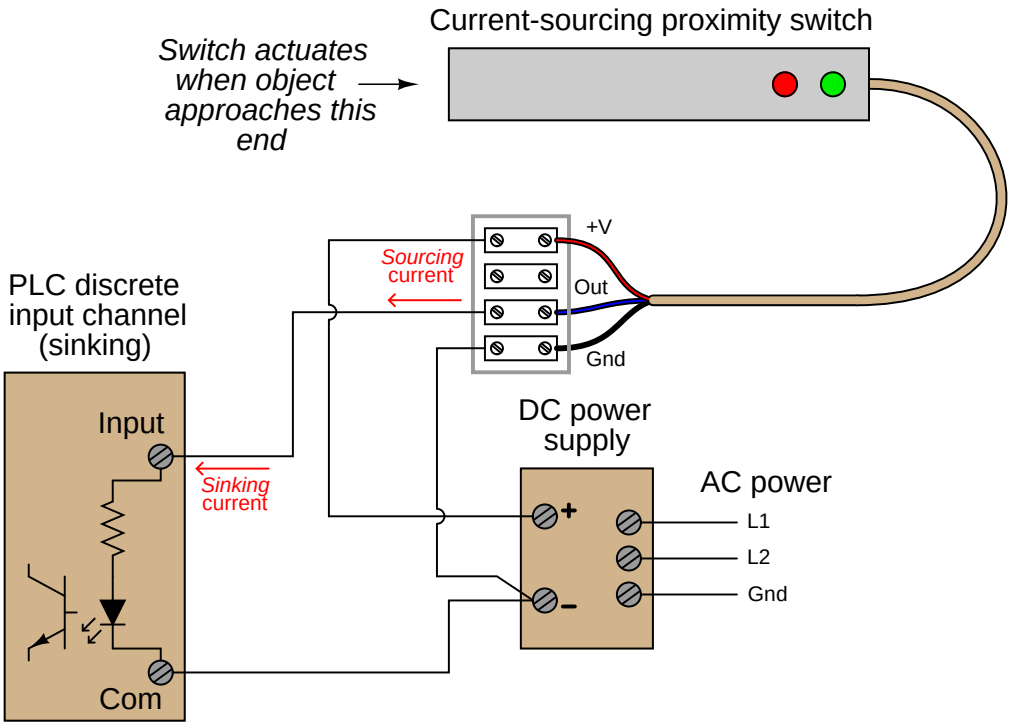




You will notice that the only difference in the two PLC input channels is the labeling of the terminals. In the "sinking" input channel, the input terminal was positive while the common ("Com") terminal was negative. These labels are reversed in the "sourcing" input channel[2].

[2]Some sourcing input modules may use a label other than "Com" to denote the positive terminal common to all input channels. DC input modules for the Allen-Bradley SLC 500 line of PLCs, for example, label the common terminal on DC sourcing inputs as "VDC" instead of "DC Com" to help avoid confusion. This way, the technician need only remember that "DC Com" always refers to the negative pole of the DC power source, while "VDC" always refers to the positive pole.

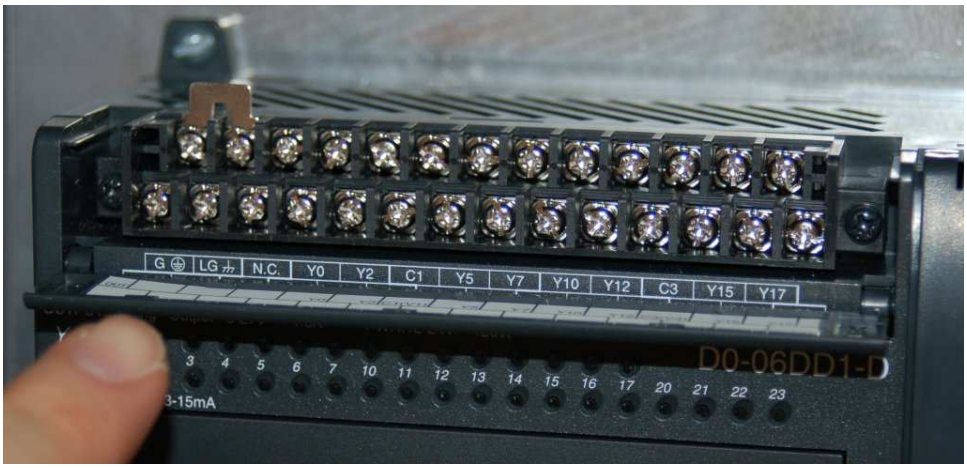
Some discrete sensing devices *are* polarity-sensitive, such as electronic proximity sensors containing transistor outputs. A "sourcing" proximity switch can only interface with a "sinking" PLC input channel, and visa-versa:



Two photographs of a DC (sinking) discrete input module for an Allen-Bradley model SLC 500 PLC are shown here: one with the plastic cover closed over the connection terminals, and the other with the plastic cover opened up for viewing the terminals. A legend on the inside of the cover shows the purpose of each screw terminal: eight input channels (numbered 0 through 7) and two redundant "DC Com" terminals for the negative pole of the DC power supply to connect:

A standard feature found on practically every PLC discrete I/O module is a set of LED indicators visually indicating the status of each bit (discrete channel). On the SLC 500 module, the LEDs appear as a cluster of eight numbered squares near the top of the module.

A photograph of discrete output terminals on another brand of PLC (a Koyo model DL06) shows somewhat different labeling:



Here, each output channel terminal is designated with a letter/number code beginning with the letter "Y". Several "common" terminals labeled with "C" codes service clusters of output channels. In this particular case, each "common" terminal is common only to four output channels. With sixteen total output channels on this PLC, this means there are four different "common" terminals. While this may seem somewhat strange (why not just have one "common" terminal for all sixteen output channels?), it more readily permits different DC power supplies to service different sets of output channels.

Electrical polarity is not an issue with AC discrete I/O, since the polarity of AC reverses periodically anyway. However, there is still the matter of whether the "common" terminal on a discrete PLC module will connect to the *neutral* (grounded) or *hot* (ungrounded) AC power conductor.

The next photograph shows a discrete AC output module for an Allen-Bradley SLC 500 PLC, using TRIACs as power switching devices rather than transistors as is customary with DC discrete output modules:

This particular eight-channel module provides two sets of TRIACs for switching power to AC loads, each set of four TRIACs receiving AC power from a "hot" terminal (VAC 1 or VAC 2), the other side of the load device being connected to the "neutral" (grounded) conductor of the AC power source.

Fortunately, the hardware reference manual supplied by the manufacturer of every PLC shows diagrams illustrating how to connect discrete input and output channels to field devices. One should always consult these diagrams before connecting devices to the I/O points of a PLC!

12.2.2 Analog I/O

In the early days of programmable logic controllers, processor speed and memory were too limited to support anything but discrete (on/off) control functions. Consequently, the only I/O capability found on early PLCs were discrete in nature[3]. Modern PLC technology, though, is powerful enough to support the measurement, processing, and output of analog (continuously variable) signals.

All PLCs are digital devices at heart. Thus, in order to interface with an analog sensor or control device, some "translation" is necessary between the analog and digital worlds. Inside every analog input module is an *ADC*, or *Analog-to-Digital Converter*, circuit designed to convert an analog electrical signal into a multi-bit binary word. Conversely, every analog output module contains a *DAC*, or *Digital-to-Analog Converter*, circuit to convert the PLC's digital command words into analog electrical quantities.

Analog I/O is commonly available for modular PLCs for many different analog signal types, including:

- Voltage (0 to 10 volt, 0 to 5 volt)
- Current (0 to 20 mA, 4 to 20 mA)
- Thermocouple (millivoltage)
- RTD (millivoltage)
- Strain gauge (millivoltage)

[3]Some modern PLCs such as the Koyo "Click" are also discrete-only. Analog I/O and processing is significantly more complex to engineer and more expensive to manufacture than discrete control, and so low-end PLCs are more likely to lack analog capability.

The following photographs show two analog I/O cards for an Allen-Bradley SLC 500 modular PLC system, an analog input card and an analog output card. Labels on the terminal cover doors indicate screw terminal assignments:

12.2.3 Network I/O

Many different digital network standards exist for PLCs to communicate with, from PLC to PLC and between PLCs and field devices. One of the earliest digital protocols developed for PLC communication was *Modbus*, originally for the Modicon brand of PLC. Modbus was adopted by other PLC and industrial device manufacturers as a *de facto* standard[4], and remains perhaps the most universal digital protocol available for industrial digital devices today.

Another digital network standard developed by a particular manufacturer and later adopted as a *de facto* standard is *Profibus*, originally developed by Siemens.

For more information on networking standards used in PLC systems, refer to the "Digital electronic instrumentation" chapter, specifically sections on specific network standards such as Modbus and Profibus.

[4] A "de facto" standard is one arising naturally out of legacy rather than by an premeditated agreement between parties. Modbus and Profibus networks are considered "de facto" standards because those networks were designed, built, and marketed by pioneering firms prior to their acceptance as standards for others to conform to. In Latin, *de facto* means "from the fact," which in this case refers to the fact of pre-existence: a standard agreed upon to conform to something already in popular use. By contrast, a standard intentionally agreed upon before its physical realization is a *de jure* standard (Latin for "from the law"). FOUNDATION Fieldbus is an example of a *de jure* standard, where a committee arrives at a consensus for a network design and specifications prior to that network being built and marketed by any firm.

12.3 Logic programming

Although it seems each model of PLC has its own idiosyncratic standard for programming, there does exist an international standard for controller programming that most PLC manufacturers at least attempt to conform to. This is the IEC 61131-3 standard, which will be the standard presented in this chapter.

One should take solace in the fact that despite differences in the details of PLC programming from one manufacturer to another and from one model to another, the basic principles are largely the same. There exist much greater disparities between different general-purpose programming languages (e.g. C/C++, BASIC, FORTRAN, Pascal, Java, Ada, etc.) than between the programming languages supported by different PLCs, and this fact does not prevent computer programmers from being "multilingual." I have personally written and/or analyzed programs for over a half-dozen different manufacturers of PLCs (Allen-Bradley, Siemens, Square D, Koyo, Fanuc, Moore Products APACS and QUADLOG, and Modicon), with multiple PLC models within most of those brands, and I can tell you the differences in programming conventions are insignificant. After learning how to program one model of PLC, it is quite easy to adapt to programming other makes and models of PLC. If you are learning to program a particular PLC that does not exactly conform to the IEC 61131-3 standard, you will still be able to apply every single principle discussed in this chapter – the fundamental concepts are truly that universal.

The IEC 61131-3 standard specifies five distinct forms of programming language for industrial controllers:

- Ladder Diagram (LD)
- Structured Text (ST)
- Instruction List (IL)
- Function Block Diagram (FBD)
- Sequential Function Chart (SFC)

Not all programmable logic controllers support all five language types, but nearly all of them support Ladder Diagram (LD), which will be the primary focus of this book.

Programming languages for many industrial devices are limited by design. One reason for this is *simplicity*: any programming language simple enough in structure for someone with no formal computer programming knowledge to understand is going to be limited in its capabilities. Another reason for programming limitations is *safety*: the more flexible and unbounded a programming language is, the more potential there will be for complicated "run-time" errors that may be very difficult to troubleshoot. The ISA safety standard number 84 classifies industrial programming languages as either *Fixed Programming Languages* (FPL), *Limited Variability Languages* (LVL), or *Full Variability Languages* (FVL). Ladder Diagram and Function Block Diagram programming are both considered to be "limited variability" languages, whereas Instruction List (and traditional computer programming languages such as C/C++, FORTRAN, BASIC, etc.) are considered "full variability" languages with all the attendant potential for complex errors.

12.3.1 Memory maps and I/O addressing

Every discrete input and output channel on a PLC is linked to a corresponding *bit* in the PLC's digital memory. Similarly, every analog input and output point on a PLC is linked to a corresponding *word* (a block of bits) in the PLC's memory. The association between I/O points and memory locations differs between different manufacturers and models of PLC, sometimes radically. Learning how the memory of any particular computer is organized – for I/O as well as for other purposes – is one of the first things any programmer should become familiar with when learning how to work with any computer system, PLCs included.

Every PLC manufacturer publishes reference manuals providing this information, usually in the form of a chart known as a *memory map*. A "memory map" shows the organization of digital memory into different sections, reserved for different uses. The following table shows a partial memory map for an Allen-Bradley SLC 500 PLC[5]:

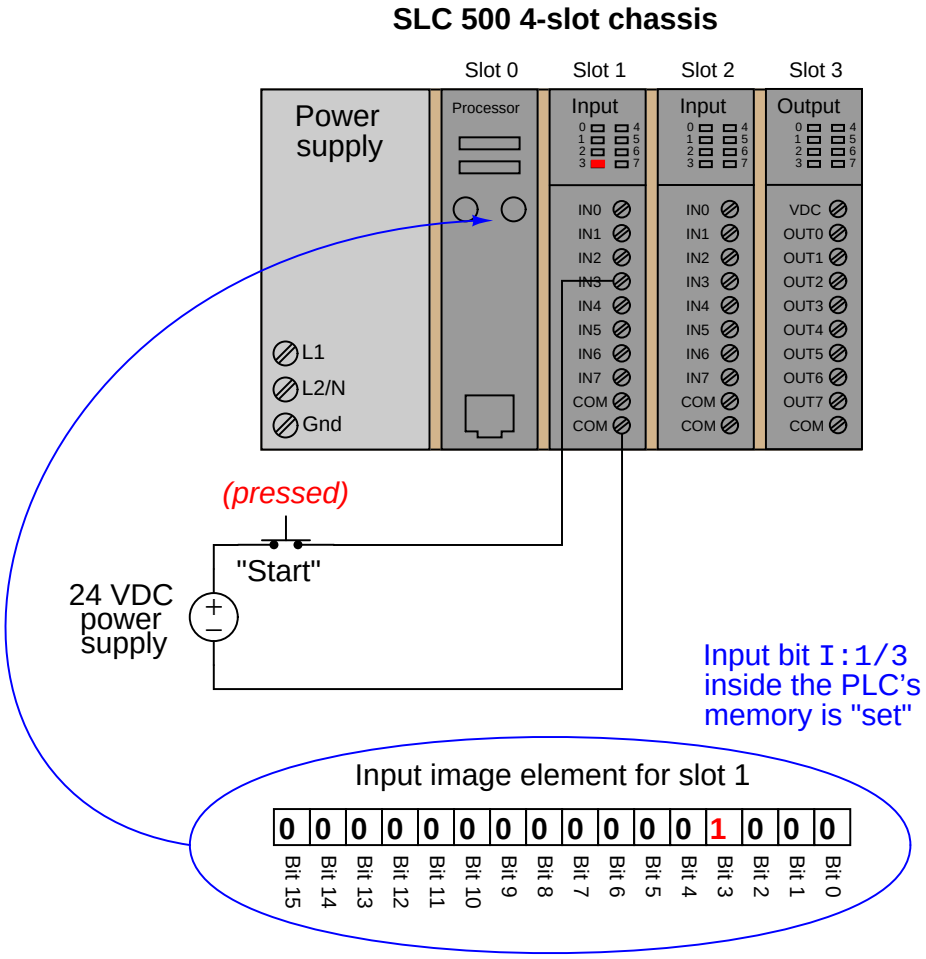
File number	Type	Logical address range
0	Output image	O:0 to O:30
1	Input image	I:0 to I:30
2	Status	S:0 to S:n
3	Binary	B3:0 to B3:255
4	Timers	T4:0 to T4:255
5	Counters	C5:0 to C5:255
6	Control	R6:0 to R6:255
7	Integer	N7:0 to N7:255
8	Floating-point	F8:0 to F8:255
9	Network	x9:0 to x9:255
10	User defined	x10:0 to x255:255

Note that Allen-Bradley's use of the word "file" differs from modern parlance. In the SLC 500 controller, a "file" is a partition of random-access memory used to store a particular type of data. By contrast, a "file" according to modern convention is a contiguous collection of data bits with collective meaning (e.g. a word processing file or a spreadsheet file). Within each of the Allen-Bradley PLC's "files" are multiple "elements," each element consisting of a set of bits (8, 16, 24, or 32) representing data. Elements are addressed by number following the colon after the file designator, and individual bits within each element addressed by a number following a slash mark. For example, the first bit (bit 0) of the second element in file 3 (Binary) would be addressed as `B3:2/0`.

A hallmark of the SLC 500's addressing scheme common to many legacy PLC systems is that the address labels for input and output bits explicitly reference the physical locations of the I/O channels. For instance, if an 8-channel discrete input card were plugged into slot 4 of an Allen-Bradley SLC 500 PLC, and you wished to specify the second bit (bit 1 out of a 0 to 7 range), you would address it with the following label: `I:4/1`. Addressing the seventh bit (bit number 6) on a discrete output card plugged into slot 3 would require the label `O:3/6`. In either case, the numerical structure of that label tells you exactly where the real-world input signal connects to the PLC. If an input or output card possesses more than 16 bits – as in the case of 24- or 36-bit discrete I/O cards – the addressing scheme further subdivides each element into *words* and bits (each "word" being 16 bits in length). Thus, the address for bit number 20 of a 32-bit input module plugged into slot 7 would be `I:7.2/4` (since bit 20 is equivalent to bit 4 of word 2 – word 1 addressing bits 0 through 15 and word 2 addressing bits 16 through 31).

[5] Called the *data table*, this map shows the addressing of memory areas reserved for programs entered by the user. Other areas of memory exist within the SLC 500 processor, but these other areas are inaccessible to the technician writing PLC programs.

To illustrate the relationship between physical I/O and bits in the PLC's memory, consider this example of an Allen-Bradley SLC 500 PLC, showing one of its discrete input channels energized (the switch being used as a "Start" switch for a mixer motor):



Legacy PLC systems typically reference each one of the I/O channels by labels such as "I:1/3" (or equivalent[6]) indicating the actual location of the input channel terminal on the PLC unit. The IEC 61131-3 programming standard refers to this channel-based addressing of I/O data points as *direct addressing*. A synonym for direct addressing is *absolute addressing*.

Addressing I/O bits directly by their card, slot, and/or terminal labels may seem simple and elegant, but it becomes very cumbersome for large PLC systems and complex programs. Every time

[6]Some systems such as the Texas Instruments 505 series used "X" labels to indicate discrete input channels and "Y" labels to indicate discrete output channels (e.g. input `X9` and output `Y14`). This same labeling convention is still used by Koyo in its DirectLogic and "Click" PLC models. Siemens continues a similar tradition of I/O addressing by using the letter "I" to indicate discrete inputs and the letter "Q" to indicate discrete outputs (e.g. input channel `I0.5` and output `Q4.1`).

a technician or programmer views the program, they must "translate" each of these I/O labels to some real-world device (e.g. "Input `I:1/3` is actually the *Start* pushbutton for the middle tank mixer motor") in order to understand the function of that bit. A later effort to enhance the clarity of PLC programming was the concept of addressing variables in a PLC's memory by arbitrary names rather than fixed codes. The IEC 61131-3 programming standard refers to this as *symbolic addressing* in contrast to "direct" (channel-based) addressing, allowing programmers arbitrarily name I/O channels in ways that are meaningful to the system as a whole. To use our simple motor "Start" switch example, it is now possible for the programmer to designate input `I:1/3` (an example of a *direct address*) as "`Motor_start_switch`" (an example of a *symbolic address*) within the program, thus greatly enhancing the readability of the PLC program. Initial implementations of this concept maintained direct addresses for I/O data points, with symbolic names appearing in addition to those absolute addresses as mere supplements for program readability. One PLC manufacturer (Koyo) still does this, calling the symbolic labels *nicknames*.

The modern trend in PLC addressing is to avoid the use of direct addresses such as `I:1/3` altogether, so they do not appear anywhere in the programming code. The Allen-Bradley "Logix" series of programmable logic controllers is the most prominent example of this new convention at the time of this writing. Each I/O point, regardless of type or physical location, is assigned a *tag name* which is meaningful in a real-world sense, and these tag names (or *symbols* as they are alternatively called) are referenced to absolute I/O channel locations by a database file. An important requirement of tag names is that they contain no space characters between words (e.g. instead of "`Motor start switch`", a tag name should use hyphens or underscore marks as spacing characters: "`Motor_start_switch`"), since spaces are generally assumed by computer programming languages to be delimiters (separators between different variables).

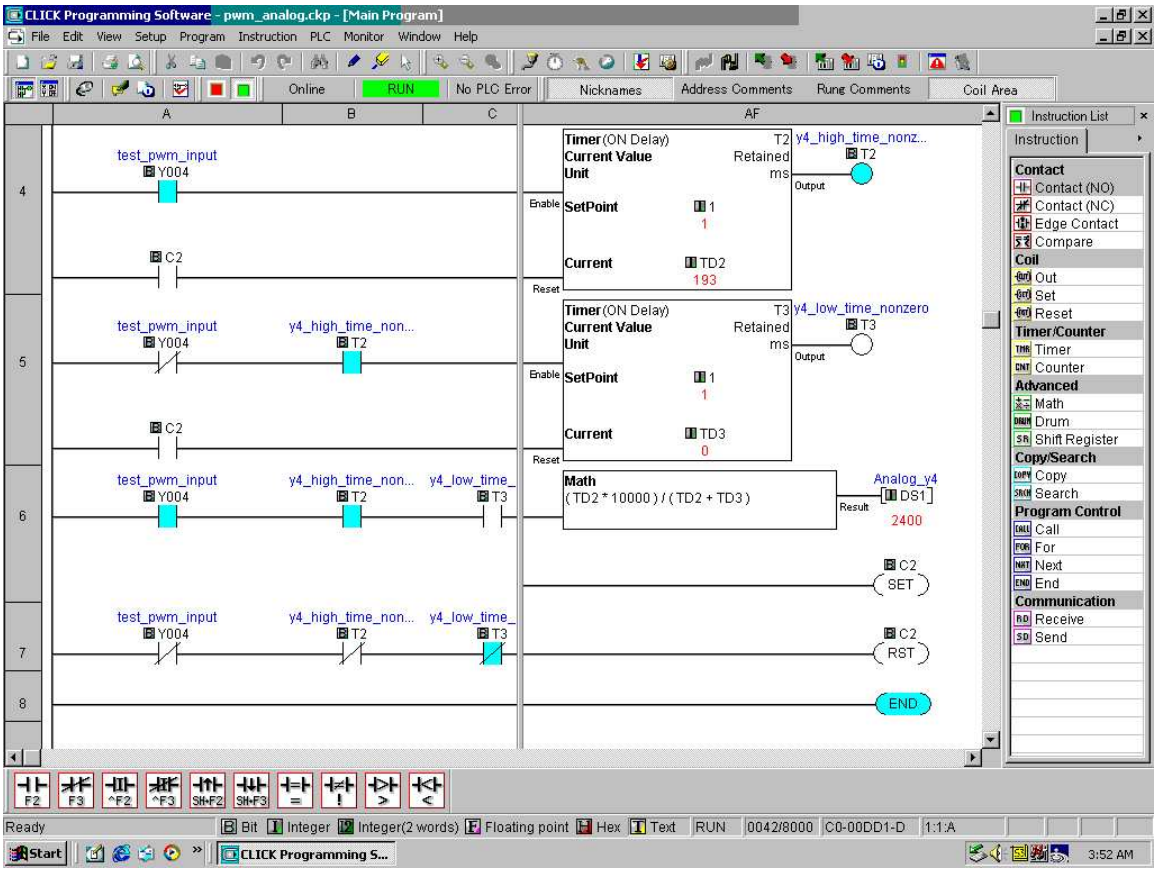
I will hold to this modern convention of symbolic addressing throughout my presentations on PLC programming. Each data point within my PLC programs will bear its own tag name rather than a direct (channel-based) address code.

12.3.2 Ladder Diagram (LD)

In the United States, the most common language used to program PLCs is *Ladder Diagram* (LD), also known as *Relay Ladder Logic* (RLL). This is a graphical language showing the logical relationships between inputs and outputs as though they were contacts and coils in a hard-wired electromechanical relay circuit. This language was invented for the express purpose of making PLC programming feel "natural" to electricians familiar with relay-based logic and control circuits. While ladder diagram programming has many shortcomings, it remains extremely popular and so will be the primary focus of this chapter.

Every ladder diagram program is arranged to resemble an electrical diagram, making this a graphical (rather than text-based) programming language. Ladder diagrams are to be thought of as *virtual circuits*, where virtual "power" flows through virtual "contacts" (when closed) to energize virtual "relay coils" to perform logical functions. None of the contacts or coils seen in a ladder diagram PLC program are real; rather, they are representations of bits in the PLC's memory, the logical inter-relationships between those bits expressed in the form of a diagram *resembling* a circuit.

The following computer screenshot shows a typical ladder diagram program[7] being edited on a personal computer:



Contacts appear just as they would in an electrical relay logic diagram – as short vertical line segments separated by a horizontal space. Normally-open contacts are empty within the space between the line segments, while normally-closed contacts have a diagonal line crossing through that space. Coils are somewhat different, appearing as either circles or pairs of parentheses. Other instructions appear as rectangular boxes.

Each horizontal line is referred to as a *rung*, just as each horizontal step on a stepladder is called a "rung." A common feature among ladder diagram program editors, as seen on this screenshot, is

[7] This particular program and editor is for the Koyo "Click" series of micro-PLCs.

the ability to color-highlight those virtual "components" in the virtual "circuit" ready to "conduct" virtual "power." In this particular editor, the color used to indicate "conduction" is light blue. Another form of status indication seen in this PLC program is the values of certain variables in the PLC's memory, shown in red text.

For example, you can see coil `T2` energized at the upper-right corner of the screen (filled with light blue coloring), while coil `T3` is not. Correspondingly, each normally-open `T2` contact appears colored, indicating its "closed" status, while each normally-closed `T2` contact is uncolored. By contrast, each normally-open `T3` contact is uncolored (since coil `T3` is unpowered) while each normally-closed `T3` contact is shown colored to indicate its conductive status. Likewise, the current count values of timers `T2` and `T3` are shown as 193 and 0, respectively. The output value of the math instruction box happens to be 2400, also shown in red text.

Color-highlighting of ladder diagram components only works, of course, when the computer running the program editing software is connected to the PLC and the PLC is in the "run" mode (and the "show status" feature of the editing software is enabled). Otherwise, the ladder diagram is nothing more than black symbols on a white background. Not only is status highlighting very useful in de-bugging PLC programs, but it also serves an invaluable diagnostic purpose when a technician analyzes a PLC program to check the status of real-world input and output devices connected to the PLC. This is especially true when the program's status is viewed remotely over a computer network, allowing maintenance staff to investigate system problems without even being near the PLC!

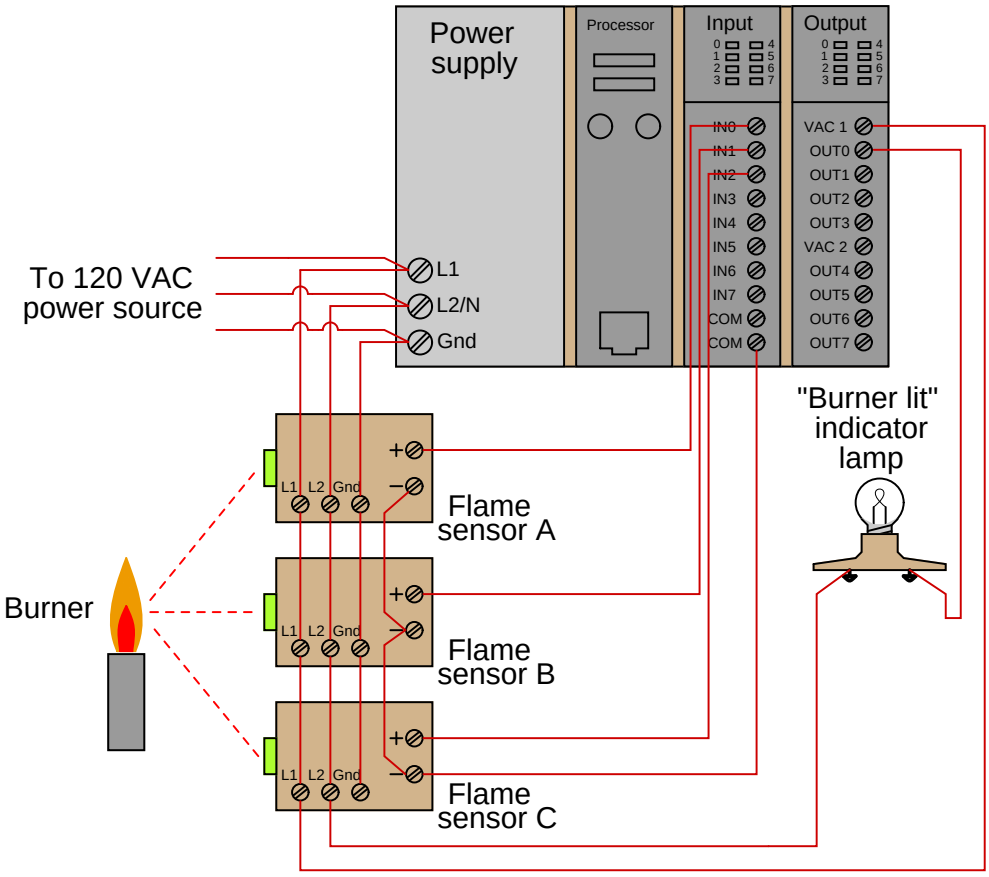
Contacts and coils

The most elementary objects in ladder diagram programming are *contacts* and *coils*, intended to mimic the contacts and coils of electromechanical relays. Contacts and coils are *discrete* programming elements, dealing with Boolean (1 and 0; on and off; true and false) variable states. Each contact in a ladder diagram PLC program represents the *reading* of a single bit in memory, while each coil represents the *writing* of a single bit in memory.

Discrete input signals to the PLC from real-world switches are read by a ladder diagram program by contacts referenced to those input channels. In legacy PLC systems, each discrete input channel has a specific address which must be applied to the contact(s) within that program. In modern PLC systems, each discrete input channel has a tag name created by the programmer which is applied to the contact(s) within the program. Similarly, discrete output channels – referenced by coil symbols in the ladder diagram – must also bear some form of address or tag name label.

To illustrate, we will imagine the construction and programming of a redundant flame-sensing system to monitor the status of a burner flame using three sensors. The purpose of this system will be to indicate a "lit" burner if at least two out of the three sensors indicate flame. If only one sensor indicates flame (or if no sensors indicate flame), the system will declare the burner to be un-lit. The burner's status will be visibly indicated by a lamp that human operators can readily see inside the control room area.

Our system's wiring is shown in the following diagram:

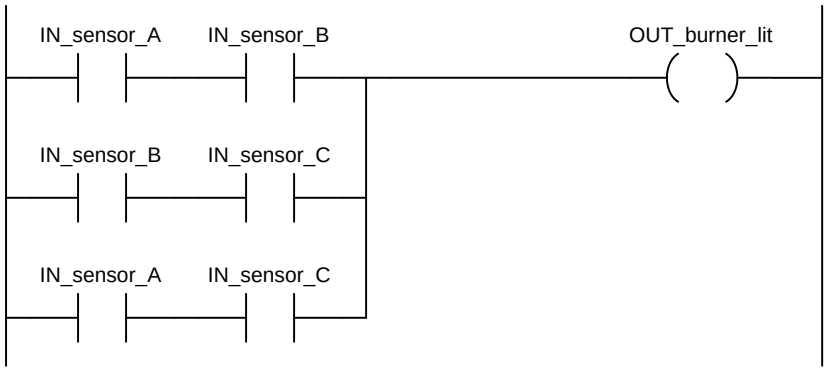


Each flame sensor outputs a DC voltage signal indicating the detection of flame at the burner, either on (24 volts DC) or off (0 volts DC). These three discrete DC voltage signals are sensed by the first three channels of the PLC's discrete input card. The indicator lamp is a 120 volt light bulb, and so must be powered by an AC discrete output card, shown here in the PLC's last slot.

To make the ladder program more readable, we will assign tag names (symbolic addresses) to each input and output bit in the PLC, describing its real-world device in an easily-interpreted format[8]. We will tag the first three discrete input channels as `IN_sensor_A`, `IN_sensor_B`, and `IN_sensor_C`, and the output as `OUT_burner_lit`.

[8] If this were a legacy Allen-Bradley PLC system using absolute addressing, we would be forced to address the three sensor inputs as `I:1/0`, `I:1/1`, and `I:1/2` (slot 1, channels 0 through 2), and the indicator lamp output as `O:2/0` (slot 2, channel 0). If this were a newer Logix5000 Allen-Bradley PLC, the default tag names would be `Local:1:I.Data.0`, `Local:1:I.Data.1`, and `Local:1:I.Data.2` for the three inputs, and `Local:2:O.Data.0` for the output. However, in either system we have the ability to assign symbolic addresses so we have a way to reference the I/O channels without having to rely on these cumbersome labels. The programs showing in this book exclusively use tag names rather than absolute addresses, since this is the more modern programming convention.

A ladder program to determine if at least two out of the three sensors detect flame is shown here, with the tag names referencing each contact and coil:



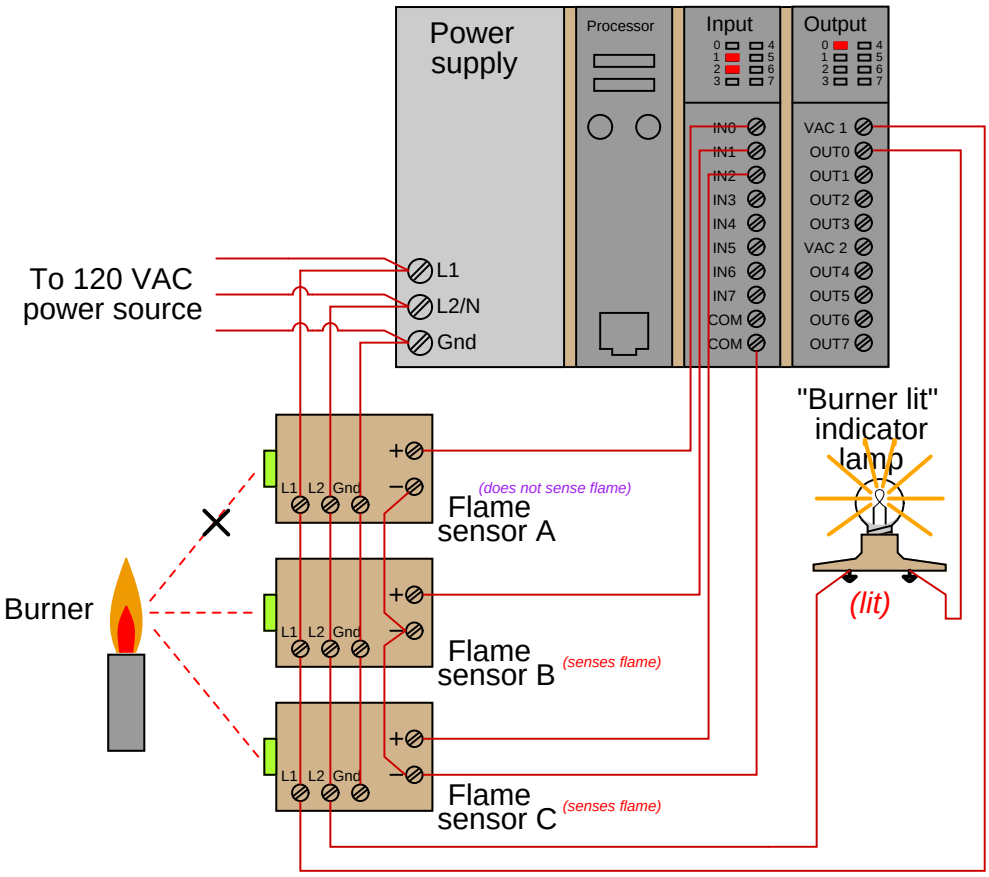
Series-connected contacts in a ladder diagram perform the logical `AND` function, while parallel contacts perform the logical `OR` function. Thus, this two-out-of-three flame-sensing program could be verbally described as:

"Burner is lit if either `A` *and* `B`, *or* either `B` *and* `C`, `or` `A` *and* `C`"

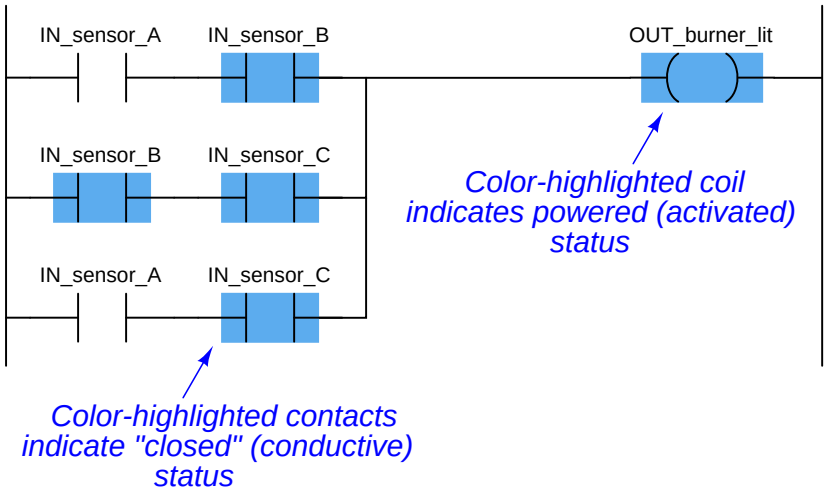
An alternate way to express this is to use the notation of *Boolean algebra*, where multiplication represents the `AND` function and addition represents the `OR` function:

$$\text{Burner\_lit} = AB + BC + AC$$

To illustrate how this program would work, we will consider a case where flame sensors B and C detect flame, but sensor A does not. This represents a two-out-of-three condition, and so we would expect the PLC to turn on the "Burner lit" indicator light as programmed. From the perspective of the PLC's rack, we would see the indicator LEDs for sensors B and C lit up, as well as the indicator LED for the lamp's output channel:

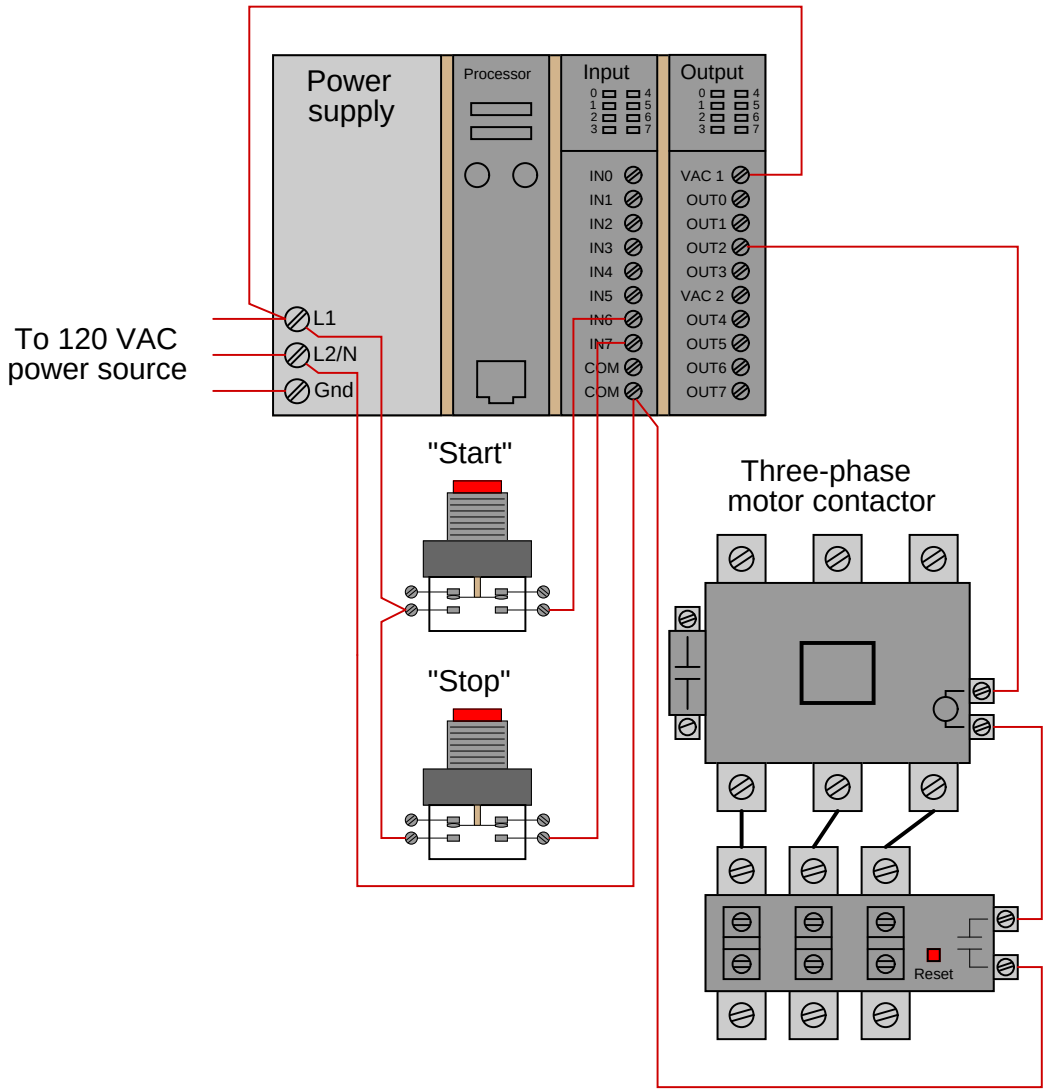


Examining the ladder diagram program with status indication enabled, we would see how just one of the series-connected contact pairs are passing "power" to the output coil:



Note that the color highlighting does *not* indicate a virtual contact is *conducting* virtual power, but merely that it is *able* to conduct power. Color highlighting around a virtual coil, however, *does* indicate the presence of virtual "power" at that coil.

Contacts and relays are not just useful for implementing simple logic functions, but they may also perform *latching* functions as well. A very common application of this in industrial PLC systems is a latching start/stop program for controlling electric motors by means of momentary-contact pushbutton switches. As before, this functionality will be illustrated by means of an hypothetical example circuit and program:

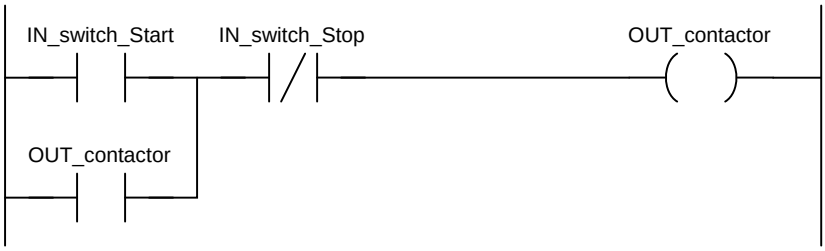


In this system, two pushbutton switches are connected to discrete inputs on a PLC, and the PLC in turn energizes the coil of a motor contactor relay by means of one of its discrete outputs[9].

[9]The particular input and output channels chosen for this example are completely arbitrary. There is no particular

An overload contact is wired directly in series with the contactor coil to provide motor overcurrent protection, even in the event of a PLC failure where the discrete output channel remains energized[10].

The ladder program for this motor control system would look like this:



Pressing the "Start" pushbutton energizes discrete input channel 6 on the PLC, which "closes" the virtual contact in the PLC program labeled `IN_switch_Start`. The normally-closed virtual contact for input channel 7 (the "Stop" pushbutton) is already closed by default when the "Stop" button is not being pressed, and so the virtual coil will receive "power" when the "Start" pushbutton is pressed and the "Stop" pushbutton is not.

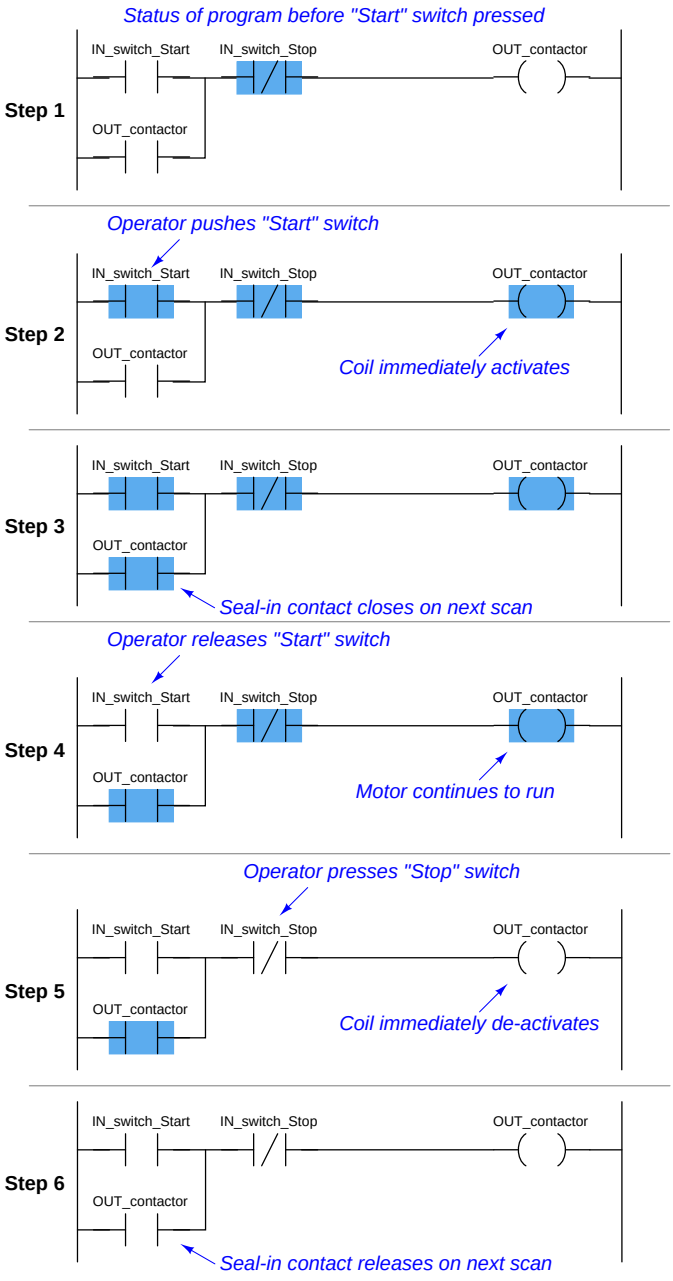
Note the *seal-in* contact bearing the exact same label as the coil: `OUT_contactor`. At first it may seem strange to have both a contact and a coil in a PLC program labeled identically, since contacts are most commonly associated with inputs and coils with outputs, but this makes perfect sense if you realize the true meaning of contacts and coils in a PLC program: as *read* and *write* operations on bits in the PLC's memory. The coil labeled `OUT_contactor` *writes* the status of that bit, while the contact labeled `OUT_contactor` *reads* the status of that same bit. The purpose of this contact, of course, is to *latch* the motor in the "on" state after a human operator has released his or her finger from the "Start" pushbutton.

This programming technique is known as *feedback*, where an output variable of a function (in this case, the feedback variable is `OUT_contactor`) is also an input to that same function. The path of feedback is *implicit* rather than *explicit* in ladder diagram programming, with the only indication of feedback being the common name shared by coil and contact. Other graphical programming languages (such as Function Block) have the ability to show feedback paths as connecting lines between function outputs and inputs, but this capacity does not exist in ladder diagram.

reason to choose input channels 6 and 7, or output channel 2, as I have shown in the wiring diagram. Any available I/O channels will serve the purpose quite adequately.

[10] While it is possible to wire the overload contact to one of the PLC's discrete input channels and then program a *virtual* overload contact in series with the output coil to stop the motor in the event of a thermal overload, this strategy would rely on the PLC to perform a safety function which is probably better performed by hard-wired circuitry.

A step-by-step sequence showing the operation and status of this simple program illustrates how the seal-in contact functions, through a start-up and shut-down cycle of the motor:



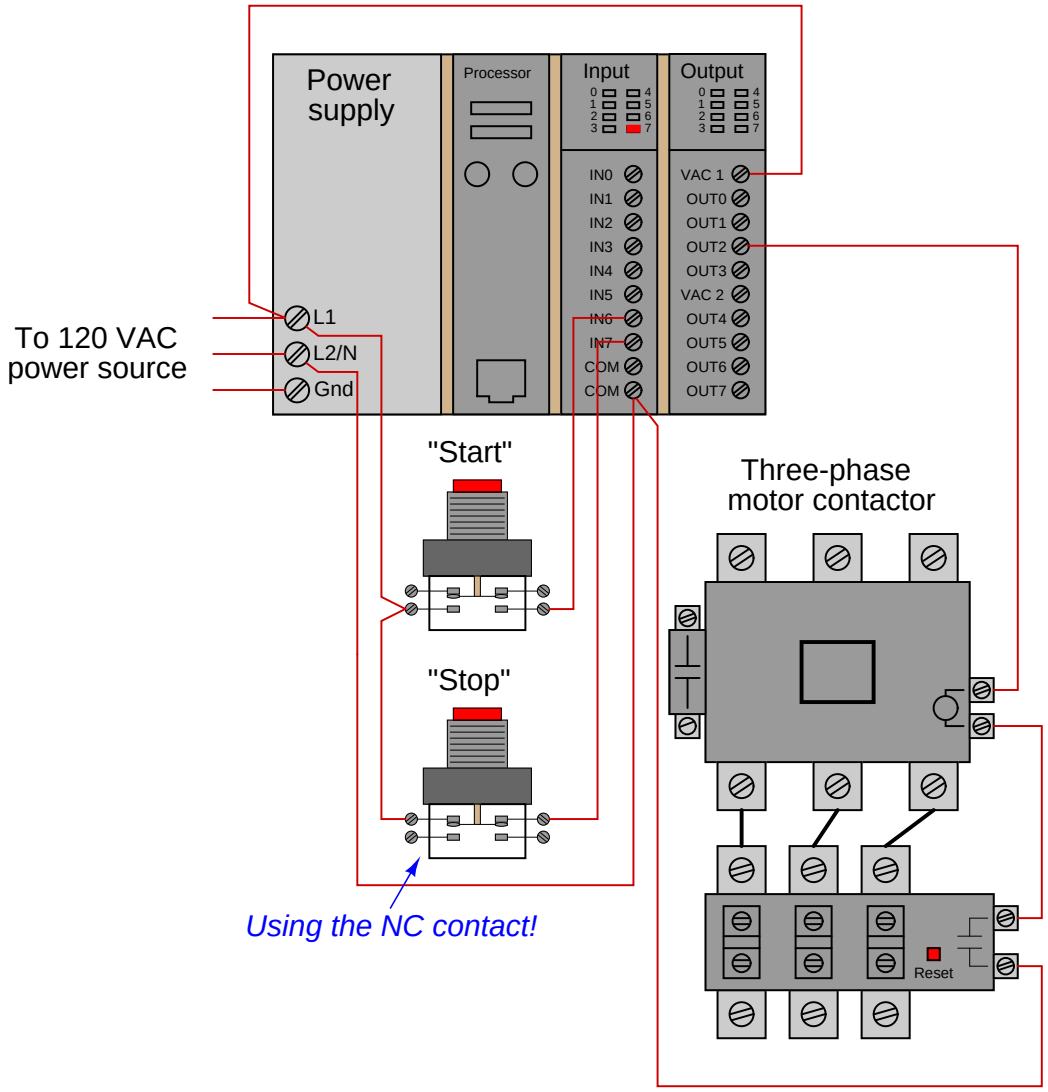
This sequence helps illustrate the *order of evaluation* or *scan order* of a ladder diagram program. The PLC reads a ladder diagram from left to right, top to bottom, in the same general order as a

human being reads sentences and paragraphs written in English. However, according to the IEC 61131-3 standard, a PLC program must evaluate (read) all inputs (contacts) to a function before determining the status of a function's output (coil or coils). In other words, the PLC does not make any decision on how to set the state of a coil until all contacts providing power to that coil have been read. Once a coil's status has been written to memory, any contacts bearing the same tag name will update with that status on subsequent rungs in the program.

Step 5 in the previous sequence is particularly illustrative. When the human operator presses the "Stop" pushbutton, the input channel for `IN_switch_Stop` becomes activated, which "opens" the normally-closed virtual contact `IN_switch_Stop`. Upon the next scan of this program rung, the PLC evaluates all input contacts (`IN_switch_Start`, `IN_switch_Stop`, and `OUT_contactor`) to check their status before deciding what status to write to the `OUT_contactor` coil. Seeing that the `IN_switch_Stop` contact has been forced open by the activation of its respective discrete input channel, the PLC writes a "0" (or "False") state to the `OUT_contactor` coil. However, the `OUT_contactor` feedback contact does not update until the next scan, which is why you still see it highlighted in blue during step 5.

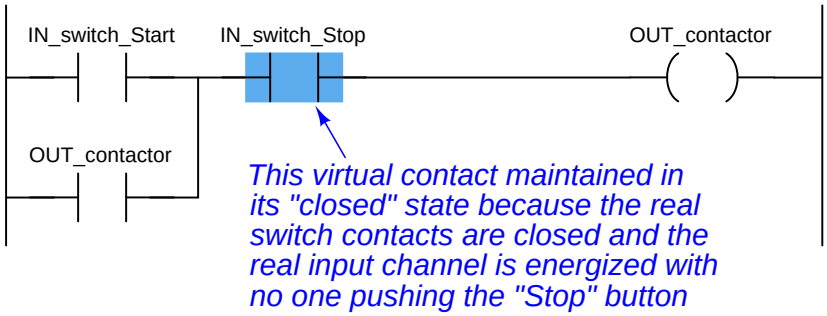
A potential problem with this system as it is designed is that the human operator loses control of the motor in the event of an "open" wiring failure in either pushbutton switch circuit. For instance, if a wire fell off a screw contact for the "Start" pushbutton switch circuit, the motor could not be started if it was already stopped. Similarly, if a wire fell off a screw contact for the "Stop" pushbutton switch circuit, the motor could not be stopped if it was already running. In either case, a broken wire connection acts the same as the pushbutton switch's "normal" status, which is to keep the motor in its present state. In some applications, this failure mode would not be a severe problem. In many applications, though, it is quite dangerous to have a running motor that cannot be stopped. For this reason, it is customary to design motor start/stop systems a bit differently from what has been shown here.

In order to build a "fail-stop" motor control system with our PLC, we must first re-wire the pushbutton switch to use its normally-closed (NC) contact:



This keeps discrete input channel 7 activated when the pushbutton is unpressed. When the operator presses the "Stop" pushbutton, the switch's contacts will be forced open, and input channel 7 will de-energize. If a wire happens to fall off a screw terminal in the "Stop" switch circuit, input channel 7 will de-energize just the same as if someone pressed the "Stop" pushbutton, which will automatically shut off the motor.

In order for the PLC program to work properly with this new switch wiring, the virtual contact for `IN_switch_Stop` must be changed from a normally-closed (NC) to a normally-open (NO).



As before, the `IN_switch_Stop` virtual contact is in the "closed" state when no one presses the "Stop" switch, enabling the motor to start any time the "Start" switch is pressed. Similarly, the `IN_switch_Stop` virtual contact will open any time someone presses the "Stop" switch, thus stopping virtual "power" from flowing to the `OUT_contactor` coil.

Although this is a very common way to build PLC-controlled motor start/stop systems – with an NC pushbutton switch and an NO "Stop" virtual contact – students new to PLC programming often find this logical reversal confusing. Perhaps the most common reason for this confusion is a mis-understanding of the "normal" concept for switch contacts, be they real or virtual. The `IN_switch_Stop` virtual contact is programmed to be normally-open (NO), but yet it is *typically* found in the closed state. Recall that the "normal" status of any switch is its status while in a condition of minimum stimulus, *not* necessarily its status while the process is in a "normal" operating mode. The "normally-open" virtual contact `IN_switch_Stop` is typically found in the closed state because its corresponding input channel is typically found energized, owing to the normally-closed pushbutton switch contacts, which pass real electrical power to the input channel while no one presses the switch. Just because a switch is configured as normally-open does not necessarily mean it will be *typically* found in the open state! The status of any switch contact, whether real or virtual, is a function of its configuration (NO versus NC) and the stimulus applied to it.

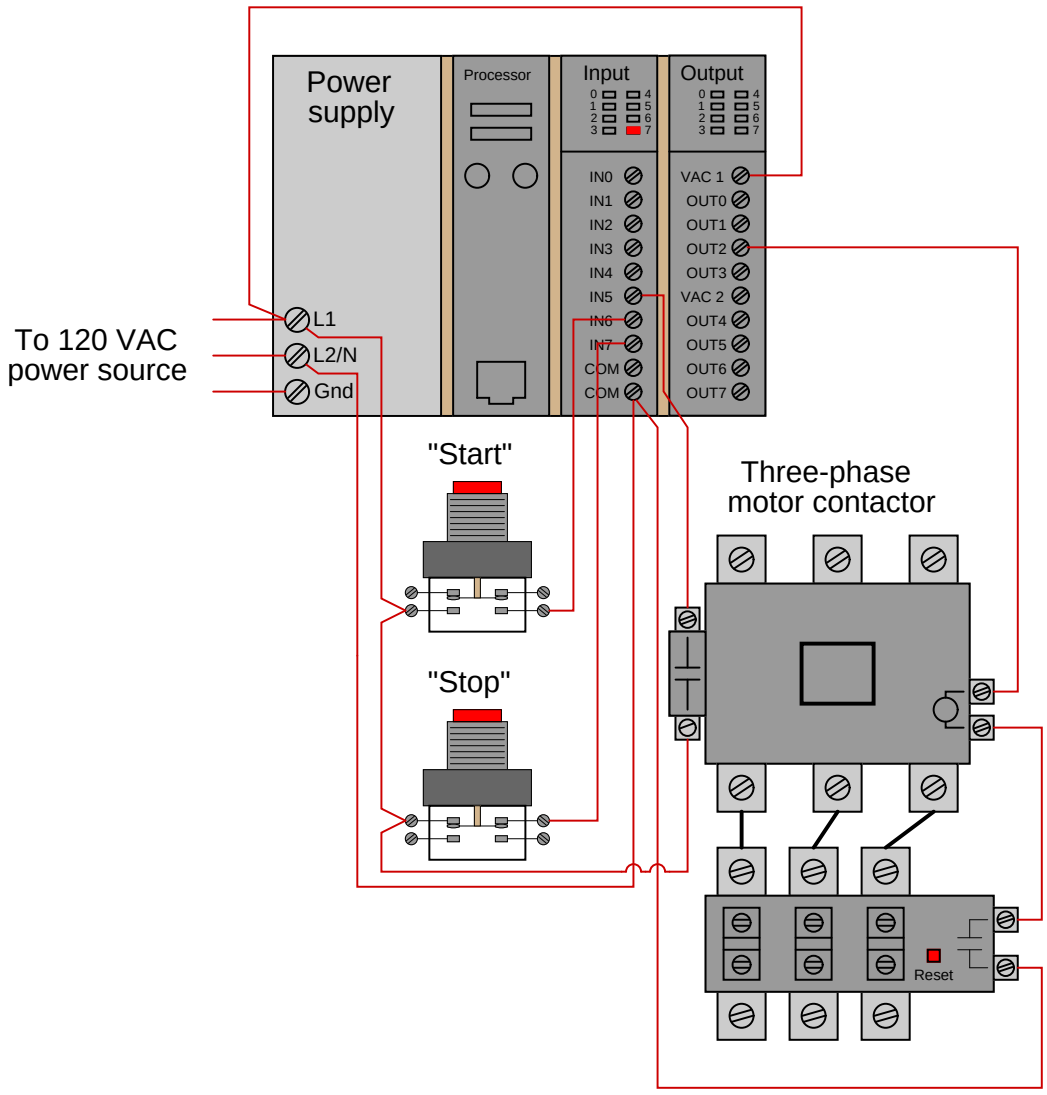
Another concern surrounding real-world wiring problems is what this system will do if the motor contactor coil circuit opens for any reason. An open circuit may develop as a result of a wire falling off a screw terminal, or it may occur because the thermal overload contact tripped open due to an over-temperature event. The problem with our motor start/stop system as designed is that it is not "aware" of the contactor's real status. In other words, the PLC "thinks" the contactor will be energized any time discrete output channel 2 is energized, but that may not actually be the case if there is an open fault in the contactor's coil circuit.

This may lead to a dangerous condition if the open fault in the contactor's coil circuit is later cleared. Imagine an operator pressing the "Start" switch but noticing the motor does not actually start. Wondering why this may be, he or she goes to look at the overload relay to see if it is tripped. If it is tripped, and the operator presses the "Reset" button on the overload assembly, the motor will immediately start because the PLC's discrete output has remained energized all the time following the pressing of the "Start" switch. Having the motor start up as soon as the thermal overload is reset may come as a surprise to operations personnel, and this could be quite dangerous if anyone

happens to be near the motor-powered machinery when it starts.

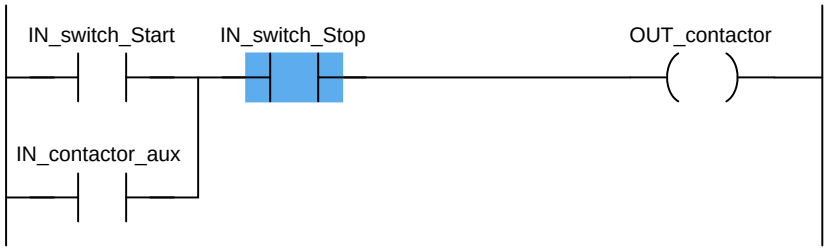
What would be safer is a motor control system that refuses to "latch" on unless the contactor actually energizes when the "Start" switch is pressed. For this to be possible, the PLC must have some way of sensing the contactor's status.

In order to make the PLC "aware" of the contactor's real status, we may connect the auxiliary switch contact to one of the unused discrete input channels on the PLC, like this:



Now, the PLC has a way of sensing the contactor's actual status.

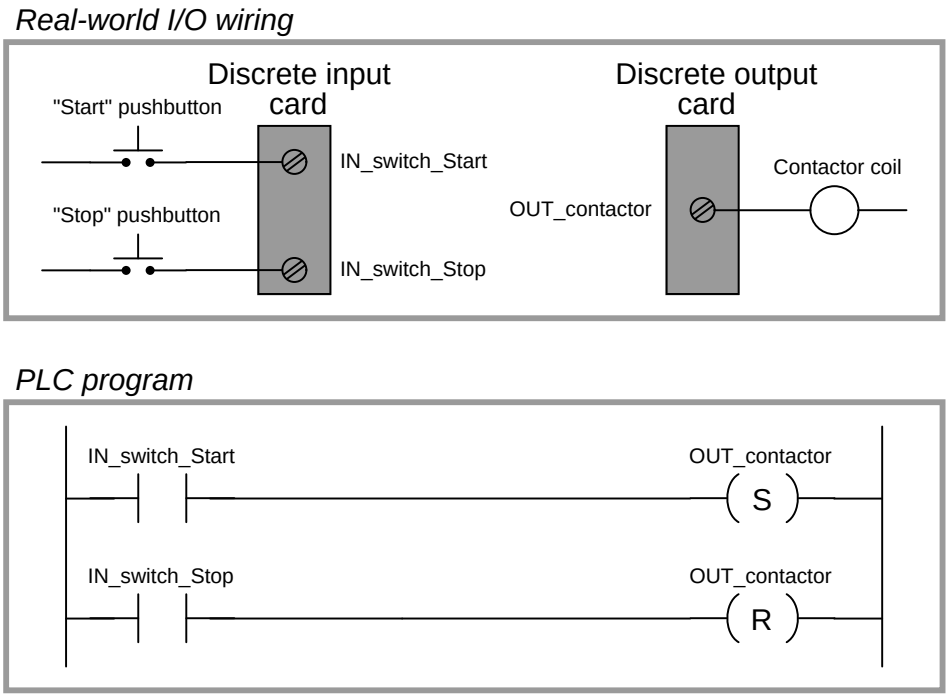
We may modify the PLC program to recognize this status by assigning a new tag name to this input (`IN_contactor_aux`) and using a normally-open virtual contact of this name as the seal-in contact instead of the `OUT_contactor` bit:



Now, if the contactor fails to energize for any reason when the operator presses the "Start" switch, the PLC's output will fail to latch when the "Start" switch is released. When the open fault in the contactor's coil circuit is cleared, the motor will *not* immediately start up, but rather wait until the operator presses the "Start" switch again, which is a much safer operating characteristic than before.

A special class of virtual "coil" used in PLC ladder programming that bears mentioning is the "latching" coil. These usually come in two forms: a *set* coil and a *reset* coil. Unlike a regular "output" coil that continually writes to a bit in the PLC's memory with every scan of the program, "set" and "reset" coils only write to a bit in memory when energized. Otherwise, the bit is allowed to retain its last value.

A very simple motor start/stop program could be written with just two input contacts and two of these latching coils (both bearing the same tag name, writing to the same bit in memory):



Note the use of normally-open (NO) pushbutton switch contacts (again!), with no auxiliary contact providing status indication of the contactor to the PLC. This is a very minimal program, shown for the strict purpose of illustrating the use of "set" and "reset" latching coils in ladder diagram PLC programming.

"Set" and "Reset" coils[11] are examples of what is known in the world of PLC programming as *retentive instructions*. A "retentive" instruction *retains* its value after being virtually "de-energized" in the ladder diagram "circuit." A standard output coil is *non-retentive*, which means it does not "latch" when de-energized. The concept of retentive and non-retentive instructions will appear again as we explore PLC programming, especially in the area of *timers*.

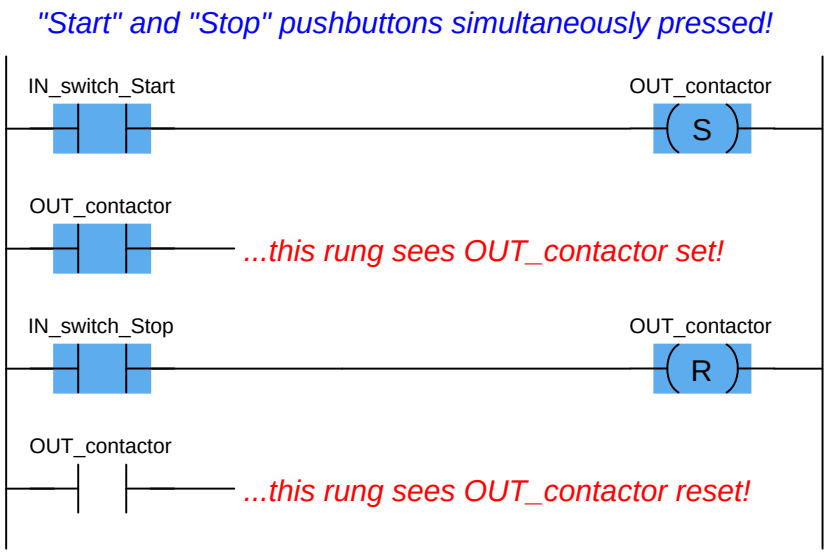
Ordinarily, we try to avoid multiple coils bearing the same label in a PLC ladder diagram program. With each coil representing a "write" instruction, multiple coils bearing the same name represents multiple "write" operations to the same bit in the PLC's memory. Here, with latching coils, there is no conflict because each of the coils only writes to the `OUT_contactor` bit when its

[11] Referred to as "Latch" and "Unlatch" coils by Allen-Bradley.

respective contact is energized. So long as only one of the pushbutton switches is actuated at a time, there is no conflict between the identically-named coils.

This begs the question: what would happen if *both* pushbutton switches were simultaneously pressed? What would happen if *both* "Set" and "Reset" coils were "energized" at the same time? The result is that the `OUT_contactor` bit would first be "set" (written to a value of 1) then "reset" (written to a value of 0) in that order as the two rungs of the program were scanned from top to bottom. PLCs typically do not typically update their discrete I/O registers while scanning the ladder diagram program (this operation takes place either before or after each program scan), so the real discrete output channel status will be whatever the *last* write operation told it to be, in this case "reset" (0, or off).

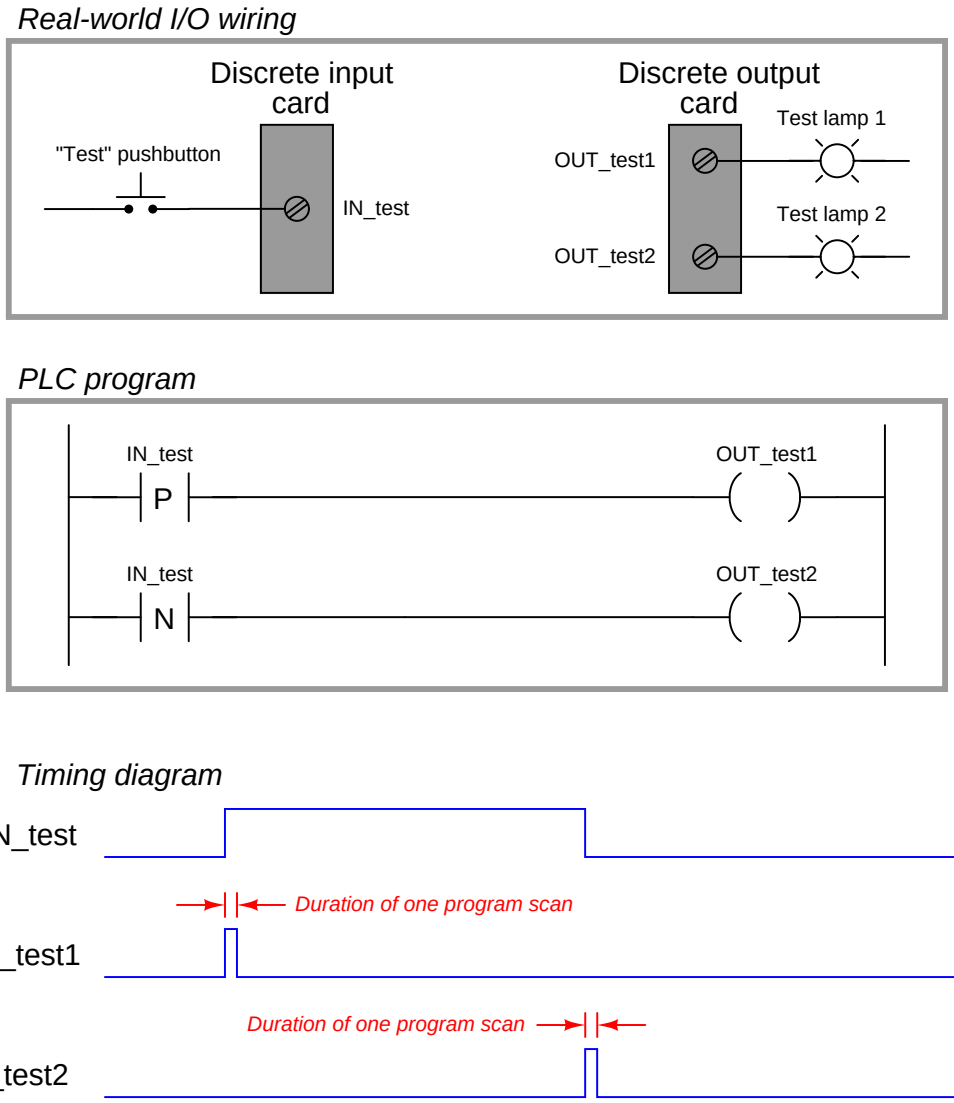
Even if the discrete output is not "confused" due to the conflicting write operations of the "Set" and "Reset" coils, other rungs of the program written between the "Set" and "Reset" rungs might be. Consider for example a case where there were other program rungs following the "Set" and "Reset" rungs reading the status of the `OUT_contactor` bit for some purpose. Those other rungs *would* indeed become "confused" because they would see the `OUT_contactor` bit in the "set" state while the actual discrete output of the PLC (and any rungs following the "Reset" rung) would see the `OUT_contactor` bit in the "reset" state:



Multiple (non-retentive) output coils with the same memory address are almost always a programming *faux pax* for this reason, but even retentive coils which are designed to be used in matched pairs can cause trouble if the implications of simultaneous energization are not anticipated. Multiple *contacts* with identical addresses are no problem whatsoever, because multiple "read" operations to the same bit in memory will never cause a conflict.

The IEC 61131-3 PLC programming standard specifies *transition-sensing* contacts as well as the more customary "static" contacts. A transition-sensing contact will "actuate" only for a duration of one program scan, even if its corresponding bit remains active. Two types of transition-sensing ladder diagram contacts are defined in the IEC standard: one for *positive* transitions and another

for *negative* transitions. The following example shows a wiring diagram, ladder diagram program, and a timing diagram demonstrating how each type of transition-sensing contact functions when stimulated by a real (electrical) input signal to a discrete channel:



When the pushbutton switch is pressed and the discrete input energized, the first test lamp will blink "on" for exactly one scan of the PLC's program, then return to its off state. The positive-transition contact (with the letter "P" inside) activates the coil `OUT_test1` only during the scan it sees the status of `IN_test` transition from "false" to "true," even though the input remains energized for many scans after that transition. Conversely, when the pushbutton switch is released and the discrete input de-energizes, the second test lamp will blink "on" for exactly one scan of the PLC's

program then return to its off state. The negative-transition contact (with the letter "N" inside) activates the coil `OUT_test2` only during the scan it sees the status of `IN_test` transition from "true" to "false," even though the input remains de-energized for many scans after that transition:

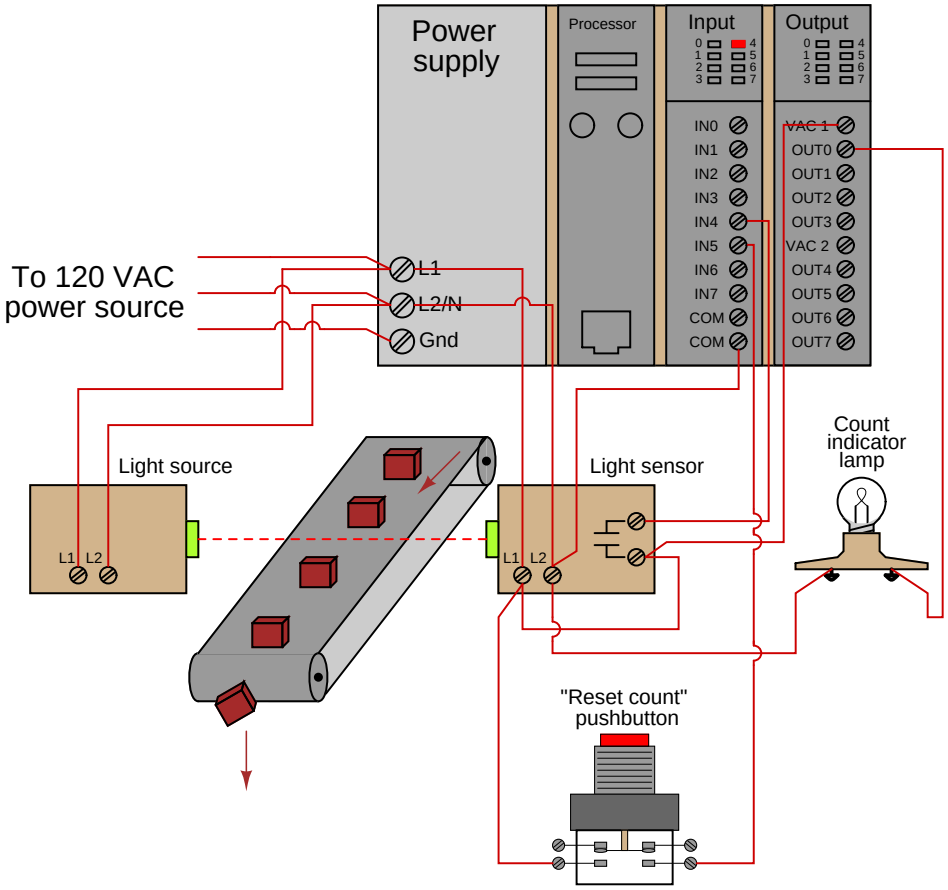
It should be noted that the duration of a single PLC program scan is typically very short: measured in milliseconds. If this program were actually tested in a real PLC, you would probably not be able to see either test lamp light up, since each pulse is so short-lived. Transitional contacts are typically used any time it is desired to execute an instruction just one time following a "triggering" event, as opposed to executing that instruction over and over again so long as the event status is maintained "true."

Contacts and coils represent only the most basic of instructions in the ladder diagram PLC programming language. Many other instructions exist, which will be discussed in the following subsections.

Counters

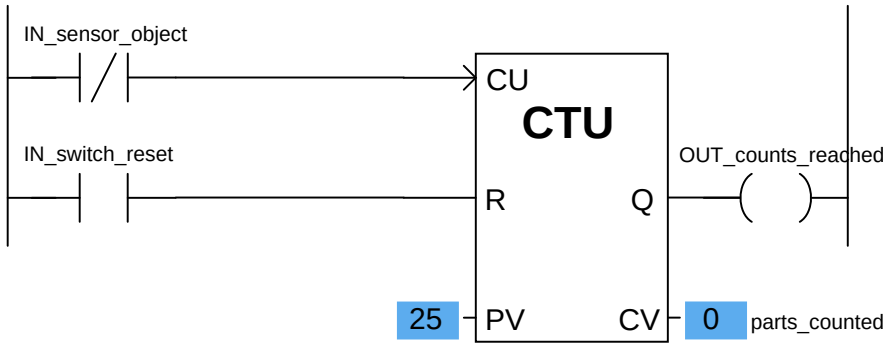
A *counter* is a PLC instruction that either increments (counts up) or decrements (counts down) an integer number value when prompted by the transition of a bit from 0 to 1 ("false" to "true"). Counter instructions come in three basic types: *up* counters, *down* counters, and *up/down* counters. Both "up" and "down" counter instructions have single inputs for triggering counts, whereas "up/down" counters have two trigger inputs: one to make the counter increment and one to make the counter decrement.

To illustrate the use of a counter instruction, we will analyze a PLC-based system designed to count objects as they pass down a conveyor belt:



In this system, a continuous (unbroken) light beam causes the light sensor to close its output contact, energizing discrete channel IN4. When an object on the conveyor belt interrupts the light beam from source to sensor, the sensor's contact opens, interrupting power to input IN4. A pushbutton switch connected to activate discrete input IN5 when pressed will serve as a manual "reset" of the count value. An indicator lamp connected to one of the discrete output channels will serve as an indicator of when the object count value has exceeded some pre-set limit.

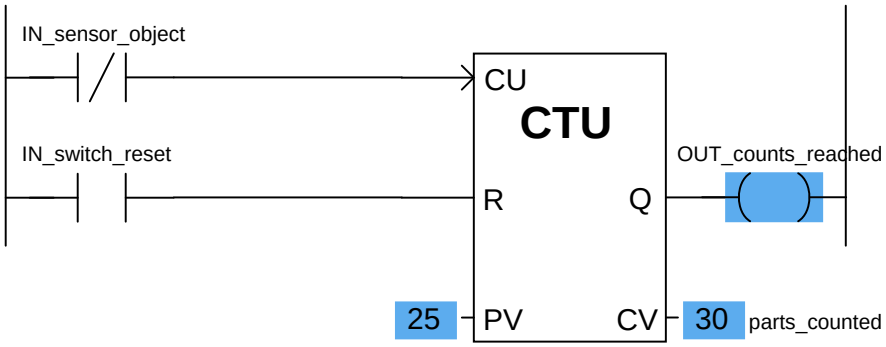
We will now analyze a simple ladder diagram program designed to increment a counter instruction each time the light beam breaks:



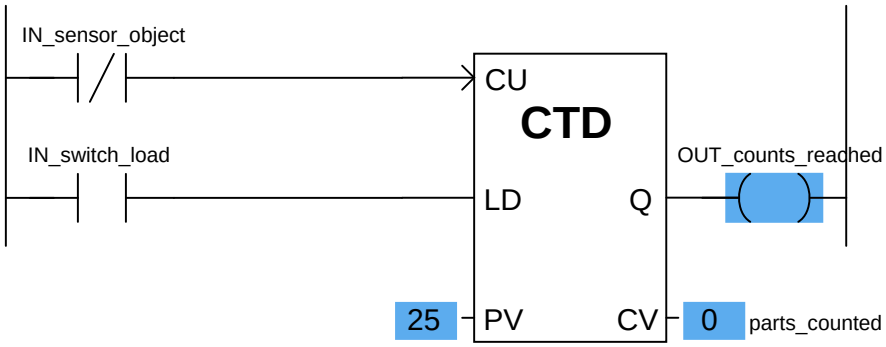
This particular counter instruction (CTU) is an incrementing counter, which means it counts "up" with each off-to-on transition input to its "CU" input. The normally-closed virtual contact (`IN_sensor_object`) is typically held in the "open" state when the light beam is continuous, by virtue of the fact the sensor holds that discrete input channel energized while the beam is continuous. When the beam is broken by a passing object on the conveyor belt, the input channel de-energizes, causing the virtual contact `IN_sensor_object` to "close" and send virtual power to the "CU" input of the counter instruction. This increments the counter just as the leading edge of the object breaks the beam. The second input of the counter instruction box ("R") is the *reset* input, receiving virtual power from the contact `IN_switch_reset` whenever the reset pushbutton is pressed. If this input is activated, the counter immediately resets its current value (CV) to zero.

Status indication is shown in this ladder diagram program, with the counter's preset value (PV) of 25 and the counter's current value (CV) of 0 shown highlighted in blue. The preset value is something programmed into the counter instruction before the system put into service, and it serves as a threshold for activating the counter's output (Q), which in this case turns on the count indicator lamp (the `OUT_counts_reached` coil). According to the IEC 61131-3 programming standard, this counter output should activate whenever the current value is equal to or greater than the preset value (Q is active if CV $\geq$ PV).

This is the status of the same program after thirty objects have passed by the sensor on the conveyor belt. As you can see, the current value of the counter has increased to 30, exceeding the preset value and activating the discrete output:



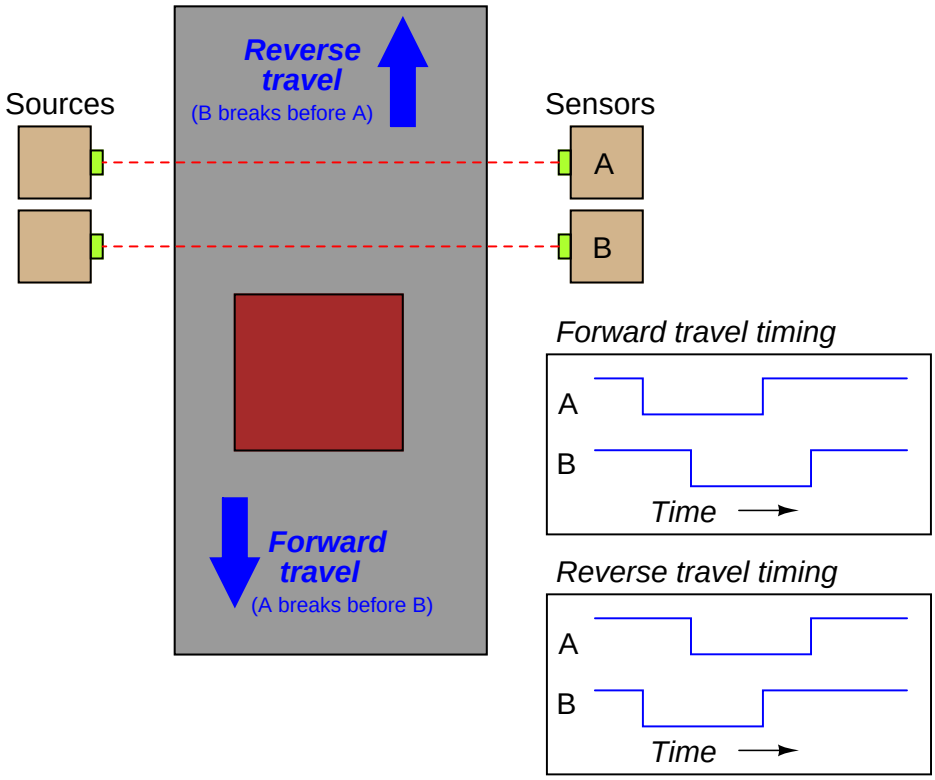
If all we did not care about maintaining an accurate total count of objects past 25 – but merely wished the program to indicate when 25 objects had passed by – we could also use a *down* counter instruction preset to a value of 25, which turns on an output coil when the count reaches zero:



Here, a "load" input causes the counter's current value to equal the preset value (25) when activated. With each sensor pulse received, the counter instruction decrements. When it reaches zero, the Q output activates.

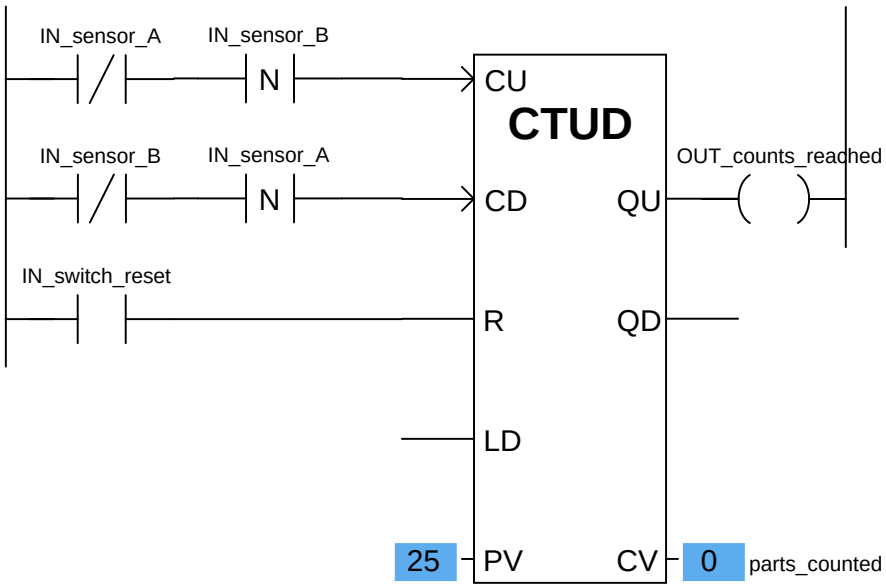
A potential problem in either version of this object-counting system is that the PLC cannot discriminate between forward and reverse motion on the conveyor belt. If, for instance, the conveyor belt were ever reversed in direction, the sensor would continue to count objects that had already passed by before (in the forward direction) as those objects retreated on the belt. This would be a problem because the system would "think" more objects had passed along the belt (indicating greater production) than actually did.

One solution to this problem is to use an up/down counter, capable of both incrementing (counting up) and decrementing (counting down), and equip this counter with two light-beam sensors capable of determining direction of travel. If two light beams are oriented parallel to each other, closer than the width of the narrowest object passing along the conveyor belt, we will have enough information to determine direction of object travel:



This is called *quadrature* signal timing, because the two pulse waveforms are approximately 90^o (one-*quarter* of a period) apart in phase. We can use these two phase-shifted signals to increment or decrement an up/down counter instruction, depending on which pulse leads and which pulse lags.

A ladder diagram PLC program designed to interpret the quadrature pulse signals is shown here, making use of negative-transition contacts as well as standard contacts:



The counter will increment (count up) when sensor B de-energizes only if sensor A is already in the de-energized state (i.e. light beam A breaks before B). The counter will decrement (count down) when sensor A de-energizes only if sensor B is already in the de-energized state (i.e. light beam B breaks before A).

Note that the up/down counter has both a "reset" (R) input and a "load" input ("LD") to force the current value. Activating the reset input forces the counter's current value (CV) to zero, just as we saw with the "up" counter instruction. Activating the load input forces the counter's current value to the preset value (PV), just as we saw with the "down" counter instruction. In the case of an up/down counter, there are two Q outputs: a QU (output up) to indicate when the current value is equal to or greater than the preset value, and a QD (output down) to indicate when the current value is equal to or less than zero.

Note how the current value (CV) of each counter shown is associated with a tag name of its own, in this case `parts_counted`. The integer number of a counter's current value (CV) is a variable in the PLC's memory just like boolean values such as `IN_sensor_A` and `IN_switch_reset`, and may be associated with a tag name or symbolic address just the same[12]. This allows other instructions in a PLC program to read (and sometimes write!) values from and to that memory location.

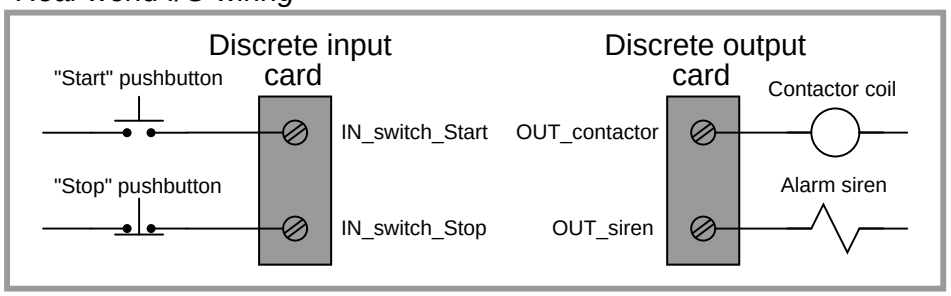
[12]This represents the IEC 61131-3 standard, where each variable within an instruction may be "connected" to its own arbitrary tag name. Other programming conventions may differ somewhat. The Allen-Bradley Logix5000 series of controllers is one of those that differs, following a convention reminiscent of structure element addressing in the C programming language: each counter is given a tag name, and variables in each counter are addressed as elements within that structure. For example, a Logix5000 counter instruction might be named `parts_count`, with the accumulated count value (equivalent to the IEC's "current value") addressed as `parts_count.ACC` (each element within the counter specified as a suffix to the counter's tag name).

Timers

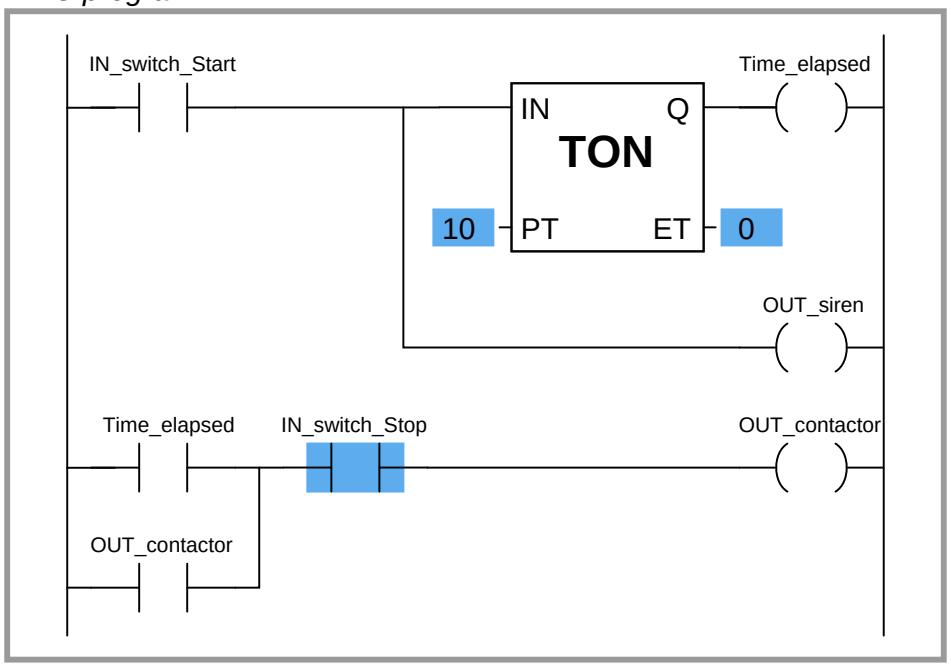
A *timer* is a PLC instruction measuring the amount of time elapsed following an event. Timer instructions come in two basic types: *on-delay* timers and *off-delay* timers. Both "on-delay" and "off-delay" timer instructions have single inputs triggering the timed function.

An "on-delay" timer activates an output only when the input has been active for a minimum amount of time. Take for instance this PLC program, designed to sound an audio alarm siren prior to starting a conveyor belt. To start the conveyor belt motor, the operator must press and hold the "Start" pushbutton for 10 seconds, during which time the siren sounds, warning people to clear away from the conveyor belt that is about to start. Only after this 10-second start delay does the motor actually start (and latch "on"):

Real-world I/O wiring



PLC program

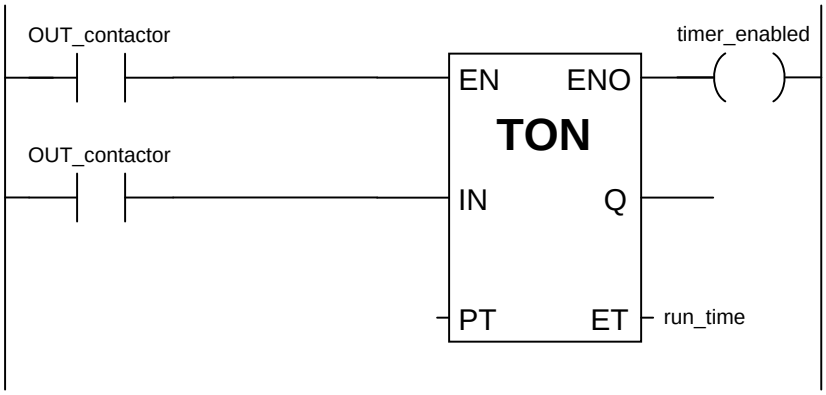


Similar to an "up" counter, the on-delay timer's elapsed time (ET) value increments once per second until the preset time (PT) is reached, at which time its output (Q) activates. In this program, the preset time value is 10 seconds, which means the Q output will not activate until the "Start" switch has been depressed for 10 seconds. The alarm siren output, which is not activated by the timer, energizes immediately when the "Start" pushbutton is pressed.

An important detail regarding this particular timer's operation is that it be *non-retentive*. This means the timer instruction should *not* retain its elapsed time value when the input is de-activated. Instead, the elapsed time value should reset back to zero every time the input de-activates. This ensures the timer resets itself when the operator releases the "Start" pushbutton. A *retentive* on-delay timer, by contrast, maintains its elapsed time value even when the input is de-activated. This makes it useful for keeping "running total" times for some event.

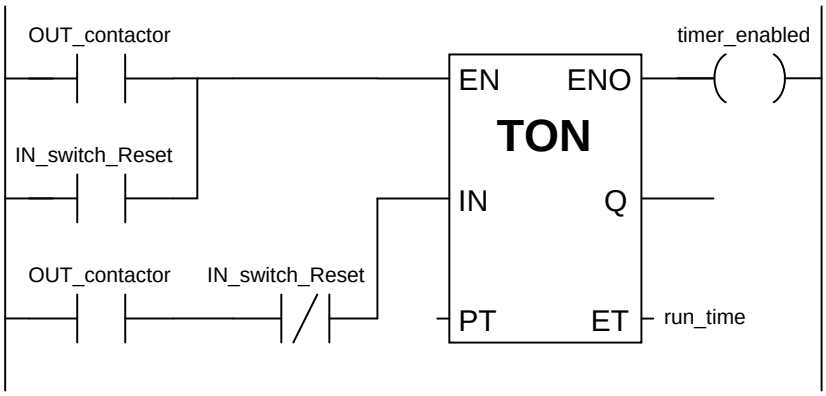
Most PLCs provide retentive and non-retentive versions of on-delay timer instructions, such that the programmer may choose the proper form of on-delay timer for any particular application. The IEC 61131-3 programming standard, however, addresses the issue of retentive versus non-retentive timers a bit differently. According to the IEC 61131-3 standard, a timer instruction may be specified with an additional *enable* input (EN) that causes the timer instruction to behave non-retentively when activated, and retentively when de-activated. The general concept of the enable (EN) input is that the instruction behaves "normally" so long as the enable input is active (in this case, non-retentive timing action is considered "normal" according to the IEC 61131-3 standard), but the instruction "freezes" all execution whenever the enable input de-activates. This "freezing" of operation has the effect of retaining the current time (CT) value even if the input signal de-activates.

For example, if we wished to add a retentive timer to our conveyor control system to record total run time for the conveyor motor, we could do so using an "enabled" IEC 61131-3 timer instruction like this:



When the motor's contactor bit (`OUT_contactor`) is active, the timer is enabled and allowed to time. However, when that bit de-activates (becomes "false"), the timer instruction as a whole is disabled, causing it to "freeze" and retain its current time (CT) value[13]. This allows the motor to be started and stopped, with the timer maintaining a tally of total motor run time.

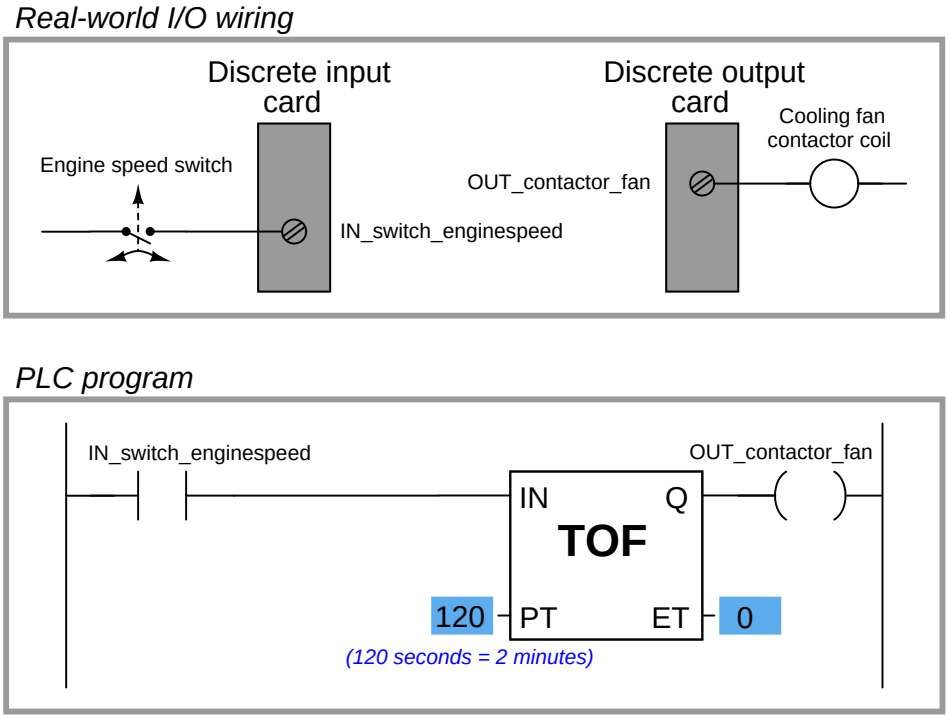
If we wished to give the operator the ability to manually reset the total run time value to zero, we could hard-wire an additional switch to the PLC's discrete input card and add "reset" contacts to the program like this:



Whenever the "Reset" switch is pressed, the timer is enabled (EN) but the timing input (IN) is disabled, forcing the timer to (non-retentively) reset its current time (CT) value to zero.

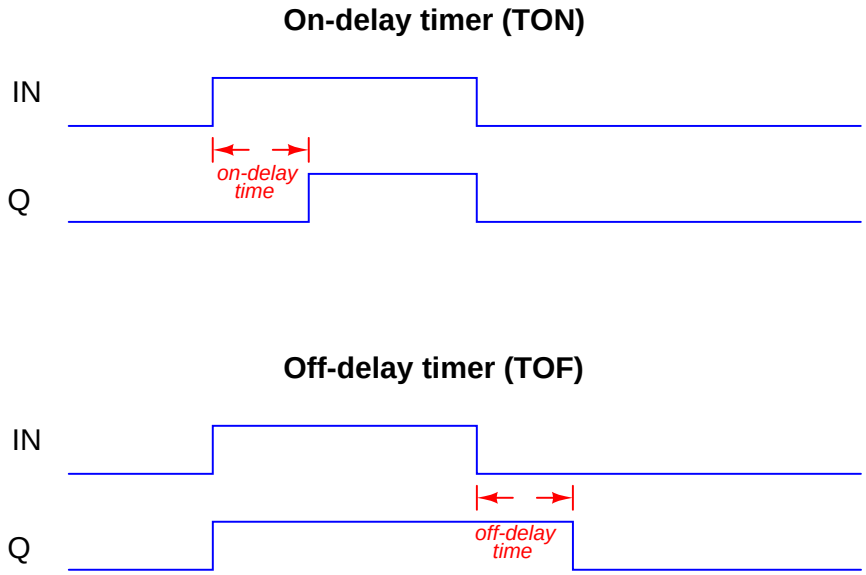
[13]The "enable out" (ENO) signal on the timer instruction serves to indicate the instruction's status: it activates when the enable input (EN) activates and de-activates when either the enable input de-activates or the instruction generates an error condition (as determined by the PLC manufacturer's internal programming). The ENO output signal serves no useful purpose in this particular program, but it is available if there were any need for other rungs of the program to be "aware" of the run-time timer's status.

The other major type of PLC timer instruction is the *off-delay* timer. This timer instruction differs from the on-delay type in that the timing function begins as soon as the instruction is de-activated, not when it is activated. An application for an off-delay timer is a cooling fan motor control for a large industrial engine. In this system, the PLC starts an electric cooling fan as soon as the engine is detected as rotating, and keeps that fan running for two minutes following the engine's shut-down to dissipate residual heat:



When the input (IN) to this timer instruction is activated, the output (Q) immediately activates (with no time delay at all) to turn on the cooling fan motor contactor. This provides the engine with cooling as soon as it begins to rotate (as detected by the speed switch connected to the PLC's discrete input). When the engine stops rotating, the speed switch returns to its normally-open position, de-activating the timer's input signal which starts the timing sequence. The Q output remains active while the timer counts from 0 seconds to 120 seconds. As soon as it reaches 120 seconds, the output de-activates (shutting off the cooling fan motor) and the elapsed time value remains at 120 seconds until the input re-activates, at which time it resets back to zero.

The following timing diagrams compare and contrast on-delay with off-delay timers:



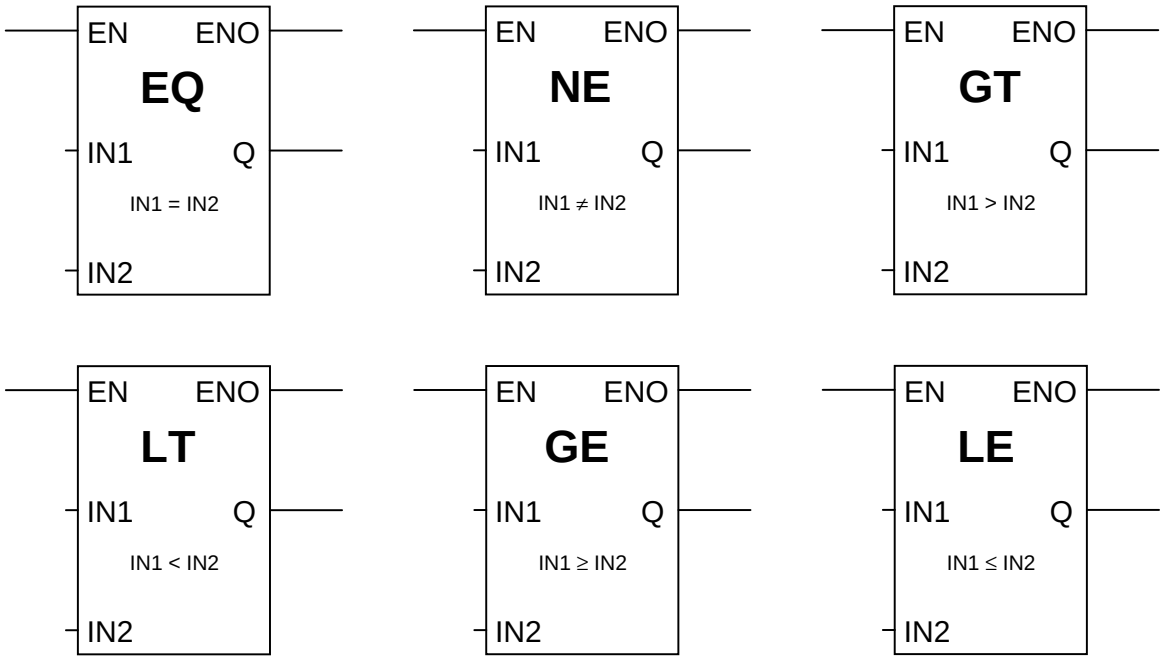
While it is common to find on-delay PLC instructions offered in both retentive and non-retentive forms within the instruction sets of nearly every PLC manufacturer and model, it is almost unheard of to find retentive off-delay timer instructions. Typically, off-delay timers are non-retentive only[14].

[14]The enable (EN) input signals specified in the IEC 61131-3 programming standard make retentive off-delay timers possible (by de-activating the enable input while maintaining the "IN" input in an inactive state), but bear in mind that most PLC implementations of timers do not have separate EN and IN inputs. This means (for most PLC timer instructions) the only input available to activate the timer is the "IN" input, in which case it is *impossible* to create a retentive off-delay timer (since such a timer's elapsed time value would be immediately re-set to zero each time the input re-activates).

Data comparison and math instructions

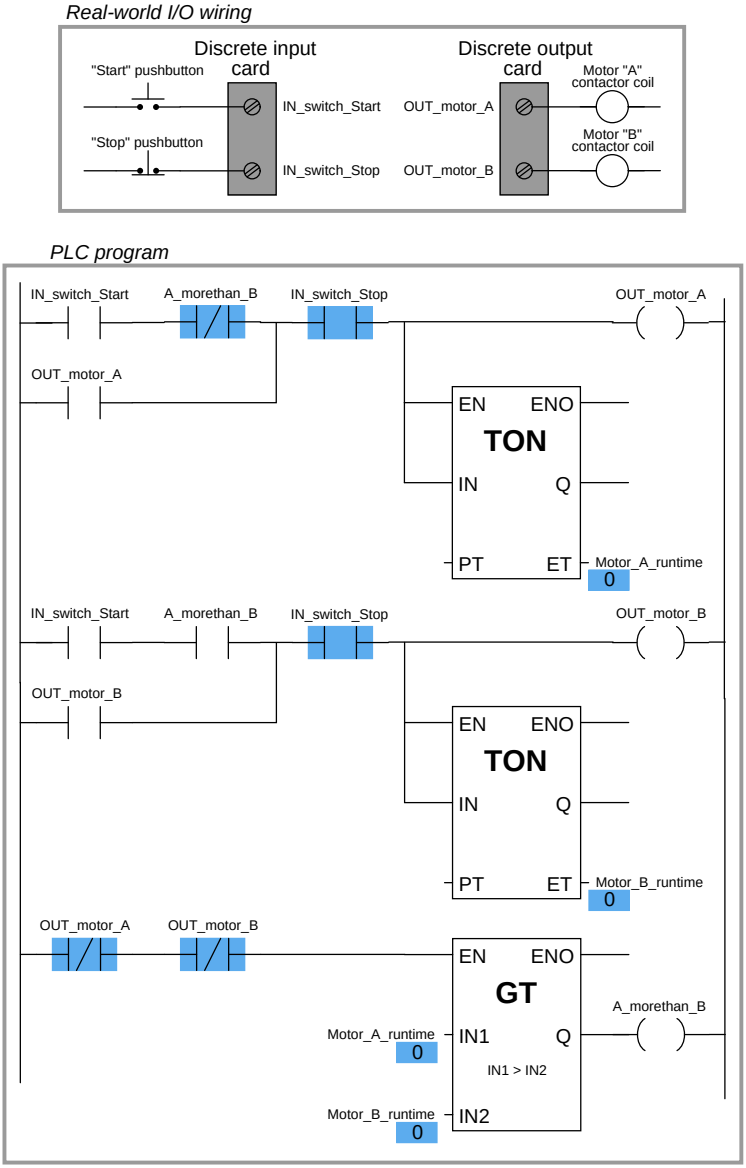
As we have seen with counter and timers, some PLC instructions generate digital values other than simple Boolean (on/off) signals. Counters have current value (CV) registers and timers have elapsed time (ET) registers, both of which are typically integer number values. Many other PLC instructions are designed to receive and manipulate non-Boolean values such as these to perform useful control functions.

The IEC 61131-3 standard specifies a variety of *data comparison* instructions for comparing two non-Boolean values, and generating Boolean outputs. The basic comparative operations of "less than" ($<$), "greater than" ($>$), "less than or equal to" ($\leq$), "greater than or equal to" ($\geq$), "equal to" ($=$), and "not equal to" ($\neq$) may be found as a series of "box" instructions in the IEC standard:



The Q output for each instruction "box" activates whenever the evaluated comparison function is "true" and the enable input (EN) is active. If the enable input remains active but the comparison function is false, the Q output de-activates. If the enable input de-de-activates, the Q output retains its last state.

A practical application for a comparative function might be to compare the total run-time of two redundant electric motors[15], the PLC determining which motor to turn on next based on which motor has run the least:



In this program, two retentive on-delay timers keep track of each electric motor's total run time, storing the run time values in two registers in the PLC's memory: `Motor_A_runtime` and `Motor_B_runtime`. These two integer values are input to the "greater than" instruction box for

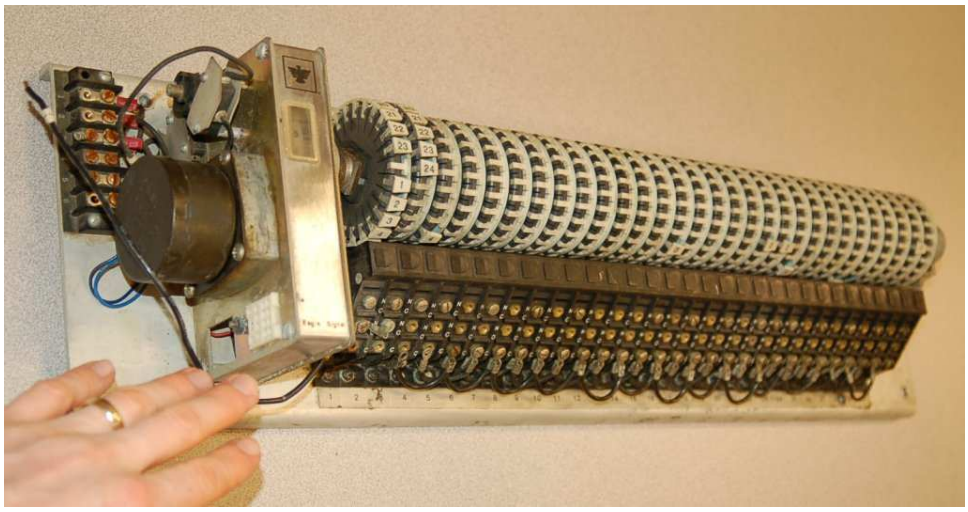
[15]Perhaps two pumps performing the same pumping function, one serving as a backup to the other.

comparison. If motor A has run longer than motor B, motor B will be the one enabled to start up next time the "start" switch is pressed. If motor A has run less time or the same amount of time as motor B (the scenario shown by the blue-highlighted status indications), motor A will be the one enabled to start. The two series-connected virtual contacts `OUT_motor_A` and `OUT_motor_B` ensure the comparison between motor run times is not made until both motors are stopped.

Sequencers

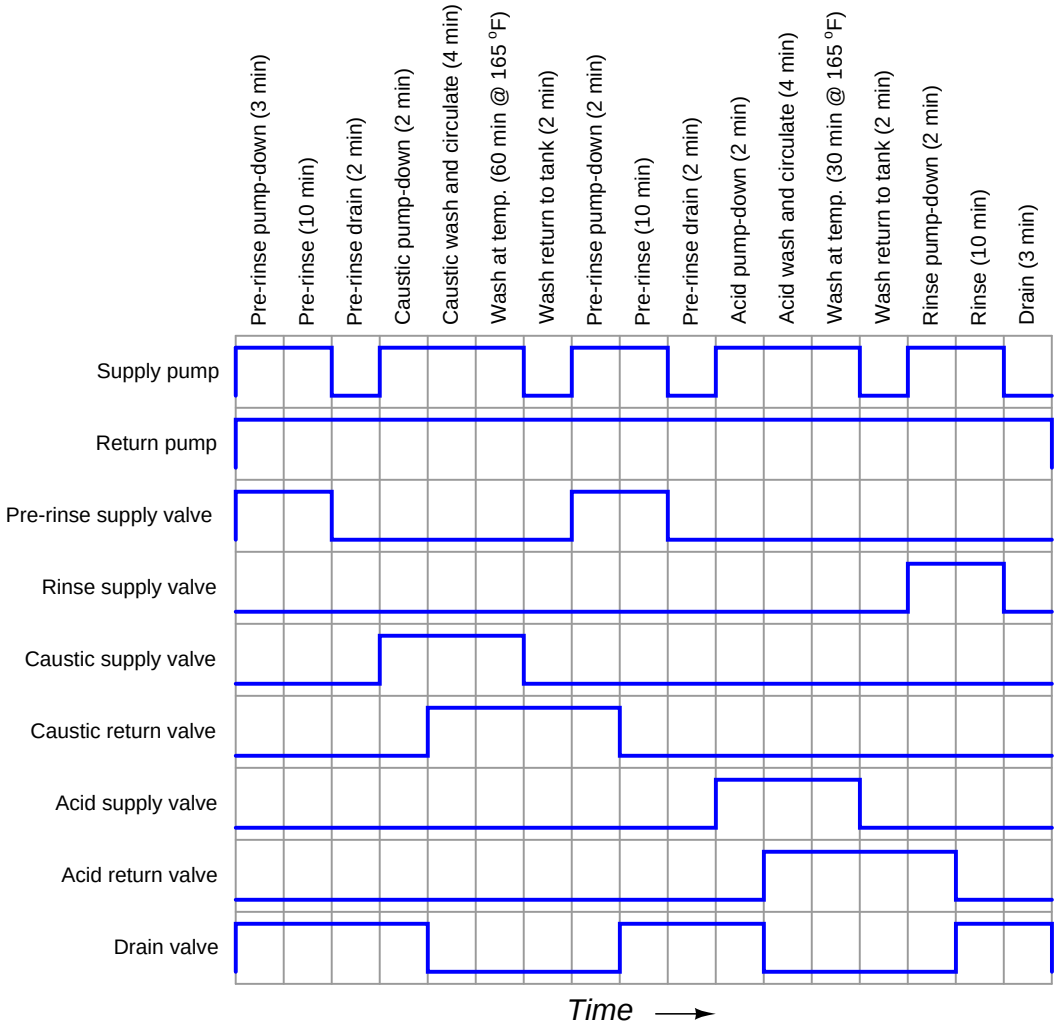
Many industrial processes require control actions to take place in certain, predefined sequences. Batch processes are perhaps the most striking example of this, where materials for making a batch must be loaded into the process vessels, parameters such as temperature and pressure controlled during the batch processing, and then discharge of the product monitored and controlled. Before the advent of reliable programmable logic devices, this form of sequenced control was usually managed by an electromechanical device known as a *drum sequencer*. This device worked on the principle of a rotating cylinder (drum) equipped with tabs to actuate switches as the drum rotated into certain positions. If the drum rotated at a constant speed (turned by a clock motor), those switches would actuate according to a timed schedule[16].

The following photograph shows a drum sequencer with 30 switches. Numbered tabs on the circumference of the drum mark the drum's rotary position in one of 24 increments. With this number of switches and tabs, the drum can control up to thirty discrete (on/off) devices over a series of twenty-four sequenced steps:



[16]The operation of the drum is not unlike that of an old *player piano*, where a strip of paper punched with holes caused hammers in the piano to automatically strike their respective strings as the strip was moved along at a set speed, thus playing a pre-programmed song.

A typical application for a sequencer is to control a *Clean In Place* (*CIP*) system for a food processing vessel, where a process vessel must undergo a cleaning cycle to purge it of any biological matter between food processing cycles. The steps required to clean the vessel are well-defined and must always occur in the same sequence in order to ensure hygienic conditions. An example timing chart is shown here:



In this example, there are nine discrete outputs – one for each of the nine final control elements (pumps and valves) – and seventeen steps to the sequence, each one of them timed. In this particular sequence, the only input is the discrete signal to commence the CIP cycle. From the initiation of the CIP to its conclusion two and a half hours (150 minutes) later, the sequencer simply steps through the programmed routine.

In a general sense, the operation of a drum sequencer is that of a *state machine*: the output

of the system depends on the condition of the machine's internal state (the drum position), not just the conditions of the input signals. Digital computers are very adept at implementing state functions, and so the general function of a drum sequencer should be (and is) easy to implement in a PLC. Other PLC functions we have seen ("latches" and timers in particular) are similar, in that the PLC's output at any given time is a function of both its present input condition(s) and its past input condition(s). Sequencing functions expand upon this concept to define a much larger number of possible states ("positions" of a "drum"), some of which may even be timed.

Unfortunately, despite the utility of drum sequence functions and their ease of implementation in digital form, there seems to be very little standardization between PLC manufacturers regarding sequencing functions. Sadly, the IEC 61131-3 standard (at least at the time of this writing, in 2009) does not specifically define a sequencing function suitable for ladder diagram programming. PLC manufacturers are left to invent sequencing instructions of their own design.

12.3.3 Structured Text (ST)

(Will be addressed in future versions of this book)

12.3.4 Instruction List (IL)

(Will be addressed in future versions of this book)

12.3.5 Function Block Diagram (FBD)

(Will be addressed in future versions of this book)

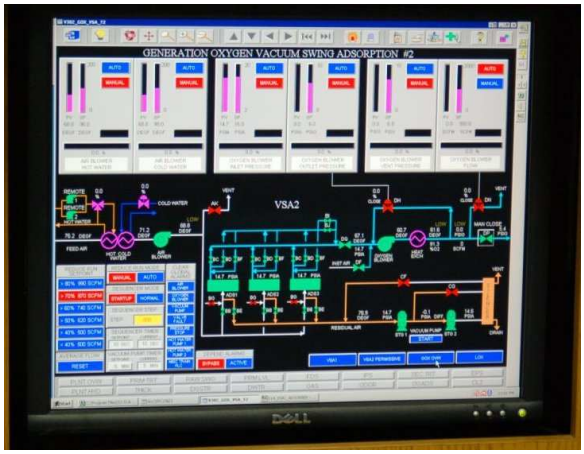
12.3.6 Sequential Function Chart (SFC)

(Will be addressed in future versions of this book)

12.4 Human-Machine Interfaces

Many modern PLC systems are equipped with computer-based interface panels which human operators use to observe data gathered by the PLC and also enter data to the PLC. These panels are generally referred to as *Human*[17]*-Machine Interfaces*, or *HMI* panels.

HMIs may take the form of general-purpose ("personal") computers running special graphic software to interface with a PLC, or as special-purpose computers designed to be mounted in sheet metal panel fronts to perform no task but the operator-PLC interface. The first photograph shows an example of the former, and the second photograph an example of the latter:



Modern HMI panels and software are almost exclusively tag-based, with each graphic object on the screen associated with at least one data tag name, which in turn is associated to data points (bits, or words) in the PLC by way of a tag name database file resident in the HMI. Graphic objects on the HMI screen either accept (read) data from the PLC to present useful information to the operator, send (write) data to the PLC from operator input, or both. The task of programming

[17] An older term for an operator interface panel was the "Man-Machine Interface" or "MMI." However, this fell out of favor due to its sexist tone.

an HMI unit consists of building a tag name database and then drawing screens to illustrate the process to as good a level of detail as operators will need to run it.

Like programmable logic controllers themselves, the capabilities of HMIs have been steadily increasing while their price decreases. Modern HMIs support graphic trending, data archival, advanced alarming, and even web server ability allowing other computers to easily access certain data over wide-area networks. The ability of HMIs to log data over long periods of time relieves the PLC of having to do this task, which is very memory-intensive. This way, the PLC merely "serves" current data to the HMI, and the HMI is able to keep a record of current and past data using its vastly larger memory reserves[18].

[18] If the HMI is based on a personal computer platform, it may even be equipped with a hard disk drive for vast amounts of data storage.

12.5 How to teach yourself PLC programming

First and foremost, you need to get your very own PLC to work with. Computer programming of any kind is not a spectator sport, and can only be learned by significant investment of time and effort at the keyboard. In many ways, learning to program is like learning a new spoken or written language: there is new vocabulary and new grammatical rules to master, and many ways to make mistakes.

Fortunately, many low-cost PLCs exist on the market for individuals to purchase. My own personal favorites are the PLC models manufactured by Koyo and marketed through Automation Direct, both for their low cost and for the outstanding quality of their documentation (User's Manuals). In the United States, where Allen-Bradley (Rockwell) holds the vast majority of market share in programmable logic controllers, it may be advantageous to learn on a low-end Allen-Bradley product such as the MicroLogix 1000. Beware the price of programming software, though!

The first document you should read once you get your PLC is something called a *Getting Started* guide. Every PLC manufacturer publishes a document with this name (or something similar such as *Quick Start* or *Getting Results*). This manual will step you through all the basic procedures for entering a simple program into your PLC and getting it to run. It is generally *far* easier to learn programming by copying and adapting a worked example than it is to start from a "blank page" on your own, just as it is easiest to learn a spoken or written language by practicing sentences spoken in that language by other people before constructing your own sentences from scratch.

Once you have learned the basic steps for entering, running, and saving a PLC program, you are ready to begin building your knowledge of the language's vocabulary and grammar. In computer programming (of all types), there are different *functions* of the language one must become familiar with in order to do useful tasks. A great way to learn how to use these functions is to create your own "demonstration" programs illustrating the use of each function.

For example, if you open up the pages of almost any computer programming book, somewhere near the beginning you will find a demonstration program called "Hello World!" The purpose of a "Hello World!" program is to do nothing more than display the words *Hello World!* on the computer screen. It is an entirely useless program to run, but entering it and running it teaches the programmer the basics of program construction and text message functionality.

By the same token, you may learn the basics of each programming function by writing simple "Hello World"-type programs illustrating each one of those functions. These demonstration programs will not serve any useful purpose (other than to help you learn), and should be kept as simple as possible in order to minimize confusion.

For example, *every* PLC provides programming functions to perform the following tasks:

- Turn discrete outputs on and off
- Count discrete events
- Time events
- Control events in a specific sequence
- Compare numerical values (greater than, less than, equal, not equal)
- Perform arithmetic functions

- Send and receive data through network connections with other PLCs (and other types of devices)

Just as every spoken or written language has verbs, nouns, adjectives, and adverbs to describe actions and things, every PLC programming language has specific functions to perform useful tasks. The details of how to perform each function will vary somewhat between PLC manufacturers and models, but the overall functions are quite similar. The reference manuals provided for your PLC will describe in detail how to use each function. Your task is to write simple demonstration programs for each function, allowing you to directly explore how each function works, and to gain an understanding of each function by observing its behavior and also by making (inevitable) mistakes.

After writing each demonstration program, you should add a lot of comments to it, so you will be able to understand what you did later when you go back to your demonstration program for reference. These comments should cover the following points:

- Proper use of the function
- A verbal description of what the function does
- A list of possible (practical) uses for the function
- Mistakes you may have made (and thus might make again!) in using the function

Years ago when I was teaching myself how to program using the *C* language, I wrote a set of "tutorial" programs demonstrating common programming functions and techniques. The following is a partial list of these tutorial programs, which I still keep to this day:

- Program that accepts and then prints alphanumeric characters (including their equivalent numerical values)
- Program demonstrating how to use command-line arguments to the `main()` function
- Program demonstrating basic "curses" commands for plotting characters at arbitrary locations on the screen
- Program illustrating the declaration and use of *data structures*
- Program illustrating how to prototype and then call *functions* (subroutines)
- Program executing an infinite loop
- Program illustrating how to return a *pointer* from a function

Each one of these tutorial programs is heavily commented, to explain to myself in my own words how they work and what they are doing. Not only did they help me learn how to write programs in C, but they also serve as a handy reference for me any time in the future I need to refresh my knowledge. The act of writing tutorial programs is akin to *journaling* as a way to work through complex problems in life – in a way, it is like having a conversation with yourself.

References

"1758 PLC-5 Programmable Controllers Addressing Reference Manual", Publication 5000-6.4.4, Allen-Bradley Company, Inc., Milwaukee, WI, 1995.

"Allen-Bradley I/O Modules Wiring Diagrams", Publication CIG-WD001A-EN-P, Rockwell Automation, Inc., Milwaukee, WI, 2005.

IEC 61131-3, "International Standard, Programmable Controllers – Part 3: Programming Languages", Edition 2.0, International Electrotechnical Commission, Geneva, Switzerland, 2003.

"Logix5000 Controllers I/O and Tag Data", Publication 1756-PM004B-EN-P, Rockwell Automation, Inc., Milwaukee, WI, 2008.

"Programming with STEP 7", Siemens AG, Nürnberg, Germany, 2006.

"S7-200 Programmable Controller System Manual", Order Number 6ES7298-8FA24-8BH0, Edition 09/2007, Siemens AG, Nürnberg, Germany, 2007.

"SLC 500 Family of Programmable Controllers Addressing Reference Manual", Publication 5000-6.4.23, Allen-Bradley Company, Inc., Milwaukee, WI, 1995.

"SLC 500 Modular Hardware Style User Manual", Publication 1747-UM011E-EN-P, Rockwell Automation, Inc., Milwaukee, WI, 2004.

Chapter 13

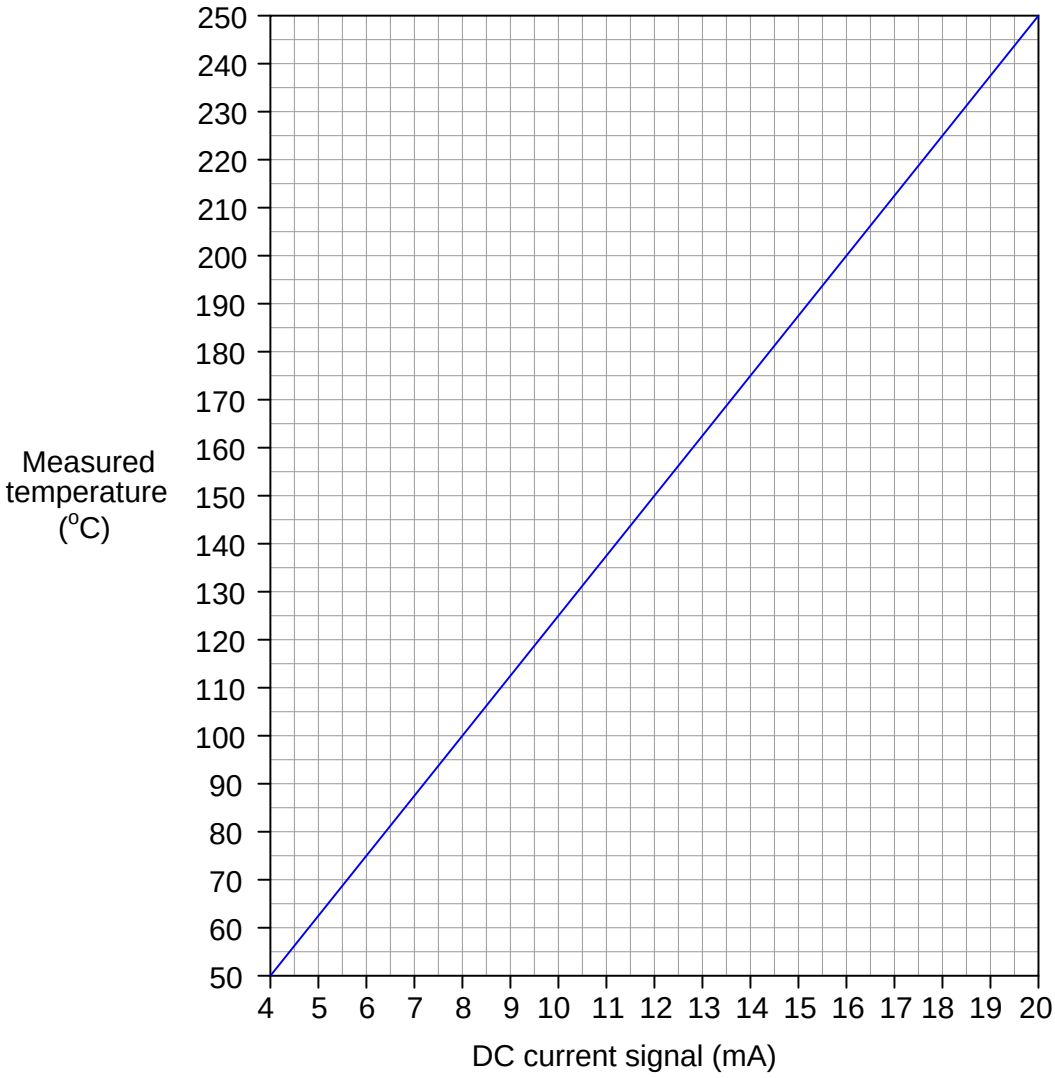
Analog electronic instrumentation

An "analog" electronic signal is a voltage or current whose magnitude represents some physical measurement or control quantity. An instrument is often classified as being "analog" simply by virtue of using an analog signal standard to communicate information, even if the internal construction and design of the instrument may be mostly digital in nature. This is to distinguish such instruments from those making use of no analog electronic signals at all (e.g. wireless or Fieldbus instruments).

13.1 4 to 20 mA analog current signals

The most popular form of signal transmission used in modern industrial instrumentation systems (as of this writing) is the 4 to 20 milliamp DC standard. This is an *analog* signal standard, meaning that the electric current is used to proportionately represent measurements or command signals. Typically, a 4 milliamp current value represents 0% of scale, a 20 milliamp current value represents 100% of scale, and any current value in between 4 and 20 milliamps represents a commensurate percentage in between 0% and 100%.

For example, if we were to calibrate a 4-20 mA temperature transmitter for a measurement range of 50 to 250 degrees C, we could relate the current and measured temperature values on a graph like this:

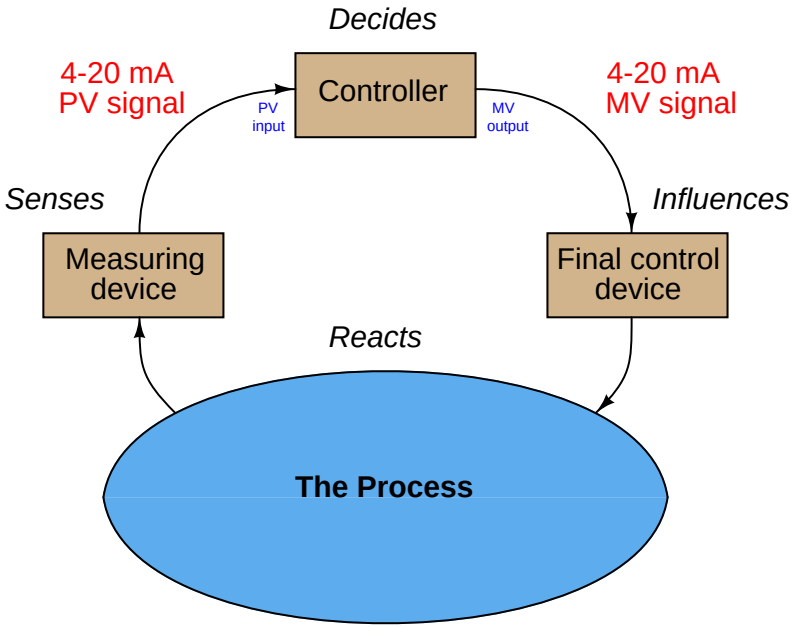


This is not unlike the pneumatic instrument signal standard or 3 to 15 pounds per square inch (PSI), where a varying air pressure signal represents some process measurement in an analog (proportional) fashion. Both signal standards are referred to as *live zero* because their ranges begin with a non-zero value (3 PSI in the case of the 3-15 PSI standard, and 4 milliamps in the case of the 4-20 mA standard). This "live" zero provides a simple means of discriminating between a legitimate

0% signal value and a failed signal (e.g. leaking tube or severed cable)[1].

DC current signals are also used in control systems to command the positioning of a final control element, such as a control valve or a variable-speed motor drive (VSD). In these cases, the milliamp value does not directly represent a process measurement, but rather how the degree to which the final control element influences the process. Typically (but not always!), 4 milliamps commands a closed (shut) control valve or a stopped motor, while 20 milliamps commands a wide-open valve or a motor running at full speed.

Thus, most industrial control systems use at least *two* different 4-20 mA signals: one to represent the process variable (PV) and one to represent the command signal to the final control element (the "manipulated variable" or MV):



The relationship between these two signals depends entirely on the response of the controller. There is no reason to ever expect the two current signals to be equal, for they represent entirely different things. In fact, if the controller is reverse-acting, it is entirely normal for the two current signals to be inversely related: as the PV signal increases going to a reverse-acting controller, the output signal will decrease. If the controller is placed into "manual" mode by a human operator, the output signal will have no automatic relation to the PV signal at all, instead being entirely determined by the operator's whim.

[1] Not all industrial measurement and control signals are "live zero" like the 3-15 PSI and 4-20 mA standards. 0 to 10 volts DC is a common "dead zero" signal standard, although far more common in environmental (building heating and cooling) control systems than industrial control systems. I once encountered an old analog control system using -10 volts to +10 volts as its analog signal range, which meant 0 volts represented a 50% signal! A failed signal path in such a system could have been very misleading indeed, as a 50% signal value is not suspicious in the least.

13.2 Relating 4 to 20 mA signals to instrument variables

Calculating the equivalent milliamp value for any given percentage of signal range is quite easy. Given the linear relationship between signal percentage and milliamps, the equation takes the form of the standard *slope-intercept* line equation $y = mx + b$. Here, y is the equivalent current in milliamps, x is the desired percentage of signal, m is the span of the 4-20 mA range (16 mA), and b is the offset value, or the "live zero" of 4 mA:

$$\text{current} = (16 \text{ mA}) \left(\frac{x}{100\%} \right) + (4 \text{ mA})$$

This equation form is identical to the one used to calculate pneumatic instrument signal pressures (the 3 to 15 PSI standard):

$$\text{pressure} = (12 \text{ PSI}) \left(\frac{x}{100\%} \right) + (3 \text{ PSI})$$

The same mathematical relationship holds for *any* linear measurement range. Given a percentage of range x, the measured variable is equal to:

$$\text{measured variable} = (\text{Span}) \left(\frac{x}{100\%} \right) + (\text{LRV})$$

Some practical examples of calculations between milliamp current values and process variable values follow:

13.2.1 Example calculation: controller output to valve

An electronic loop controller outputs a signal of 8.55 mA to a direct-responding control valve (where 4 mA is shut and 20 mA is wide open). How far open should the control valve be at this MV signal level?

We must convert the milliamp signal value into a percentage of valve travel. This means determining the percentage value of the 8.55 mA signal on the 4-20 mA range. First, we need to manipulate the percentage-milliamp formula to solve for percentage (x):

$$(16 \text{ mA}) \left(\frac{x}{100\%} \right) + (4 \text{ mA}) = \text{ current}$$

$$(16 \text{ mA}) \left(\frac{x}{100\%} \right) = \text{ current} - (4 \text{ mA})$$

$$\frac{x}{100\%} = \frac{\text{current} - (4 \text{ mA})}{(16 \text{ mA})}$$

$$x = \left(\frac{\text{current} - (4 \text{ mA})}{(16 \text{ mA})} \right) 100\%$$

Next, we plug in the 8.55 mA signal value and solve for x:

$$x = \left(\frac{8.55 \text{ mA} - (4 \text{ mA})}{(16 \text{ mA})} \right) 100\%$$

$$x = 28.4\%$$

Therefore, the control valve should be 28.4 % open when the MV signal is at a value of 8.55 mA.

13.2.2 Example calculation: flow transmitter

A flow transmitter is ranged 0 to 350 gallons per minute, 4-20 mA output, direct-responding. Calculate the current signal value at a flow rate of 204 GPM.

First, we convert the flow value of 204 GPM into a percentage of range. This is a simple matter of division, since the flow measurement range is zero-based:

$$\frac{204 \text{ GPM}}{350 \text{ GPM}} = 0.583 = 58.3\%$$

Next, we take this percentage value and translate it into a milliamp value using the formula previously shown:

$$(16 \text{ mA}) \left(\frac{x}{100\%} \right) + (4 \text{ mA}) = \text{ current}$$

$$(16 \text{ mA}) \left(\frac{58.3\%}{100\%} \right) + (4 \text{ mA}) = 13.3 \text{ mA}$$

Therefore, the transmitter should output a PV signal of 13.3 mA at a flow rate of 204 GPM.

13.2.3 Example calculation: temperature transmitter

An electronic temperature transmitter is ranged 50 to 140 degrees Fahrenheit and has a 4-20 mA output signal. Calculate the current output by this transmitter if the measured temperature is 79 degrees Fahrenheit.

First, we convert the temperature value of 79 degrees into a percentage of range based on the knowledge of the temperature range span (140 degrees − 50 degrees = 90 degrees) and lower-range value (LRV = 50 degrees). We may do so by manipulating the general formula for any linear measurement to solve for x:

$$\text{measured variable} = (\text{Span})\left(\frac{x}{100\%}\right) + (\text{LRV})$$

$$\text{measured variable} - (\text{LRV}) = (\text{Span})\left(\frac{x}{100\%}\right)$$

$$\frac{\text{measured variable} - (\text{LRV})}{(\text{Span})} = \frac{x}{100\%}$$

$$x = \left(\frac{\text{measured variable} - (\text{LRV})}{(\text{Span})}\right) 100\%$$

$$x = \left(\frac{79^o\text{F} - 50^o\text{F}}{90^o\text{F}}\right) 100\%$$

$$x = 32.2\%$$

Next, we take this percentage value and translate it into a 4-20 mA current value using the formula previously shown:

$$(16\text{ mA})\left(\frac{x}{100\%}\right) + (4\text{ mA}) = \text{ current}$$

$$(16\text{ mA})\left(\frac{32.2\%}{100\%}\right) + (4\text{ mA}) = 9.16\text{ mA}$$

Therefore, the transmitter should output a PV signal of 9.16 at a temperature of 79^o F.

13.2.4 Example calculation: pH transmitter

A pH transmitter has a calibrated range of 4 pH to 10 pH, with a 4-20 mA output signal. Calculate the pH sensed by the transmitter if its output signal is 11.3 mA.

First, we must convert the milliamp value into a percentage. Following the same technique we used for the control valve problem:

$$\left(\frac{\text{current} - (4\text{ mA})}{(16\text{ mA})}\right) 100\% = \text{ percent of range}$$

$$\left(\frac{11.3\text{ mA} - (4\text{ mA})}{(16\text{ mA})}\right) 100\% = 0.456 = 45.6\%$$

Next, we take this percentage value and translate it into a pH value, given the transmitter's measurement span of 6 pH (10 pH − 4 pH) and offset of 4 pH:

$$(10\text{ pH} - 4\text{ pH})\left(\frac{x}{100\%}\right) + (4\text{ pH}) = \text{ pH value}$$

$$(6\text{ pH})\left(\frac{45.6\%}{100\%}\right) + (4\text{ pH}) = 6.74\text{ pH}$$

Therefore, the transmitter's 11.3 mA output signal reflects a measured pH value of 6.74 pH.

13.2.5 Example calculation: reverse-acting I/P transducer signal

A current-to-pressure transducer is used to convert a 4-20 mA electronic signal into a 3-15 PSI pneumatic signal. This particular transducer is configured for **reverse action** instead of direct, meaning that its pressure output at 4 mA should be 15 PSI and its pressure output at 20 mA should be 3 PSI. Calculate the necessary current signal value to produce an output pressure of 12.7 PSI.

Reverse-acting instruments are still linear, and therefore still follow the slope-intercept line formula $y = mx + b$. The only differences are a negative slope and a different intercept value. Instead of $y = 16x + 4$ as is the case for direct-acting instruments, this reverse-acting instrument follows the linear equation $y = -16x + 20$:

$$(-16 \text{ mA}) \left(\frac{x}{100\%} \right) + (20 \text{ mA}) = \text{ current}$$

First, we need to to convert the pressure signal value of 12.7 PSI into a percentage of 3-15 PSI range. We will manipulate the percentage-pressure formula to solve for x:

$$(12 \text{ PSI}) \left(\frac{x}{100\%} \right) + (3 \text{ PSI}) = \text{ pressure}$$

$$(12 \text{ PSI}) \left(\frac{x}{100\%} \right) = \text{ pressure} - (3 \text{ PSI})$$

$$\frac{x}{100\%} = \frac{\text{pressure} - (3 \text{ PSI})}{(12 \text{ PSI})}$$

$$x = \left(\frac{\text{pressure} - (3 \text{ PSI})}{(12 \text{ PSI})} \right) 100\%$$

Next, we plug in the 12.7 PSI signal value and solve for x:

$$x = \left(\frac{12.7 \text{ PSI} - (3 \text{ PSI})}{(12 \text{ PSI})} \right) 100\%$$

$$x = 80.8\%$$

This tells us that 12.7 PSI represents 80.8 % of the 3-15 PSI signal range. Plugging this percentage value into our modified (negative-slope) percentage-current formula will tell us how much current is necessary to generate this 12.7 PSI pneumatic output:

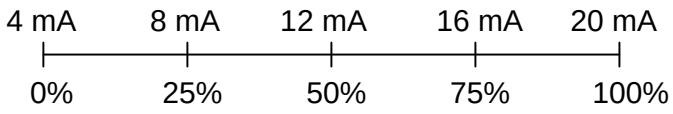
$$(-16 \text{ mA}) \left(\frac{x}{100\%} \right) + (20 \text{ mA}) = \text{ current}$$

$$(-16 \text{ mA}) \left(\frac{80.8\%}{100\%} \right) + (20 \text{ mA}) = 7.07 \text{ mA}$$

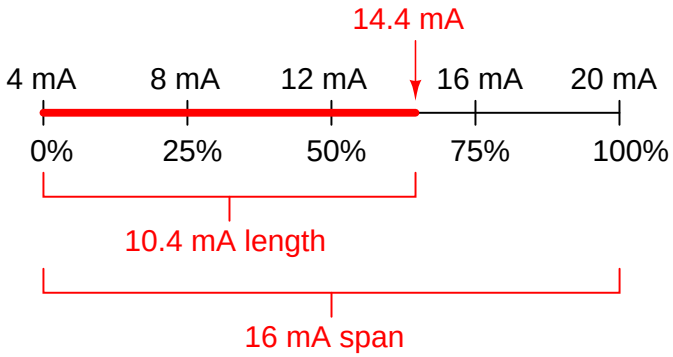
Therefore, a current signal of 7.07 mA is necessary to drive the output of this reverse-acting I/P transducer to a pressure of 12.7 PSI.

13.2.6 Graphical interpretation of signal ranges

A helpful illustration for students in understanding analog signal ranges is to consider the signal range to be expressed as a length on a number line. For example, the common 4-20 mA analog current signal range would appear as such:



If one were to ask the percentage corresponding to a 14.4 mA signal on a 4-20 mA range, it would be as simple as determining the length of a line segment stretching from the 4 mA mark to the 14.4 mA mark:

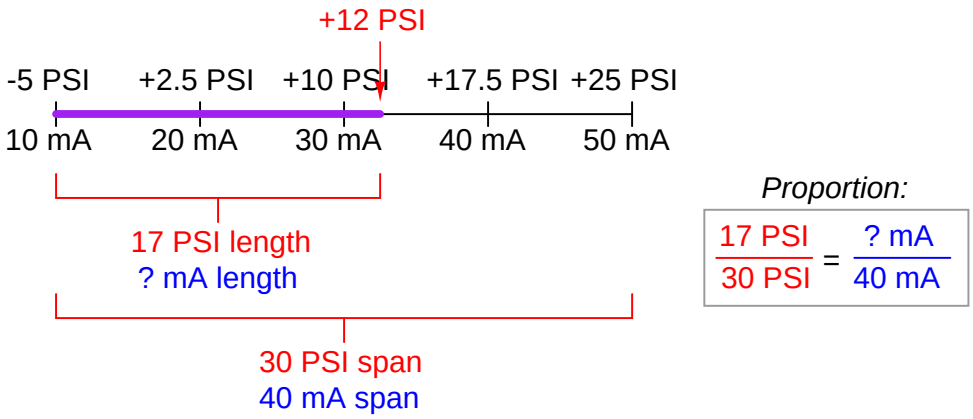


As a percentage, this thick line is 10.4 mA long (the distance between 14.4 mA and 4 mA) over a total (possible) length of 16 mA (the total span between 20 mA and 4 mA). Thus:

$$\text{Percentage} = \left(\frac{14.4 \text{ mA} - 4 \text{ mA}}{20 \text{ mA} - 4 \text{ mA}} \right) 100\%$$

$$\text{Percentage} = 65\%$$

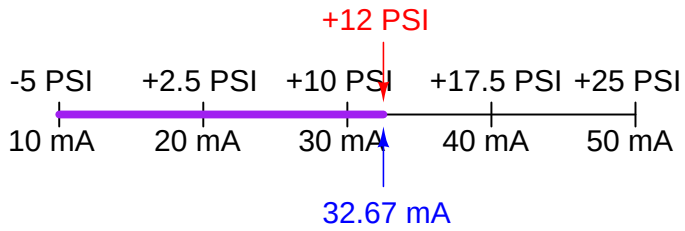
This same "number line" approach may be used to visualize any conversion from one analog scale to another. Consider the case of an electronic pressure transmitter calibrated to a pressure range of -5 to +25 PSI, having an (obsolete) current signal output range of 10 to 50 mA. The appropriate current signal value for an applied pressure of +12 PSI would be represented on the number line as such:



Finding the "length" of this line segment in units of milliamps is as simple as setting up a proportion between the length of the line in units of PSI over the total (span) in PSI, to the length of the line in units of mA over the total (span) in mA:

$$\frac{17 \text{ PSI}}{30 \text{ PSI}} = \frac{? \text{ mA}}{40 \text{ mA}}$$

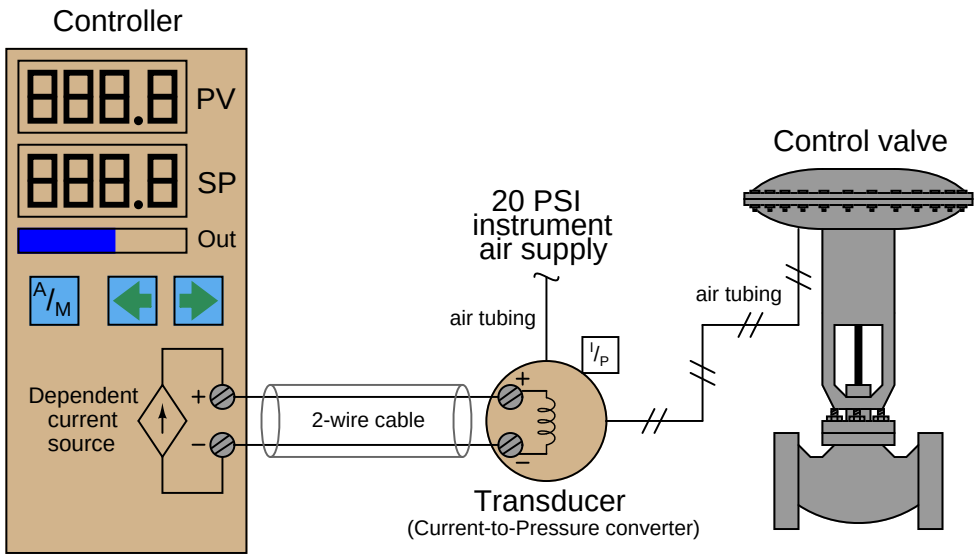
Solving for the unknown (?) current by cross-multiplication and division yields a value of 22.67 mA. Of course, this value of 22.67 mA only tells us the length of the line segment on the number line; it does not directly tell us the current signal value. To find that, we must add the "live zero" offset of 10 mA, for a final result of 32.67 mA.



Thus, an applied pressure of +12 PSI to this transmitter should result in a 32.67 mA output signal.

13.3 Controller output current loops

The simplest form of 4-20 mA current loop is the type used to represent the output of a process controller, sending a command signal to a final control element. Here, the controller both supplies the electrical power and regulates the DC current to the final control element, which acts as an electrical load. To illustrate, consider the example of a controller sending a 4-20 mA signal to an I/P (current-to-pressure) signal converter, which then pneumatically drives a control valve:



This particular controller has two digital displays, one for process variable (PV) and one for setpoint (SP), with a bargraph for displaying the output value (Out). One pushbutton provides the operator with a way to switch between Automatic and Manual modes (A/M), while two other pushbuttons provide means to decrement and increment either the setpoint value (in Automatic mode) or the Output value (in Manual mode).

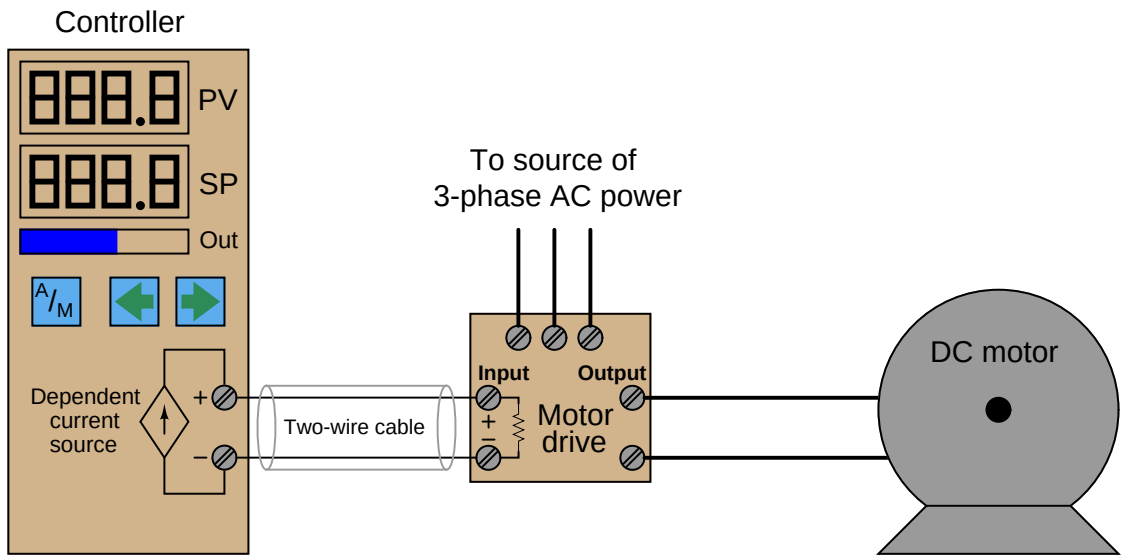
Inside the controller, a *dependent current source* provides the 4-20 mA DC current signal to the I/P transducer. Like all current sources, its purpose is to maintain current in the "loop" circuit regardless of circuit resistance or any external voltage sources. Unlike a constant current source, a "dependent" current source (represented by a diamond shape instead of a circle shape) varies its current value according to the dictates of some external stimulus. In this case, either the mathematical function of the controller (Automatic mode) or the arbitrary setting of the human operator (Manual mode) tells the current source how much DC current it should maintain in the circuit.

For example, if the operator happened to switch the controller into Manual mode and set the output value at 50%, the proper amount of DC current for this signal percentage would be 12 mA (exactly half-way between 4 mA and 20 mA). If everything is working properly, the current in the "loop" circuit to the I/P transducer should remain exactly at 12 mA regardless of slight changes in wire resistance, I/P coil resistance, or anything else: the current source inside the controller will "fight" as hard as it has to in order to maintain this set amount of current. This current, as it flows

through the wire coil of the I/P transducer mechanism, creates a magnetic field inside the I/P to actuate the pneumatic mechanism and produce a 9 PSI pressure signal output to the control valve (9 PSI being exactly half-way between 3 PSI and 15 PSI in the 3-15 PSI signal standard range). This should move the control valve to the half-way position.

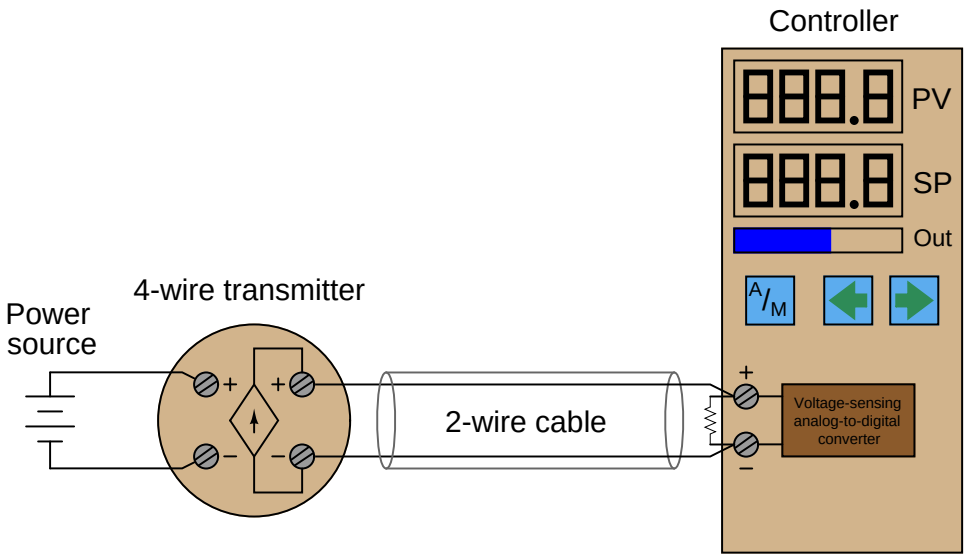
The details of the controller's internal current source are not terribly important. Usually, it takes the form of an operational amplifier circuit driven by the voltage output of a DAC (Digital-to-Analog Converter). The DAC converts a binary number (either from the controller's automatic calculations, or from the human operator's manual setting) into a small DC voltage, which then commands the op-amp circuit to regulate output current at a proportional value.

The scenario is much the same if we replace the I/P and control valve with a variable-speed motor drive. From the controller's perspective, the only difference it sees is a resistive load instead of an inductive load. The input resistance of the motor drive circuit converts the 4-20 mA signal into an analog voltage signal (typically 1-5 V, but not always). This voltage signal then constitutes a command to the rest of the drive circuitry, telling it to modulate the power going to the electric motor in order to drive it at the desired speed:



13.4 4-wire ("self-powered") transmitter current loops

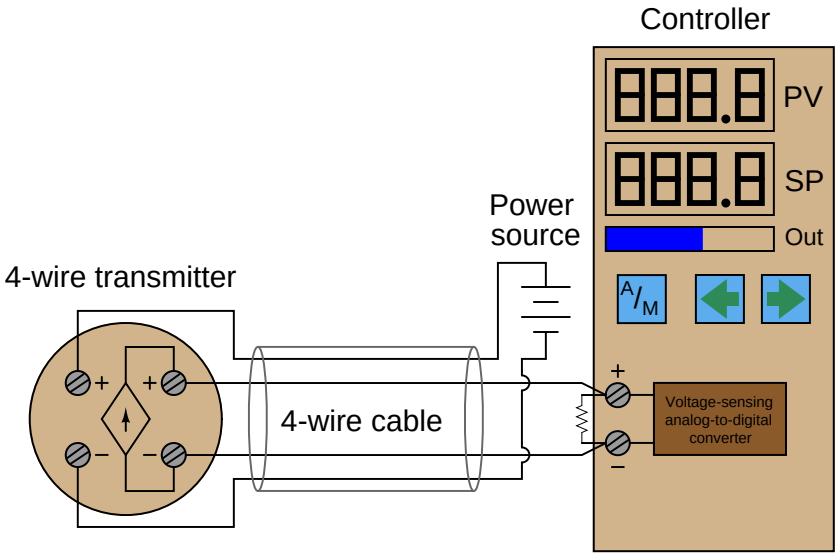
DC electric current signals may also be used to communicate process measurement information from transmitters to controllers, indicators, recorders, alarms, and other input devices. The simplest form of 4-20 mA measurement loop is one where the transmitter has two terminals for the 4-20 mA signal wires to connect, and two more terminals where a power source connects. These transmitters are called "4-wire" or self-powered. The current signal from the transmitter connects to the *process variable input* terminals of the controller to complete the loop:



Typically, process controllers are not equipped to directly accept milliamp input signals, but rather voltage signals. For this reason we must connect a precision resistor across the input terminals to convert the 4-20 mA signal into a standardized analog voltage signal that the controller can understand. A voltage signal range of 1 to 5 volts is standard, although some models of controller use different voltage ranges and therefore require different precision resistor values. If the voltage range is 1-5 volts and the current range is 4-20 mA, the precision resistor value must be 250 ohms.

Since this is a digital controller, the input voltage at the controller terminals is interpreted by an analog-to-digital converter (ADC) circuit, which converts the measured voltage into a digital number that the controller's microprocessor can work with.

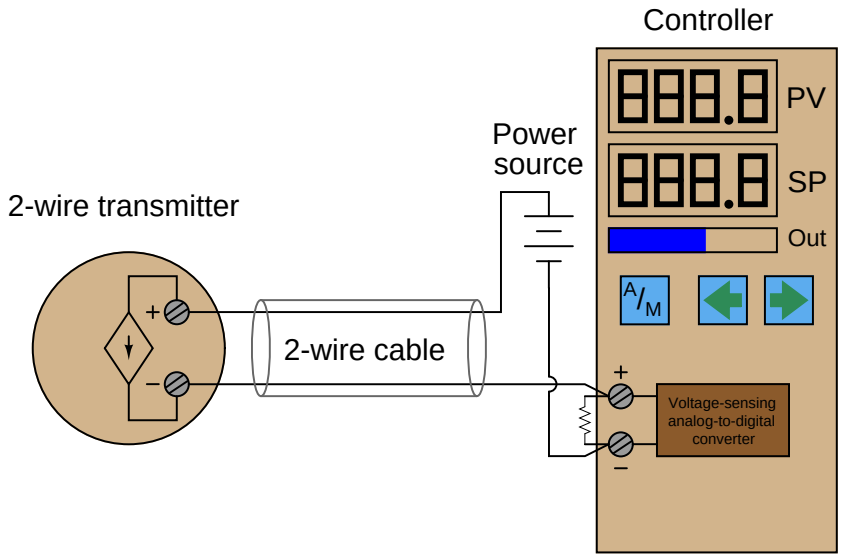
In some installations, transmitter power is supplied through additional wires in the cable from a power source located in the same panel as the controller:



The obvious disadvantage of this scheme is the requirement of two more conductors in the cable. More conductors means the cable will be larger-diameter and more expensive for a given length. Cables with more conductors will require larger electrical conduit to fit in to, and all field wiring panels will have to contain more terminal blocks to marshal the additional conductors. If no suitable electrical power source exists at the transmitter location, though, a 4-wire cable is necessary to service a 4-wire transmitter.

13.5 2-wire ("loop-powered") transmitter current loops

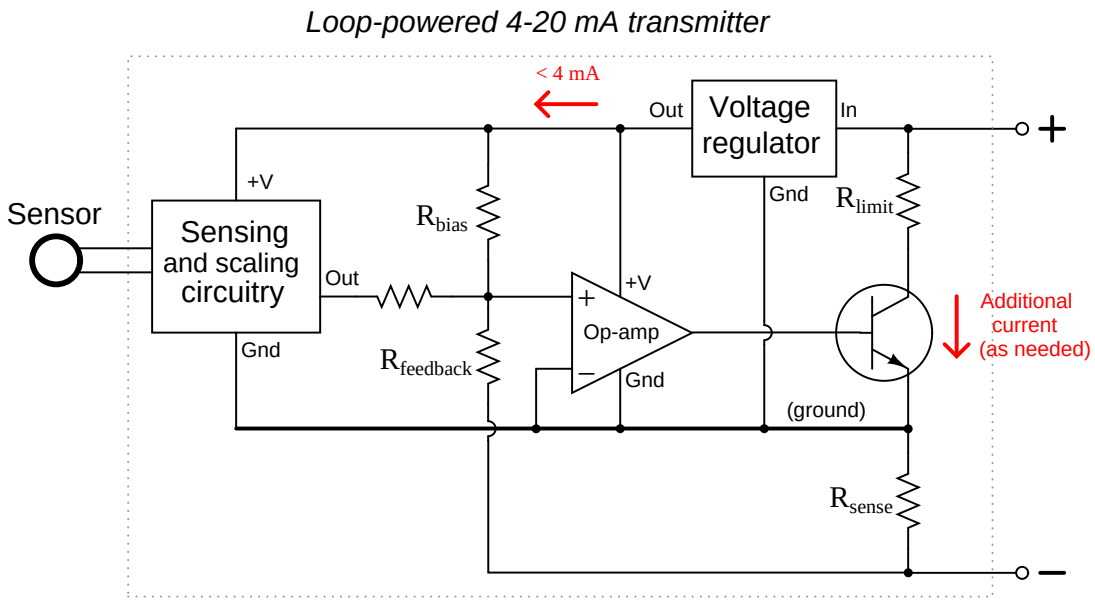
It is possible to convey electrical power *and* communicate analog information over the same two wires using 4 to 20 milliamps DC, if we design the transmitter to be *loop-powered*. A loop-powered transmitter connects to a process controller in the following manner:



Here, the transmitter is not really a current *source* in the sense that a 4-wire transmitter is. Instead, a 2-wire transmitter's circuitry is designed to act as a current *regulator*, limiting current in the series loop to a value representing the process measurement, while relying on a remote source of power to motivate current to flow. Please note the direction of the arrow in the transmitter's dependent current source symbol, and how it relates to the voltage polarity marks. Refer back to the illustration of a 4-wire transmitter circuit for comparison. The current "source" in this loop-powered transmitter actually behaves as an electrical *load*, while the current source in the 4-wire transmitter functions as a true electrical source.

A loop-powered transmitter gets its operating power from the minimum terminal voltage and current available at its two terminals. With the typical source voltage being 24 volts DC, and the maximum voltage dropped across the controller's 250 ohm resistor being 5 volts DC, the transmitter should always have at least 19 volts available at its terminals. Given the lower end of the 4-20 mA signal range, the transmitter should always have at least 4 mA of current to run on. Thus, the transmitter will always have a certain minimum amount of electrical power available on which to operate, while regulating current to signal the process measurement.

Internally, the loop-powered transmitter circuitry looks something like this:



All sensing, scaling, and output conditioning circuitry inside the transmitter must be designed to run on less then 4 mA of DC current, and at a modest terminal voltage. In order to create loop currents exceeding 4 mA – as the transmitter must do in order to span the entire 4 to 20 milliamp signal range – the transmitter circuitry uses a transistor to shunt (bypass) extra current from one terminal to the other as needed to make the total current indicative of the process measurement. For example, if the transmitter's internal operating current is only 3.8 mA, and it must regulate loop current at a value of 16 mA to represent a condition of 75% process measurement, the transistor will bypass 12.2 mA of current.

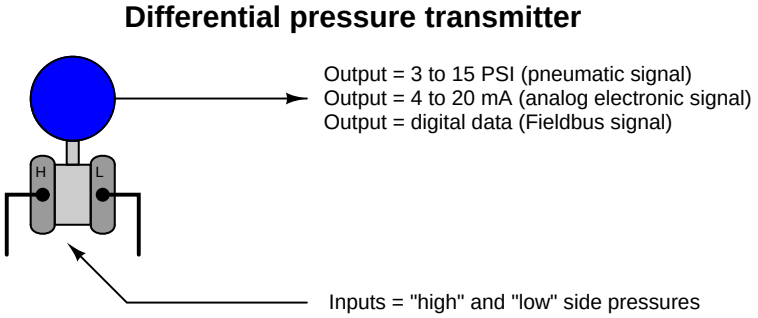
Early current-based industrial transmitters were not capable of operating on such low levels of electrical power, and so used a different current signal standard: 10 to 50 milliamps DC. Loop power supplies for these transmitters ranged upwards of 90 volts to provide enough power for the transmitter. Safety concerns made the 10-50 mA standard unsuitable for some industrial installations, and modern microelectronic circuitry with its reduced power consumption made the 4-20 mA standard practical for nearly all types of process transmitters.

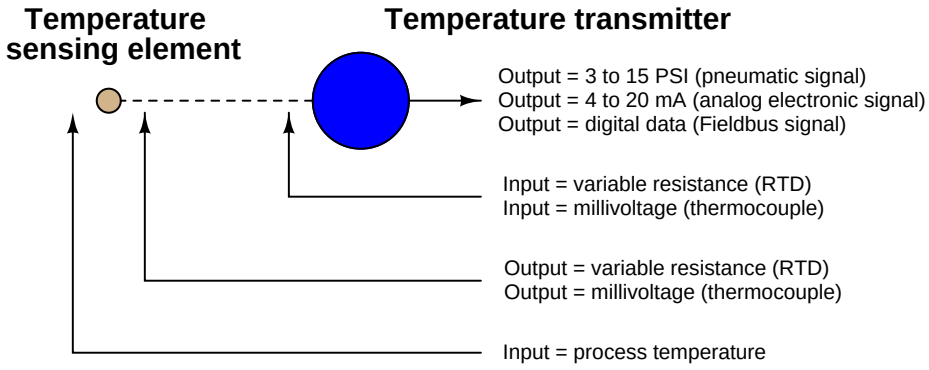
13.6 Troubleshooting current loops

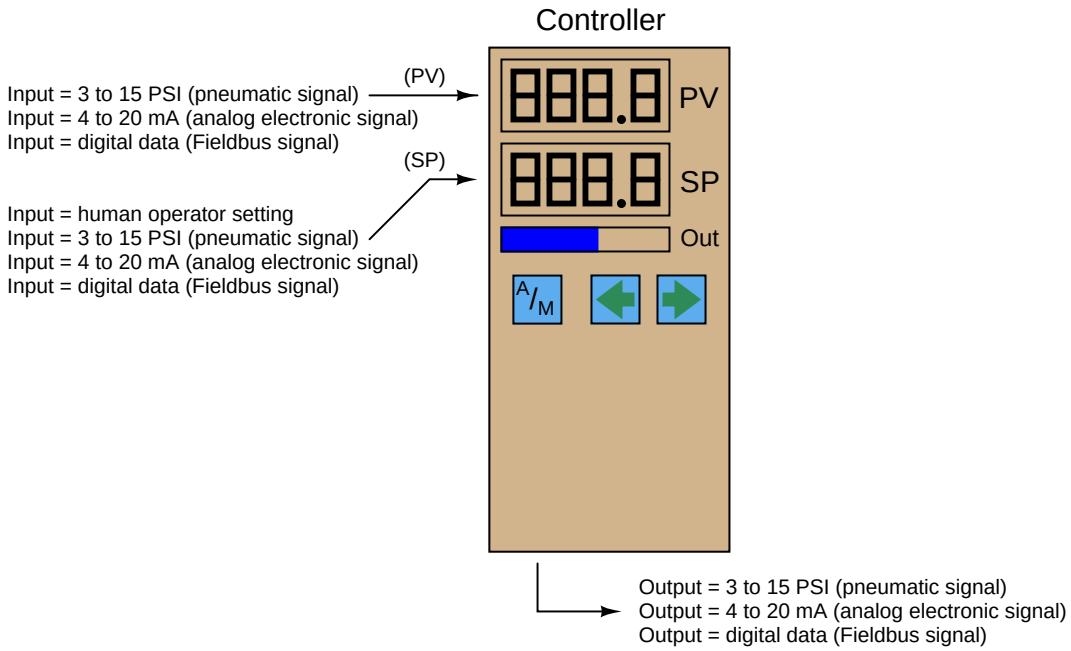
A fundamental principle in instrumentation system troubleshooting is that every instrument has at least one input and at least one output, and that the output(s) should accurately correspond to the input(s). If an instrument's output is not properly corresponding to its input according to the instrument's design function, there must be something wrong with that instrument.

Consider the inputs and outputs of several common instruments: transmitters, controllers, indicators, and control valves. Each of these instruments takes in (input) data in some form, and generates (output) data in some form. In any instrument "loop," the output of one instrument feeds into the input of the next, such that information is passed from one instrument to another. By intercepting the data communicated between components of an instrument system, we are able to locate and isolate faults. In order to properly understand the intercepted data, we must understand the inputs and outputs of the respective instruments and the basic functions of those instruments.

The following illustrations highlight inputs and outputs for instruments commonly found in control systems:







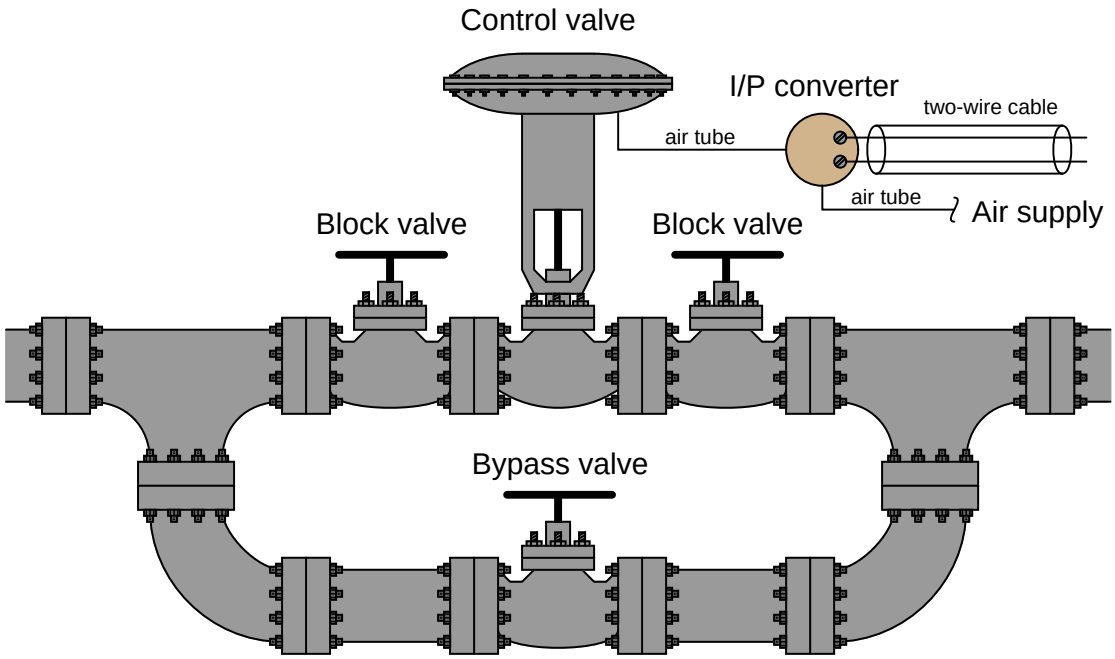
In order to check for proper correspondence between instrument inputs and outputs, we must be able to use appropriate test equipment to intercept the signals going into and out of those instruments. For 4-20 mA analog signal-based instruments, this means we must be able to use electrical meters capable of accurately measuring current and voltage.

13.6.1 Using a standard milliammeter to measure loop current

Since the signal of interest is represented by an electric current in an instrumentation current "loop" circuit, the obvious tool to use for troubleshooting is a multimeter capable of accurately measuring DC milliamperes. Unfortunately, though, there is a major disadvantage to the use of a milliammeter: the circuit must be "broken" at some point to connect the meter in series with the current, and this means the current will fall to 0 mA until the meter is connected (then fall to 0 mA when the meter is removed from the circuit). Interrupting the current means interrupting the flow of information conveyed by that current, be it a process measurement or a command signal to a final control element. This *will* have adverse effects on a control system unless certain preparatory steps are taken.

Before "breaking the loop" to connect your meter, one must first warn all appropriate personnel that the signal will be interrupted at least twice, falling to a value of -25% each time. If the signal to be interrupted is coming from a process transmitter to a controller, the controller should be placed in Manual mode so it will not cause an upset in the process (by moving the final control element in response to the sudden loss of PV signal). Also, process alarms should be temporarily disabled so they do not cause panic. If this current signal also drives process shutdown alarms, these should be temporarily disabled so that nothing shuts down upon interruption of the signal.

If the current signal to be interrupted is a command signal from a controller to a final control element, the final control element either needs to be manually overridden so as to hold a fixed setting while the signal varies, or it needs to be bypasses completely by some other device(s). If the final control element is a control valve, this typically takes the form of opening a bypass valve and closing at least one block valve:



Since the manually-operated bypass valve now performs the job that the automatic control valve used to, a human operator must remain posted at the bypass valve to carefully throttle it and maintain control of the process.

Block and bypass valves for a large gas flow control valve may be seen in the following photograph:

In consideration of the labor necessary to safely interrupt the current signal to a control valve in a live process, we see that the seemingly simple task of connecting a milliammeter in series with a 4-20 mA current signal is not as easy as it may first appear. Better ways must exist, no?

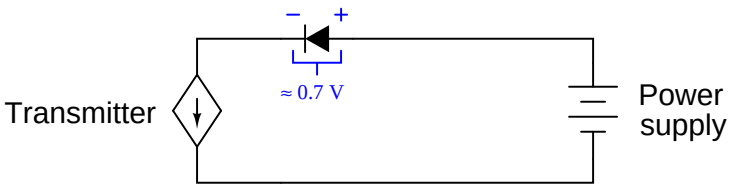
13.6.2 Using a clamp-on milliammeter to measure loop current

One better way to measure a 4-20 mA signal without interrupting it is to do so magnetically, using a clamp-on milliammeter. Modern Hall-effect sensors are sensitive and accurate enough to monitor the weak magnetic fields created by the passage of small DC currents in wires. Ammeters using Hall-effect sensors have are completely non-intrusive because they merely clamp around the wire, with no need to "break" the circuit. An example of a such a clamp-on current meter is the Fluke model 771, shown in this photograph:

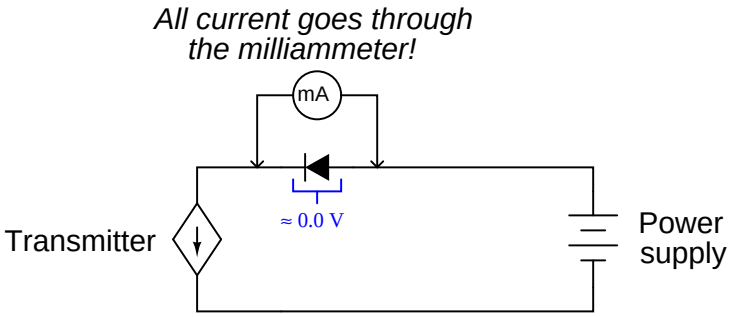
Note how this milliammeter not only registers loop current (3.98 mA as shown in the photograph), but it also converts the milliamp value into a percentage of range, following the 4 to 20 mA signal standard. One disadvantage to be aware of for clamp-on milliammeters is the susceptibility to error from strong external magnetic fields. Steady magnetic fields (from permanent magnets or DC-powered electromagnets) may be compensated for by performing a "zero" adjustment with the instrument held in a similar orientation prior to measuring loop current through a wire.

13.6.3 Using "test" diodes to measure loop current

Another way to measure a 4-20 mA signal without interrupting it involves the use of a rectifying diode, originally installed in the loop circuit when it was commissioned. A "test" diode may be placed anywhere in series within the loop in such a way that it will be forward-biased. During normal operation, the diode will drop approximately 0.7 volts, as is typical for any silicon rectifying diode when forward biased. The following schematic diagram shows such a diode installed in a 2-wire transmitter loop circuit:



If someone connects a milliammeter in parallel with this diode, however, the very low input resistance of the ammeters "shorts past" the diode and prevents any substantial voltage drop from forming across it. Without the necessary forward voltage drop, the diode effectively turns off and conducts 0 mA, leaving the entire loop current to pass through the ammeter:



When the milliammeter is disconnected, the requisite 0.7 volt drop appears to turn on the diode, and all loop current flows through the diode again. At no time is the loop current ever interrupted, which means a technician may take current measurements this way and never have to worry about generating false process variable indications, setting off alarms, or upsetting the process.

Such a diode may be installed at the nearest junction box, between terminals on a terminal strip, or even incorporated into the transmitter itself. Some process transmitters have an extra pair of terminals labeled "Test" for this exact purpose. A diode is already installed in the transmitter, and these "test" terminals serve as points to connect the milliammeter across.

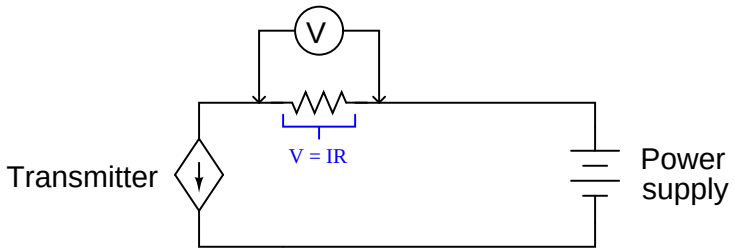
The following photograph shows an example of this on a Rosemount model 3051 differential pressure transmitter:

Note the two test points labeled "TEST" below and to the right of the main screw terminals where the loop wiring attaches. Connecting an ammeter to these two test points allows for direct measurement of the 4-20 mA current signal without having to un-do any wire connections in the circuit.

Transmitters equipped with analog meter movements for direct visual indication of the 4-20 mA signal usually connect the analog milliammeter in parallel with just such a diode. The reason for doing this is to maintain loop continuity in the event that the fine-wire coil inside the milliammeter movement were to accidently break open.

13.6.4 Using shunt resistors to measure loop current

A similar method for non-invasively measuring current in a 4-20 mA instrumentation circuit is to install a precision resistor in series. If the resistance value is precisely known, the technician merely needs to measure voltage across it with a voltmeter and use Ohm's Law to calculate current:



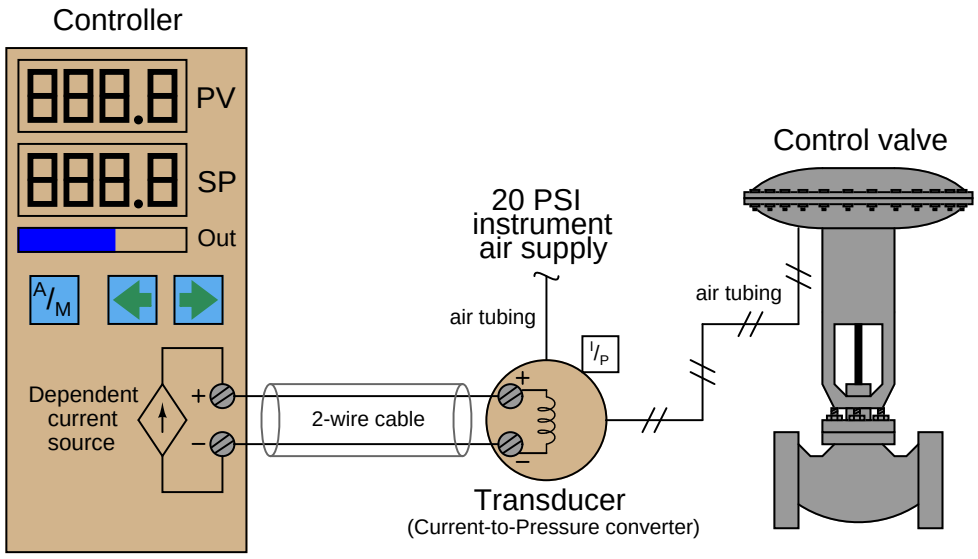
In electronics, such a precision resistor used for measuring current is often referred to as a *shunt* resistor. Shunt resistor values are commonly very small, for their purpose is to assist in current measurement without imposing undue voltage drop within a circuit. It is rare to find a 250 ohm resistor used strictly as a diagnostic shunt resistor, because the extra voltage drop (1 to 5 volts, depending on the current signal level) may "starve" loop-powered instruments of voltage necessary to operate. Shunt resistor values as low as 1 ohm may be found installed in 4-20 mA current loops at strategic locations where technicians may need to measure loop current[2].

[2]Of course, a 1 ohm resistor would drop 4 mV at 4 mA loop current, and drop 20 mV at 20 mA loop current. These small voltage values necessitate a highly accurate DC voltmeter for field measurement!

13.6.5 Troubleshooting current loops with voltage measurements

If neither component (diode nor shunt resistor) is pre-installed in the circuit, and if a Hall-effect (clamp-on) precision milliammeter is unavailable, a technician may still perform useful troubleshooting measurements using nothing but a DC voltmeter. Here, however, one must be careful of how to interpret these voltage measurements, for they may not directly correspond to the loop current as was the case with measurements taken in parallel with the precision resistor.

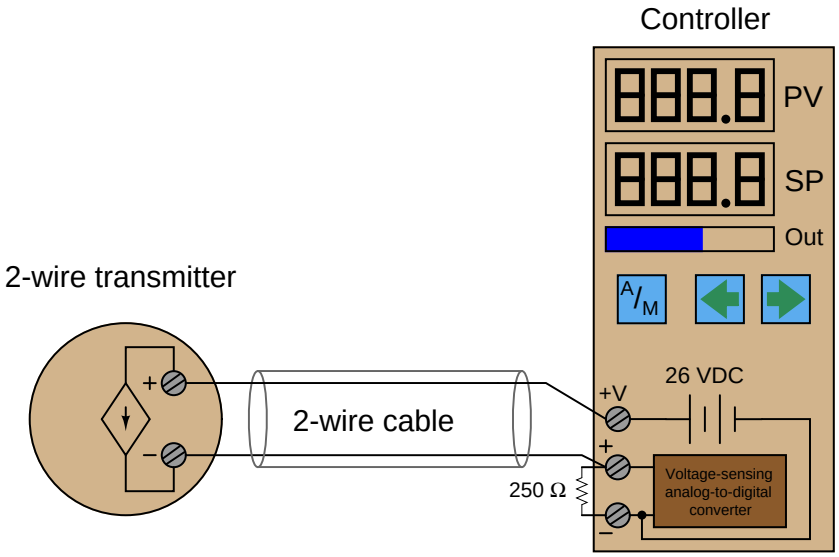
Take for example this 4-20 mA loop where a controller sends a command signal to an I/P transducer:



There is no standardized resistance value for I/P transducer coils, and so the amount of voltage dropped across the I/P terminals for any given amount of loop current will be unique for every different model of I/P. The Fisher model 567 I/P transducer built for 4-20 mA signals has a nominal coil resistance of 176 ohms. Thus, we would expect to see a voltage drop of approximately 0.7 volts at 4 mA and a drop of approximately 3.5 volts at 20 mA across the I/P terminals. Since the controller output terminals are directly in parallel with the I/P terminals, we would expect to see approximately the same voltage there as well (slightly greater due to wire resistance). The lack of known precision in the I/P coil resistance makes it difficult to tell exactly how much current is in the loop for any given voltage measurement we take with a voltmeter. However, if we do know the approximate coil resistance of the I/P, we can at least obtain an estimate of loop current, which is usually good enough for diagnostic purposes.

If the I/P coil resistance is completely unknown, voltage measurements become useless for quantitative determination of loop current. Voltage measurements would be useful only for qualitatively determining loop continuity (i.e. whether there is a break in the wiring between the controller and I/P).

Another example for consideration is this loop-powered 4-20 mA transmitter and controller circuit, where the controller supplies DC power for the loop:



It is very common to find controllers with their own built-in loop power supplies, due to the popularity of loop-powered (2-wire) 4-20 mA transmitters. If we know the transmitter requires a DC voltage source somewhere in the circuit to power it up, it makes sense to include one in the controller, right?

The only voltage measurement that directly and accurately correlates to loop current is the voltage directly across the 250 ohm precision resistor. A loop current of 4 mA will yield a voltage drop of 1 volt, 12 mA will drop 3 volts, 20 mA will drop 5 volts, etc.

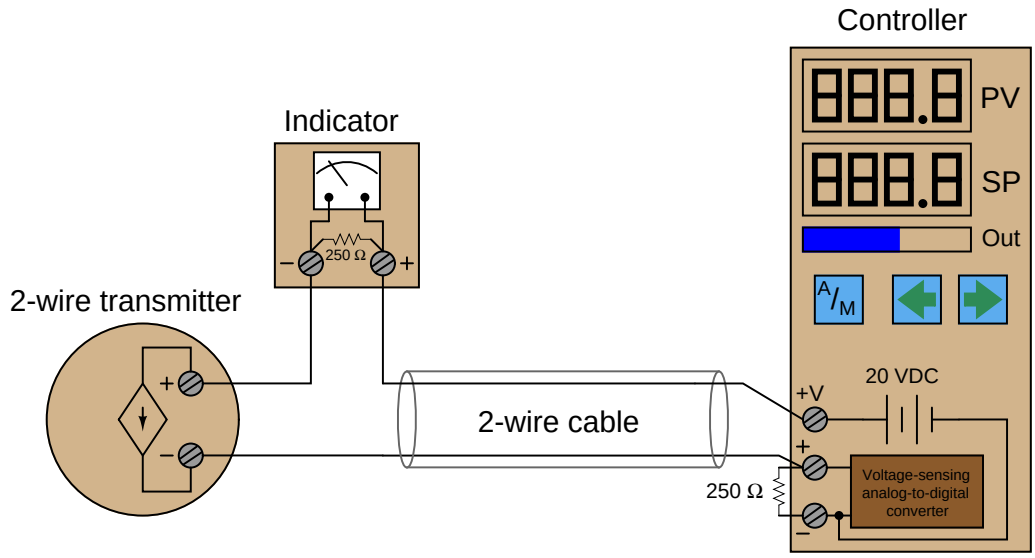
A voltage measurement across the transmitter terminals will show us the *difference* in voltage between the 26 volt power supply and the voltage dropped across the 250 ohm resistor. In other words, the transmitter's terminal voltage is simply what is left over from the source voltage of 26 volts after subtracting the resistor's voltage drop. This makes the transmitter terminal voltage inversely proportional to loop current: the transmitter sees approximately 25 volts at 4 mA loop current (0% signal) and approximately 21 volts at 20 mA loop current (100% signal).

The use of the word "approximate" is very intentional here, for loop power supplies are usually non-regulated. In other words, the "26 volt" rating is approximate and subject to change! One of the advantages of the loop-powered transmitter circuit is that the source voltage is largely irrelevant, so long as it exceeds the minimum value necessary to ensure adequate power to the transmitter. If the source voltage drifts for any reason, it will have no impact on the measurement signal at all, because the transmitter is built as a *current regulator*, regulating current in the loop to whatever value represents the process measurement, regardless of slight changes in loop source voltage, wire resistance, etc. This rejection of power supply voltage changes means that the loop power supply need not be regulated, and so in practice it rarely is.

This brings us to a common problem in loop-powered 4-20 mA transmitter circuits: maintaining sufficient operating voltage at the transmitter terminals. Recall that a loop-powered transmitter

relies on the voltage dropped across its terminals (combined with a current of less than 4 mA) to power its internal workings. This means the terminal voltage must not be allowed to dip below a certain minimum value, or else the transmitter will not have enough electrical power to continue its normal operation. This makes it possible to "starve" the transmitter of voltage if the loop power supply voltage is insufficient, and/or if the loop resistance is excessive.

To illustrate how this can be a problem, consider the following 4-20 mA measurement loop, where the controller supplies only 20 volts DC to power the loop, and an indicator is included in the circuit to provide operators with field-located indication of the transmitter's measurement:



The indicator contains its own 250 ohm resistor to provide a 1-5 volt signal for the meter mechanism to sense. This means the total loop resistance is now 500 ohms (plus any wire resistance). At full current (20 mA), this total resistance will drop (at least) 10 volts, leaving 10 volts or less at the transmitter terminals to power the transmitter's internal workings. 10 volts may not be enough for the transmitter to successfully operate, though. The Rosemount model 3051 pressure transmitter, for example, requires a minimum of 10.5 volts at the terminals to operate.

However, the transmitter *will* operate just fine at lower loop current levels. When the loop current is only 4 mA, for example, the combined voltage drop across the two 250 ohm resistors will be only 2 volts, leaving about 18 volts at the transmitter terminals: more than enough for practically any model of 4-20 mA loop-powered transmitter to successfully operate. Thus, the problem of insufficient supply voltage only manifests itself when the process measurement nears 100% of range. This could be a difficult problem to diagnose, since it appears only during certain process conditions. A technician looking only for wiring faults (loose connections, corroded terminals, etc.) would never find the problem.

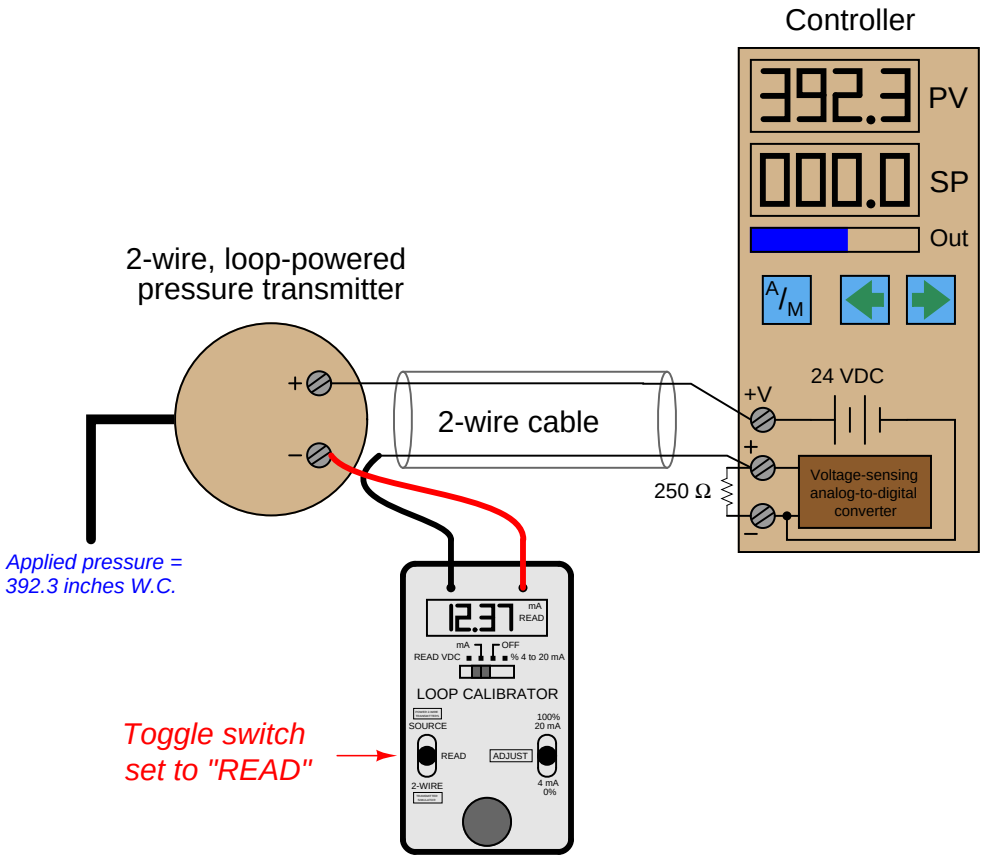
When a loop-powered transmitter is starved of voltage, its behavior becomes erratic. This is especially true of "smart" transmitters with built-in microprocessor circuitry. If the terminal voltage dips below the required minimum, the microprocessor circuit shuts down. When the circuit shuts down, the current draw decreases accordingly. This causes the terminal voltage to rise again, at

which point the microprocessor has enough voltage to start up. As the microprocessor "boots" back up again, it increases loop current to reflect the near-100% process measurement. This causes the terminal voltage to sag, which subsequently causes the microprocessor to shut down again. The result is a slow on/off cycling of the transmitter's current, which makes the process controller think the process variable is surging wildly. The problem disappears, though, as soon as the process measurement decreases enough that the transmitter is allowed enough terminal voltage to operate normally.

13.6.6 Using loop calibrators

Special-purpose electronic test instruments called *loop calibrators* are manufactured for the express purpose of 4-20 mA current loop circuit troubleshooting. These versatile instruments are generally capable of not only measuring current, but also *sourcing* current to unpowered devices in a loop, and also *simulating* the operation of loop-powered 4-20 mA transmitters.

A very popular loop calibrator unit is the Altek model 334A, a battery-powered, hand-held unit with a rotary knob for current adjustment and toggle switches for mode setting. The following illustration shows how this calibrator would be used to measure current in a functioning input signal loop[3]:

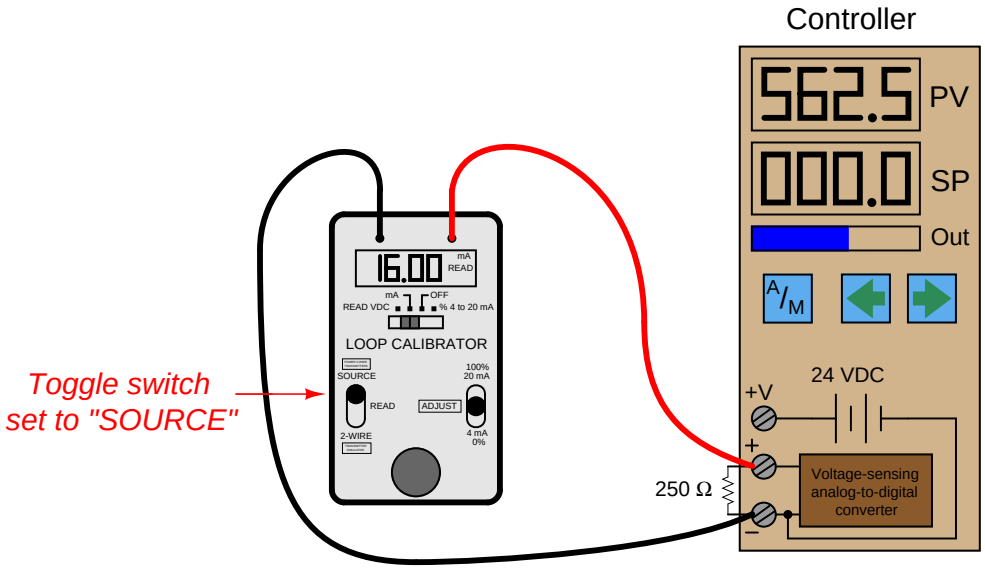


Here, the loop wiring is broken at the negative terminal of the loop-powered transmitter, and the calibrator connected in series to measure current. If this loop had a test diode installed, the calibrator could be connected in parallel with the diode to achieve the same function. Note the polarity of the calibrator's test leads in relation to the circuit being tested: the calibrator is acting

[3] In the following illustrated examples, the transmitter is assumed to be a pressure transmitter with a calibrated range of 0 to 750 inches of water column, 4-20 mA. The controller's PV (process variable) display is appropriately ranged to display 0 to 750 as well.

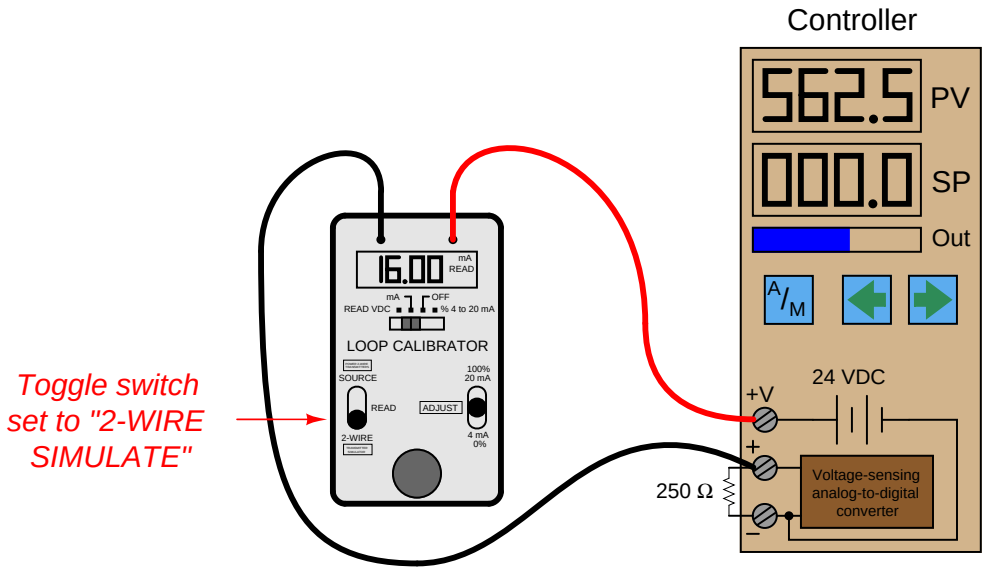
as an unpowered device (a load rather than a source), with the more positive loop terminal connected to the calibrator's red test lead and the more negative terminal connected to the black test lead.

The same loop calibrator may be used to *source* (or *drive*) a 4-20 mA signal into an indicating instrument to test the function of that instrument independently. Here, we see the Altek calibrator used as a current source to send a 16.00 mA signal to the PV (process variable) input of the controller:



No transmitter need be included in this illustration, because the calibrator takes its place. Note how the calibrator is used here as an active *source* of current rather than a passive load as it was in the last example. The calibrator's red test lead connects to the controller's positive input terminal, while the black test lead connects to the negative terminal. The DC power source inside the controller is not used for loop power, because the calibrator in "source" mode provides the necessary power to drive current through the 250 ohm resistor.

An alternative method of sourcing a known current signal into an indicating instrument that provides loop power is to set the loop calibrator to a mode where it mimics the electrical behavior of a loop-powered 2-wire transmitter. In this mode, the calibrator serves to regulate loop current at a user-determined value, but it provides no motivating voltage to drive this current. Instead, it passively relies on some external voltage source in the loop circuit to provide the necessary electromotive force:



Note the polarity of the calibrator's test leads in relation to the controller: the red test lead connects to the positive loop power terminal while the black lead connects to the positive input terminal. Here, the calibrator acts as a load, just as a loop-powered transmitter acts as an electrical load. The only source of electrical power in this test circuit is the 24 VDC source inside the controller: the same one normally providing energy to the circuit when a loop-powered transmitter is connected.

This *simulate transmitter* mode is especially useful for testing a 4-20 mA loop at the end of the cable where the transmitter is physically located. After disconnecting the cable wires from the transmitter and re-connecting them to the loop calibrator (set to "simulate" mode), the calibrator may be used to simulate a transmitter measuring any value within its calibrated range.

A legacy loop calibrator still familiar to many instrument technicians at the time of this writing is the classic Transmation model 1040:

Other examples of vintage loop calibrator technology include the Nassau model 8060 (left) and the Biddle Versa-Cal (right):

A modern loop calibrator manufactured by Fluke is the model 705:

With this calibrator, the *measure*, *source*, and *simulate* modes are accessed by repeatedly pushing a button, with the current mode displayed on the screen:

13.6.7 NAMUR signal levels

Signal level	Fault condition
Output $\leq$ 3.6 mA	Sensing transducer failed low
3.6 mA $<$ Output $<$ 3.8 mA	Sensing transducer failed (detected) low
3.8 mA $\leq$ Output $<$ 4.0 mA	Measurement under-range
21.0 $>$ Output $\geq$ 20.5 mA	Measurement over-range
Output $\geq$ 21.0 mA	Sensing transducer failed high

4-20 mA process transmitters compliant with the NAMUR recommendations for fault signal levels will limit their output signals between 3.8 mA and less than 21 mA when functioning properly. Signals lying outside this range indicate some form of failure has occurred within the transmitter[4].

Any control system programmed to respond to these specific fault-induced current levels may act upon them by forcing controllers into manual mode, initiating shutdown procedures, or taking some other form of safe action appropriate to the knowledge of a failed process transmitter.

References

Lipták, Béla G., *Instrument Engineers' Handbook – Process Software and Digital Networks*, Third Edition, CRC Press, New York, NY, 2002.

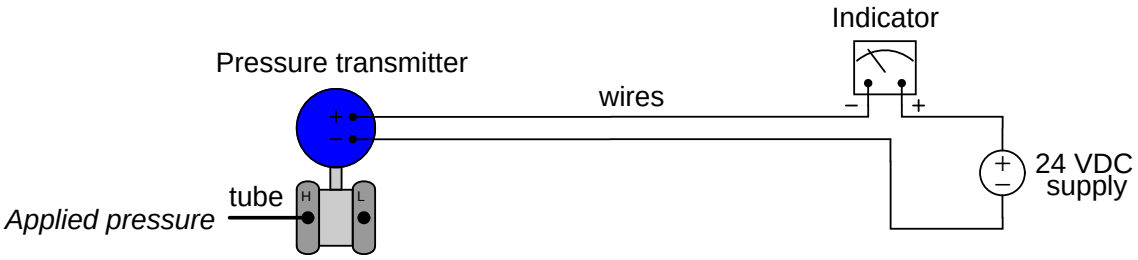
"NAMUR" whitepaper, Emerson Process Management, 2007.

[4]Of course, an open or shorted fault in the wiring connecting the transmitter with the rest of the system could easily force the current beyond this range, through no fault of the transmitter itself.

Chapter 14

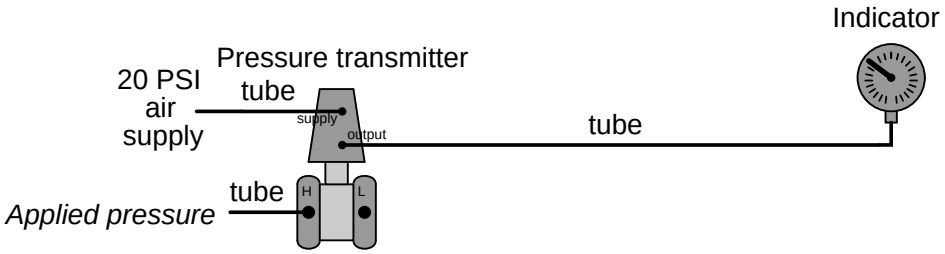
Pneumatic instrumentation

While electricity is commonly used as a medium for transferring energy across long distances, it is also used in instrumentation to transfer *information*. A simple 4-20 mA current "loop" uses direct current to represent a process measurement in percentage of span, such as in this example:



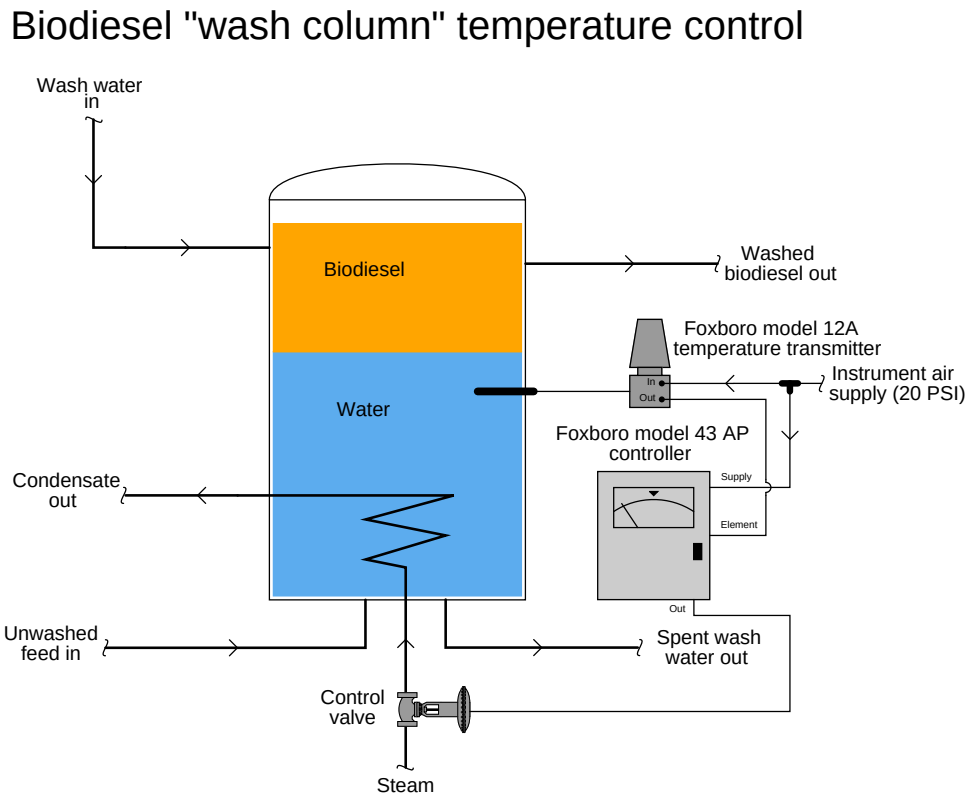
The transmitter senses an applied fluid pressure from the process being measured, regulates electric current in the series circuit according to its calibration (4 mA = no pressure ; 20 mA = full pressure), and the indicator (ammeter) registers this measurement on a scale calibrated to read in pressure units. If the calibrated range of the pressure transmitter is 0 to 250 PSI, then the indicator's scale will be labeled to read from 0 to 250 PSI as well. No human operator reading that scale need worry about how the measurement gets from the process to the indicator – the 4-20 mA signal medium is transparent to the end-user as it should be.

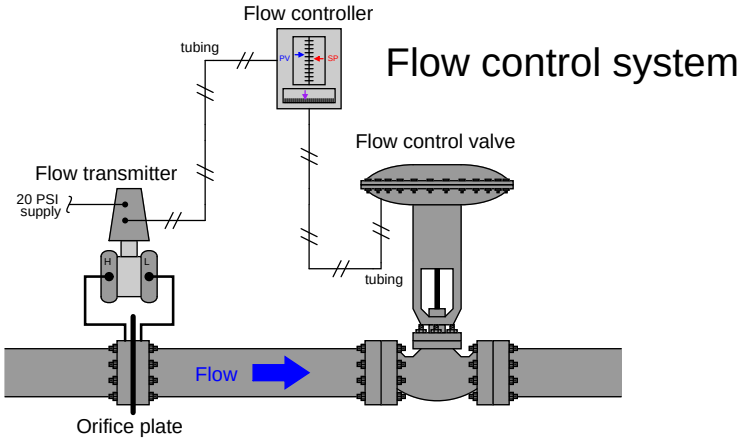
Air pressure may be used as an alternative signaling medium to electricity. Imagine a pressure transmitter designed to output a variable air pressure according to its calibration rather than a variable electric current. Such a transmitter would have to be supplied with a source of constant-pressure compressed air instead of an electric voltage, and the resulting output signal would be conveyed to the indicator via tubing instead of wires:



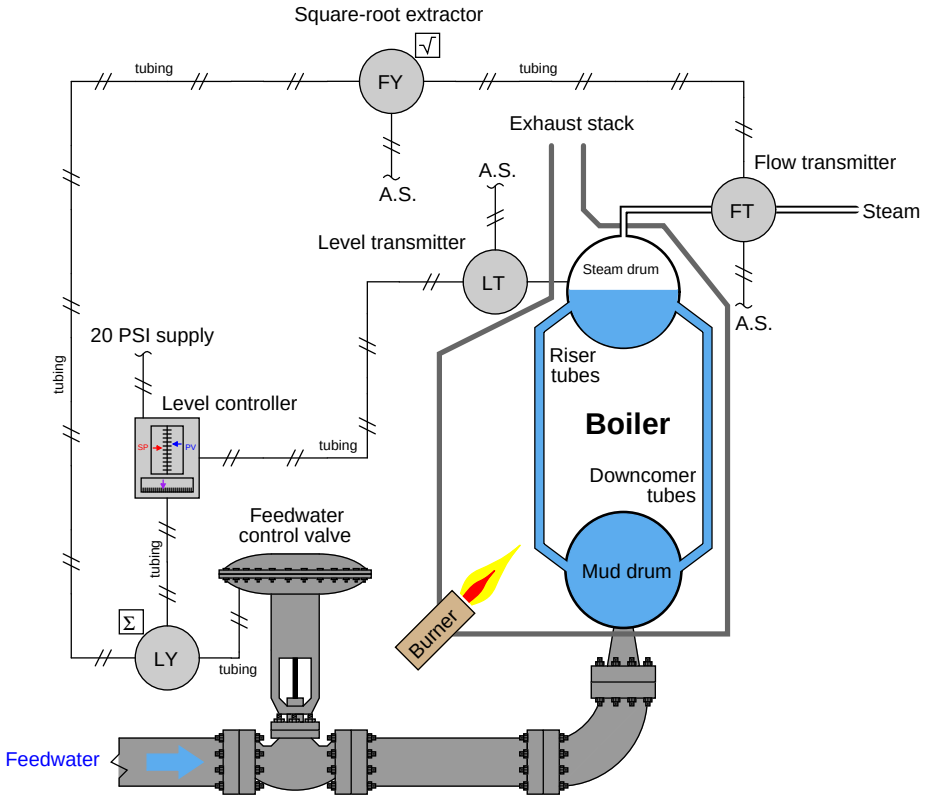
The indicator in this case would be a special pressure gauge, calibrated to read in units of process pressure although actuated by the pressure of clean compressed air from the transmitter instead of directly by process fluid. The most common range of air pressure for industrial pneumatic instruments is 3 to 15 PSI. An output pressure of 3 PSI represents the low end of the process measurement scale and an output pressure of 15 PSI represents the high end of the measurement scale. Applied to the previous example of a transmitter calibrated to a range of 0 to 250 PSI, a lack of process pressure would result in the transmitter outputting a 3 PSI air signal and full process pressure would result in an air signal of 15 PSI. The face of this special "receiver" gauge would be labeled from 0 to 250 PSI, while the actual mechanism would operate on the 3 to 15 PSI range output by the transmitter. Just like the 4-20 mA loop, the end-user need not know how the information gets transmitted from the process to the indicator. The 3-15 PSI signal medium is once again transparent to the operator.

Pneumatic temperature, flow, and level control systems have all been manufactured to use the same principle of 3-15 PSI air pressure signaling. In each case, the transmitter and controller are both supplied clean compressed air at some nominal pressure (20 to 25 PSI, usually) and the instrument signals travel via tubing. The following illustrations show what some of these applications look like:





Two-element boiler steam drum level control



Instruments functioning on compressed air, and process measurement signals transmitted as air

pressures through long runs of metal tubing, was the norm for industrial instrumentation prior to the advent of reliable electronics. In honor of this paradigm, instrument technicians were often referred to as *instrument mechanics*, for these air-powered devices were mechanically complex and in frequent need of adjustment to maintain high accuracy.

Back in the days of control room panels populated by rows and rows of pneumatic indicators, recorders, and controllers, clean and organized routing of all the instrument signal tubes was a significant concern. By contrast, electrical wires are relatively easy to organize through the use of marshaling panels and terminal blocks – bundles of tubes (especially metal tubes!) are not. A photograph taken of the upper rear portion of an old control room panel shows a portion of a marshaling board where dozens of bulkhead-style 1/4 inch instrument tube fittings are organized in neat rows[1], where a multitude of pneumatic instrument signal lines once attached:

Each bulkhead fitting bears a numbered tag[2], for easy identification and documentation of tube connections. Loop diagrams of pneumatic control systems documented each bulkhead fitting where an instrument signal passed, in the same way that modern loop diagrams document each terminal block where an electrical signal connection is made.

Pneumatic instruments still find wide application in industry, although it is increasingly rare to

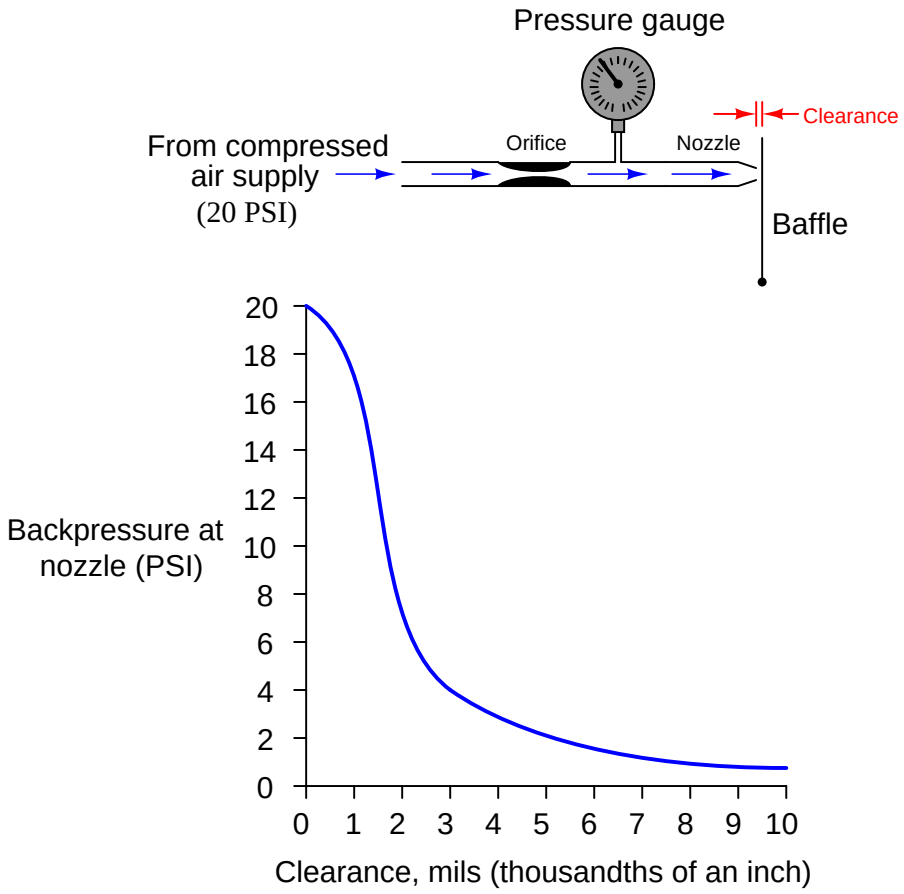
[1] Note the staggered layout of the tube fittings, intended to improve access to each one. Remember that the technician used a 9/16 inch wrench to loosen and tighten the tube fitting nuts, so it was important to have working room between fittings in which to maneuver a wrench.

[2] The numbers are difficult to see here, because the entire panel has been painted in a thick coat of grey paint. This particular panel was stripped of all pneumatic instruments and outfitted with electronic instruments, so the rows of bulkhead fittings no longer serve a purpose, but to remind us of legacy technology. I must wonder if some day in the future I will include a photograph of an empty terminal strip in another chapter of this book, as I explain how wired "legacy" instruments have all but been replaced by wireless (radio) instruments! Let the ghosts of the past speak to you, dear reader, testifying to the endless march of technological evolution.

encounter completely pneumatic control loops. One of the most common applications for pneumatic control system components is control valve actuation, where pneumatic technology still dominates. Not only is compressed air used to create the actuation force in many control valve mechanisms, it is still often the signal medium employed to command the valve's position. Quite often this pneumatic signal originates from a device called an *I/P transducer*, or *current-to-pressure converter*, taking a 4-20 mA control signal from the output of an electronic controller and translating that information as a pneumatic 3-15 PSI signal to the control valve's positioner or actuator.

14.1 Pneumatic sensing elements

Most pneumatic instruments use a simple but highly sensitive mechanism for converting mechanical motion into variable air pressure: the *baffle-and-nozzle* assembly (sometimes referred to as a *flapper-and-nozzle* assembly). A baffle is nothing more than a flat object obstructing the flow of air out of a small nozzle by close proximity:

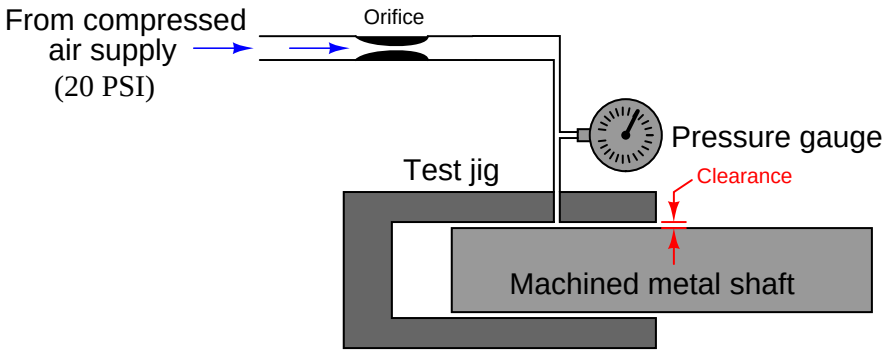


The physical distance between the baffle and the nozzle alters the resistance of air flow through the nozzle. This in turn affects the air pressure built up inside the nozzle (called the nozzle *backpressure*). Like a voltage divider circuit formed by one fixed resistor and one variable resistor, the baffle/nozzle mechanism "divides" the pneumatic source pressure to a lower value based on the ratio of restrictiveness between the nozzle and the fixed orifice.

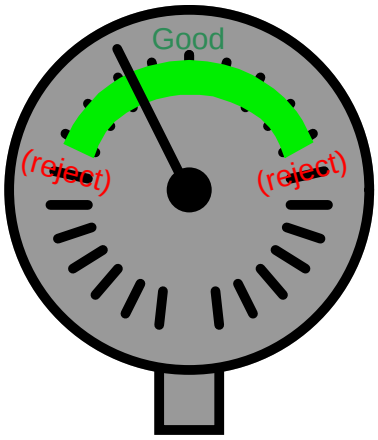
This crude assemblage is surprisingly sensitive, as shown by the graph. With a small enough orifice, just a few thousandths of an inch of motion is enough to drive the pneumatic output between its saturation limits. Pneumatic transmitters typically employ a small sheet-metal lever as the baffle. The slightest motion imparted to this baffle by changes in the process variable (pressure, temperature, flow, level, etc.) detected by some sensing element will cause the air pressure to

change in response.

The principle behind the operation of a baffle/nozzle mechanism is often used directly in quality-control work, checking for proper dimensioning of machined metal parts. Take for instance this shaft diameter checker, using air to determine whether or not a machined shaft inserted by a human operator is of the proper diameter after being manufactured on an assembly line:



If the shaft diameter is too small, there will be excessive clearance between the shaft and the inside diameter of the test jig, causing less air pressure to register on the gauge. Conversely, if the shaft diameter is too large, the clearance will be less and the gauge will register a greater air pressure because the flow of air will be obstructed by the reduced clearance. The exact pressure is of no particular consequence to the quality-control operator reading the gauge. What does matter is that the pressure falls within an acceptable range, reflecting proper manufacturing tolerances for the shaft. In fact, just like the 3-15 PSI "receiver gauges" used as pneumatic instrument indicators, the face of this pressure gauge might very well lack pressure units (such as kPa or PSI), but rather be labeled with a colored band showing acceptable limits of mechanical fit:

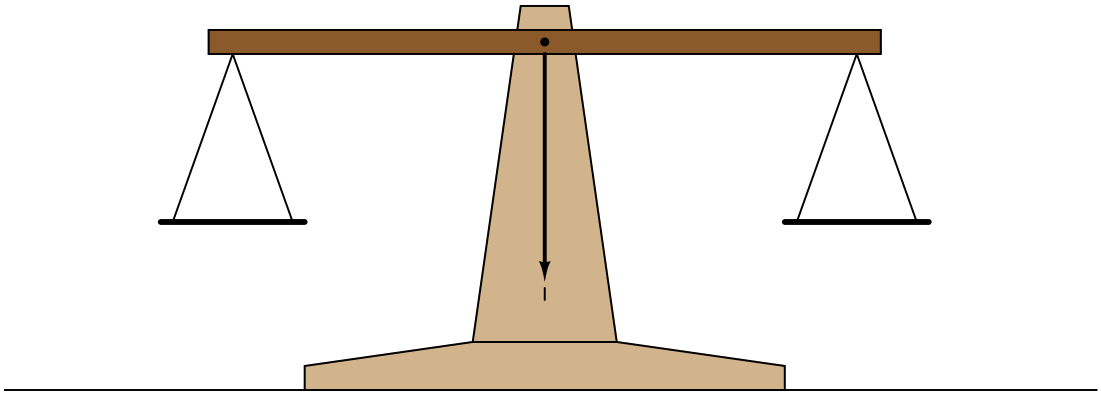


This is another example of the *analogue* nature of pneumatic pressure signals: the pressure registered by this gauge *represents* a completely different variable, in this case the mechanical fit of the shaft to the test jig.

Although it is possible to construct a pneumatic instrument consisting *only* of a baffle/nozzle mechanism, this is rarely done. Usually the baffle/nozzle mechanism is but one of several components that comprise a "balancing" mechanism in a pneumatic instrument. It is this concept of self-balancing that we will study next.

14.2 Self-balancing pneumatic instrument principles

A great many precision instruments use the principle of *balance* to measure some quantity. Perhaps the simplest example of a balance-based instrument is the common balance-beam scale used to measure mass in a laboratory:

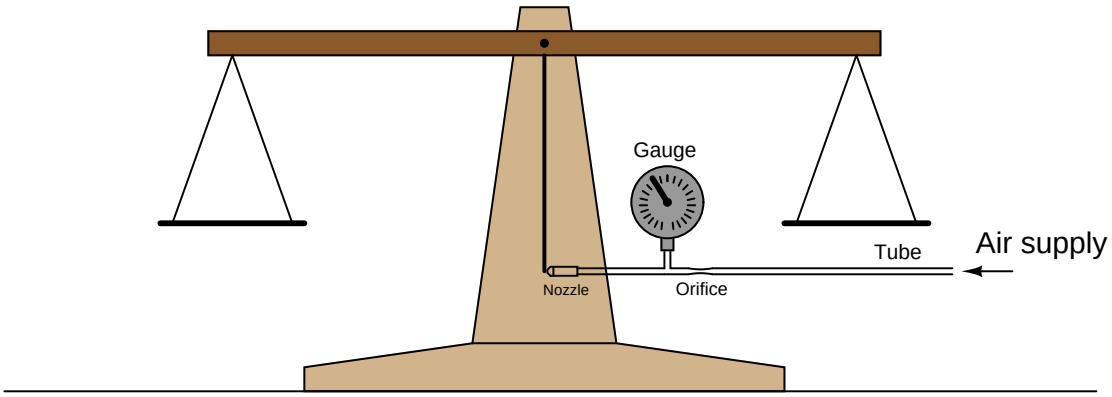


A specimen of unknown mass is placed in one pan of the scale, and precise weights are placed in the other pan until the scale achieves a condition of balance. When balance is achieved, the mass of the specimen is known to be equal to the sum total of mass in the other pan. An interesting detail to note about the scale itself is that it has no need of routine calibration. There is nothing to "drift" out of spec which would cause the scale to read inaccurately. In fact, the scale itself doesn't even have a gauge to register the mass of the specimen: all it has is a single mark indicating a condition of balance. To express this more precisely, the balance beam scale is actually a *differential mass* comparison device, and it only needs to be accurate at a single point: zero. In other words, it only has to be correct when it tells you there is zero difference in mass between the specimen and the standard masses piled on the other pan.

The elegance of this mechanism allows it to be quite accurate. The only real limitation to accuracy is the certainty to which we know the masses of the balancing weights.

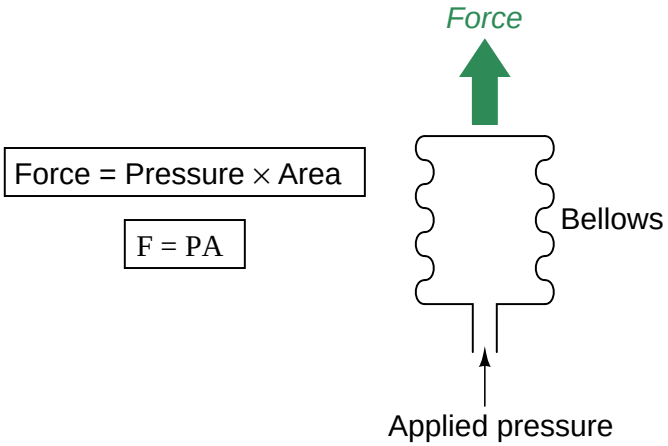
Imagine being tasked with the challenge of automating this laboratory scale. Suppose we grew weary of having to pay a lab technician to place standard weights on the scale to balance it for every new measurement, and we decided to find a way for the scale to balance itself. Where would we start? Well, we would need some sort of mechanism to tell when the scale was out of balance, and another mechanism to change weight on the other pan whenever an out-of-balance condition was detected.

The baffle/nozzle mechanism previously discussed would suffice quite well as a detection mechanism. Simply attach a baffle to the end of the pointer on the scale, and attach a nozzle adjacent to the baffle at the "zero" position (where the pointer should come to a rest at balance):



Now we have a highly sensitive means of indicating when the scale is balanced, but we still have not yet achieved full automation. The scale cannot balance itself, at least not yet.

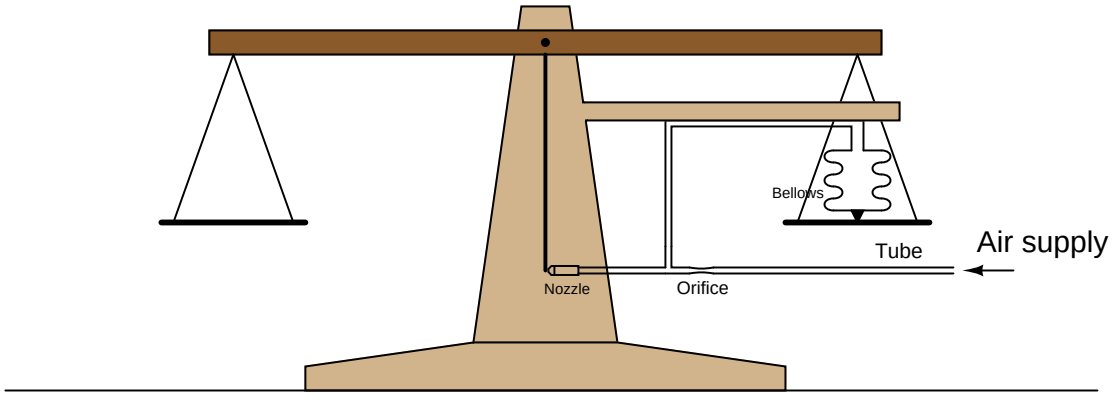
What if, instead of using precise, machined, brass weights placed on the other pan to counter the mass of the specimen, we used a pneumatically-actuated force generator operated by the backpressure of the nozzle? An example of such a "force generator" is a *bellows*: a device made of thin sheet metal with circular corrugations in it, such that it resembles the bellows fabric on an accordion. Pneumatic pressure applied to the interior of the bellows causes it to elongate. If the metal of the bellows is flexible enough so it does not naturally restrain the motion of expansion, the force generated by the expansion of the bellows will almost exactly equal that predicted by the force-pressure-area equation:



A photograph of a brass bellows unit appears here, the bellows taken from a Foxboro model 130 pneumatic controller:

If the bellows' expansion is externally restrained so it does not stretch appreciably – and therefore the metal never gets the opportunity to act as a restraining spring – the force exerted by the bellows on that restraining object will *exactly* equal the pneumatic pressure multiplied by the cross-sectional area of the bellows' end.

Applying this to the problem of the self-balancing laboratory scale, imagine fixing a bellows to the frame of the scale so it presses downward on the pan where the brass weights normally go, then connecting the bellows to the nozzle backpressure:



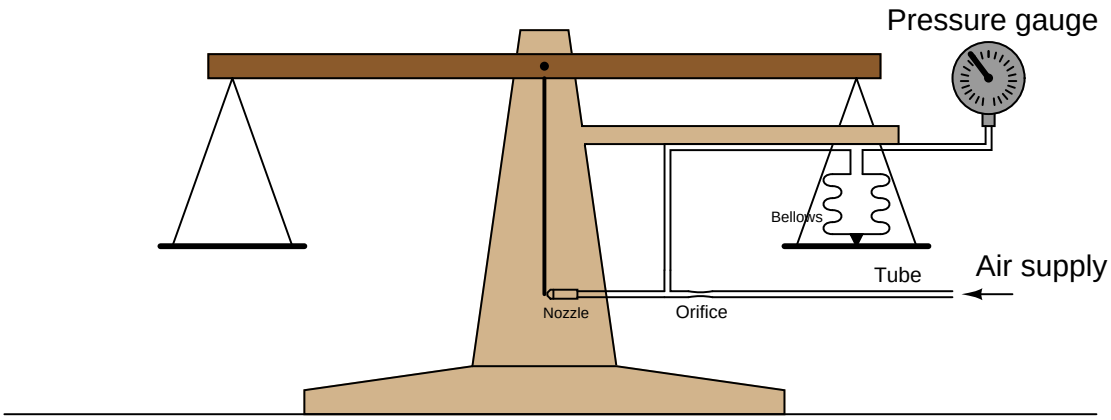
Now the scale *will* self-balance. When mass is added to the left-hand pan, the pointer (baffle) will move ever so slightly toward the nozzle until enough backpressure builds up behind the nozzle to make the bellows exert the proper amount of balancing force and bring the pointer back (very close) to its original balanced condition. This balancing action is entirely automatic: the nozzle backpressure

adjusts to whatever it needs to be in order to keep the pointer at the balanced position, applying or venting pressure to the bellows as needed to keep the system in a condition of equilibrium. What we have created is a *negative feedback system*, where the output of the system (nozzle backpressure) continuously adjusts to match and balance the input (the applied mass).

This is all well and good, but how does this help us determine the mass of the specimen in the left-hand pan? What good is this self-balancing scale if we cannot *read* the balancing force? All we have achieved so far is to make the scale self-balancing. The next step is making the balancing force readable to a human operator.

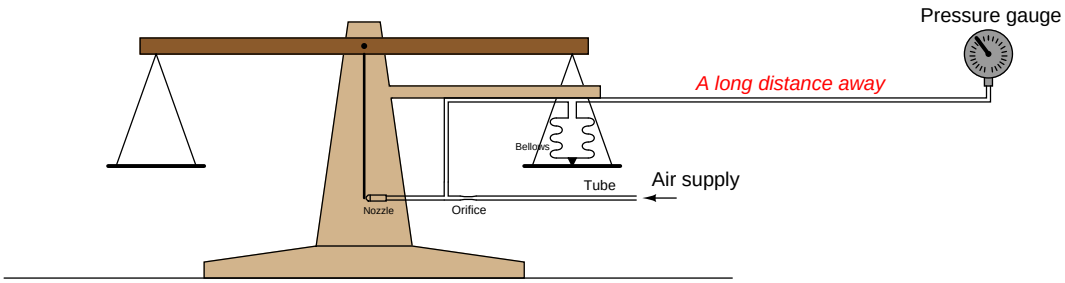
Before we add the final piece to this automated scale, it is worthwhile to reflect on what has been done so far. By adding the baffle/nozzle and bellows mechanisms to the scale, we have abolished the need for brass weights and instead have substituted air pressure. In effect, the scale translates the specimen's mass into a proportional, *analogue*, air pressure. What we really need is a way to now translate that air pressure into a human-readable indication of mass.

The solution is simple: add the pressure gauge back to the system. The gauge will register air pressure, but this time the air pressure will be proportionately equivalent to specimen mass. In honor of this proportionality, we may label the face of the pressure gauge in units of grams (mass) instead of PSI or kPa (pressure):



It is very important to note how the pressure gauge performs an entirely different function now than when it did prior to the addition of the feedback bellows. With just a baffle-nozzle mechanism at work, the pressure gauge was hyper-sensitive to any motion of the scale's balance beam – it served only as a highly sensitive indicator of balance. Now, with the addition of the feedback bellows, the pressure gauge actually indicates how much mass is in the specimen pan, not merely whether the scale is balanced or not. As we add more mass to the specimen pan, the gauge's indication proportionately increases. As we take away mass from the specimen pan, the gauge's indication proportionately decreases.

Although it may seem as though we are done with the task of fully automating the laboratory scale, we can go a step further. Building this pneumatic negative-feedback balancing system provides us with a capability the old manually-operated scale never had: *remote indication.* There is no reason why the indicating gauge must be located near the scale. Nothing prevents us from locating the receiver gauge some distance from the scale, and using long lengths of tubing to connect the two:



By equipping the scale with a pneumatic self-balancing apparatus, we have turned it into a *pneumatic mass transmitter*, capable of relaying the mass measurement in pneumatic, analog form

to an indicating gauge far away. This is the basic *force-balance* principle used in most pneumatic industrial transmitters to convert some process measurement into a 3-15 PSI pneumatic signal.

14.3 Pilot valves and pneumatic amplifying relays

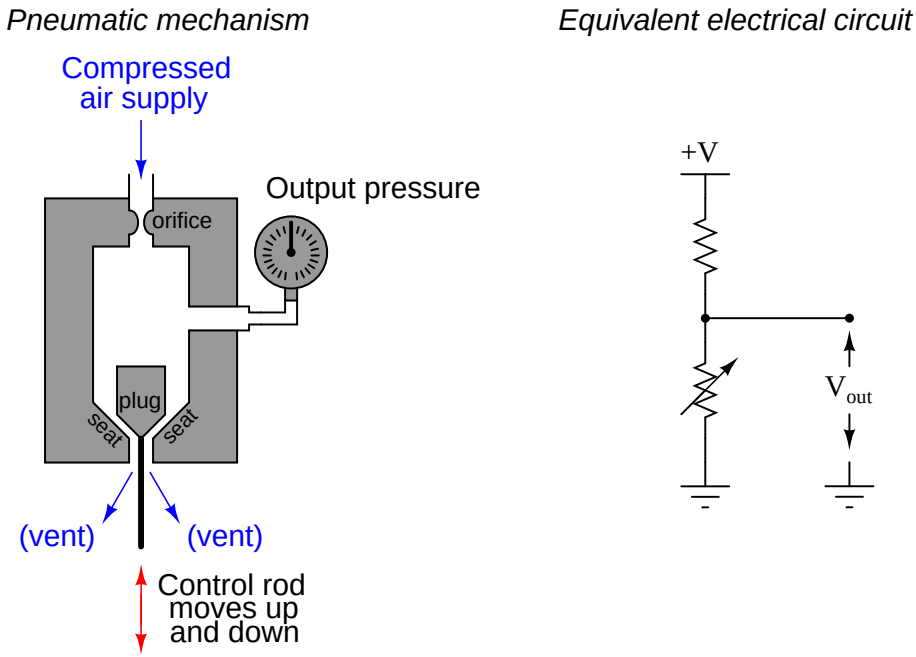
Self-balancing mechanisms such as the fictitious pneumatic laboratory scale in the previous section are most accurate when the imbalance detection mechanism is most sensitive. In other words, the more aggressively the baffle/nozzle mechanism responds to slight out-of-balance conditions, the more precise will be the relationship between measured variable (mass) and output signal (air pressure to the gauge).

A plain baffle/nozzle mechanism may be made extremely sensitive by reducing the size of the orifice. However, a problem caused by decreasing orifice size is a corresponding decrease in the nozzle's ability to provide increasing backpressure to fill a bellows of significant volume. In other words, a smaller orifice will result in greater sensitivity to baffle motion, but it also limits the air *flow rate* available to fill the bellows, which makes the system slower to respond. Another disadvantage of smaller orifices is that they become more susceptible to plugging due to impurities in the compressed air.

An alternative technique to making the baffle/nozzle mechanism more sensitive is to amplify its output pressure using some other pneumatic device. This is analogous to increasing the sensitivity of a voltage-generating electrical detector by passing its output voltage signal through an electronic amplifier. Small changes in detector output become bigger changes in amplifier output which then causes our self-balancing system to be even more precise.

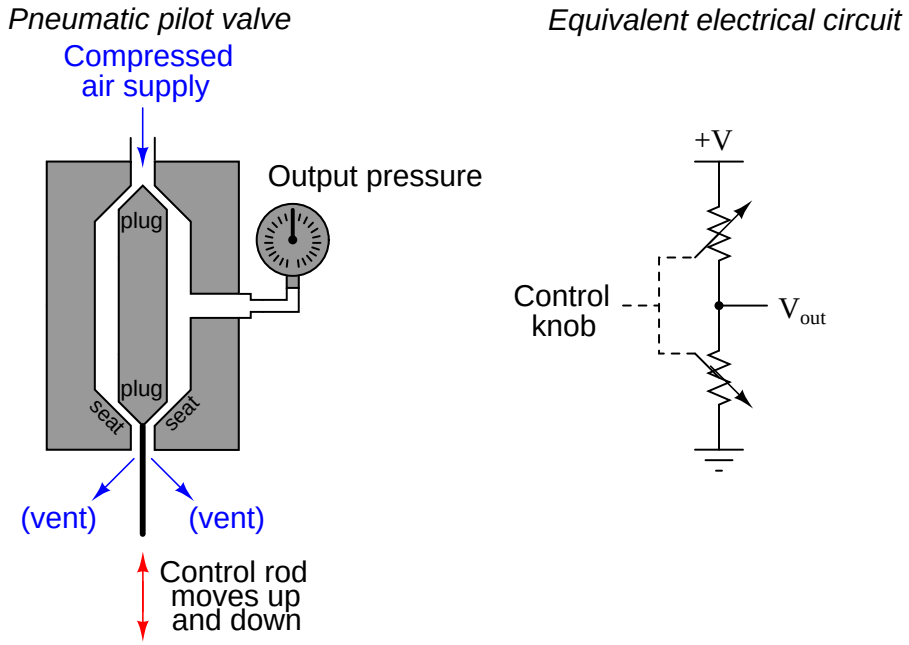
What we need, then, is a pneumatic amplifier: a mechanism to amplify small changes in air pressure and convert them into larger changes in air pressure. In essence, we need to find a pneumatic equivalent of the electronic *transistor*: a device that lets one signal control another.

First, let us analyze the following pneumatic mechanism and its electrical analogue (as shown on the right):



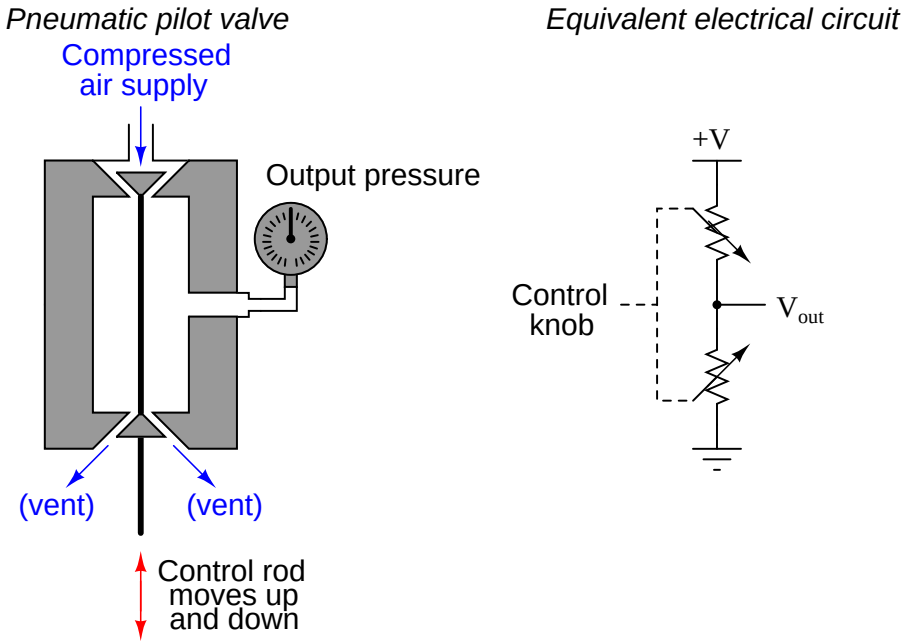
As the control rod is moved up and down by an outside force, the distance between the plug and the seat changes. This changes the amount of resistance experienced by the escaping air, thus causing the pressure gauge to register varying amounts of pressure. There is little functional difference between this mechanism and a baffle/nozzle mechanism. Both work on the principle of one variable restriction and one fixed restriction (the orifice) "dividing" the pressure of the compressed air source to some lesser value.

The sensitivity of this pneumatic mechanism may be improved by extending the control rod and adding a second plug/seat assembly. The resulting mechanism, with dual plugs and seats, is known as a pneumatic *pilot valve*. An illustration of a pilot valve is shown here, along with its electrical analogue (on the right):

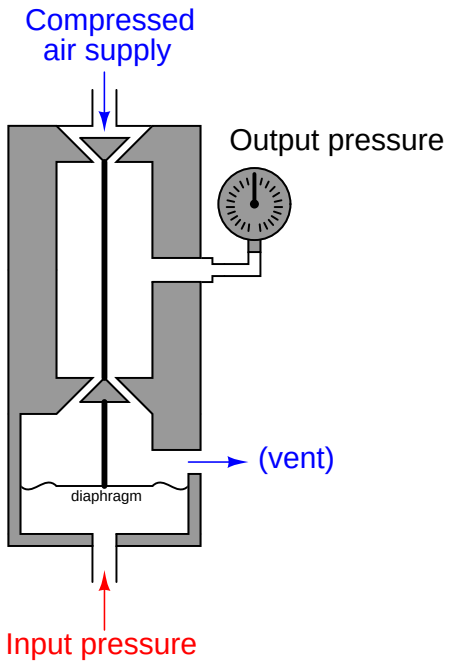


As the control rod is moved up and down, *both* variable restrictions change in complementary fashion. As one restriction opens up, the other pinches shut. The combination of two restrictions changing in opposite direction results in a much more aggressive change in output pressure as registered by the gauge.

A similar design of pilot valve reverses the directions of the two plugs and seats. The only operational difference between this pilot valve and the previous design is an inverse relationship between control rod motion and pressure:



At this point, all we've managed to accomplish is build a better baffle/nozzle mechanism. We still do not yet have a pneumatic equivalent of an electronic transistor. To accomplish that, we must have some way of allowing an air pressure signal to control the motion of a pilot valve's control rod. This is possible with the addition of a *diaphragm*, as shown in this illustration:

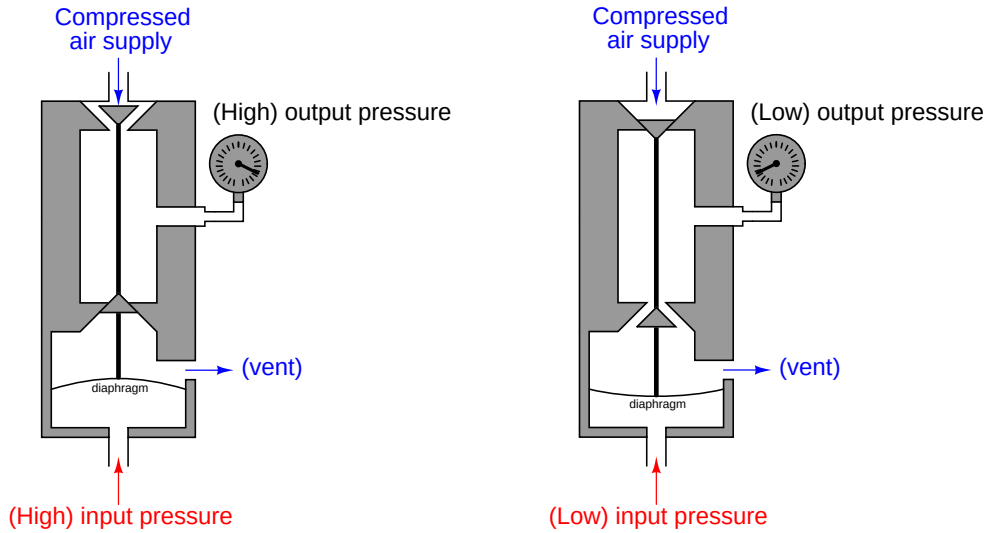


The diaphragm is nothing more than a thin disk of sheet metal, upon which an incoming air pressure signal presses. Force on the diaphragm is a simple function of signal pressure (P) and diaphragm area (A), as described by the standard force-pressure-area equation:

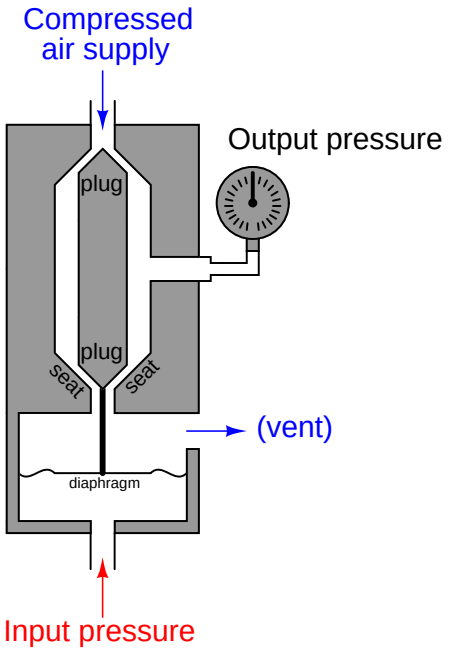
$$F = PA$$

If the diaphragm is taut, the elasticity of the metal allows it to also function as a spring. This allows the force to translate into displacement (motion), forming a definite relationship between applied air pressure and control rod position. Thus, the applied air pressure input will exert control over the output pressure. The addition of an actuating mechanism to the pilot valve turns it into a *pneumatic relay*, which is the pneumatic equivalent of the electronic transistor we were looking for.

It is easy to see how the input air signal exerts control over the output air signal in these two illustrations:



Since there is a direct relationship between input pressure and output pressure in this pneumatic relay, we classify it as a *direct-acting relay*. If we were to add an actuating diaphragm to the first pilot valve design, we would have a *reverse-acting relay* as shown here:

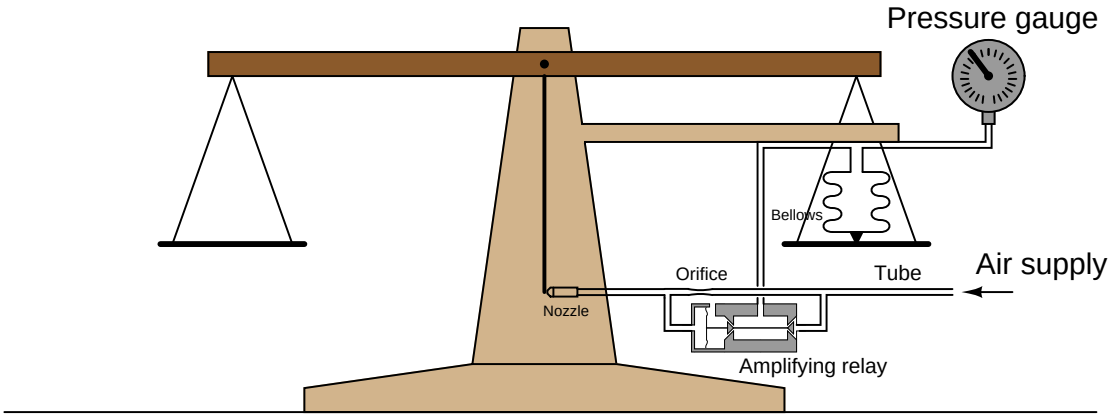


The *gain* (A) of any pneumatic relay is defined just the same as the gain of any electronic amplifier circuit, the ratio of output change to input change:

$$A = \frac{\Delta\text{Output}}{\Delta\text{Input}}$$

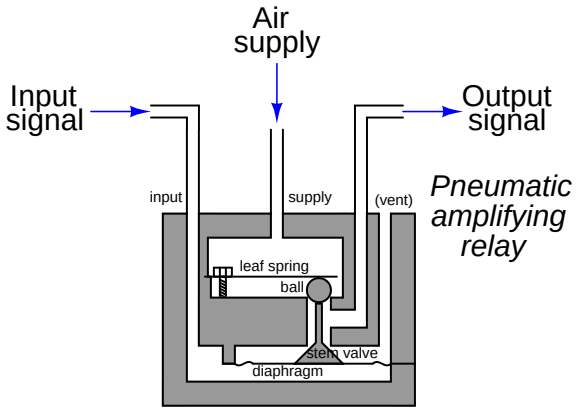
For example, if an input pressure change of $\Delta 2$ PSI results in an output pressure change of $\Delta 12$ PSI, the gain of the pneumatic relay is 6.

Adding a pneumatic pressure-amplifying relay to a force-balance system such as our hypothetical laboratory scale improves the performance of that pneumatic system:



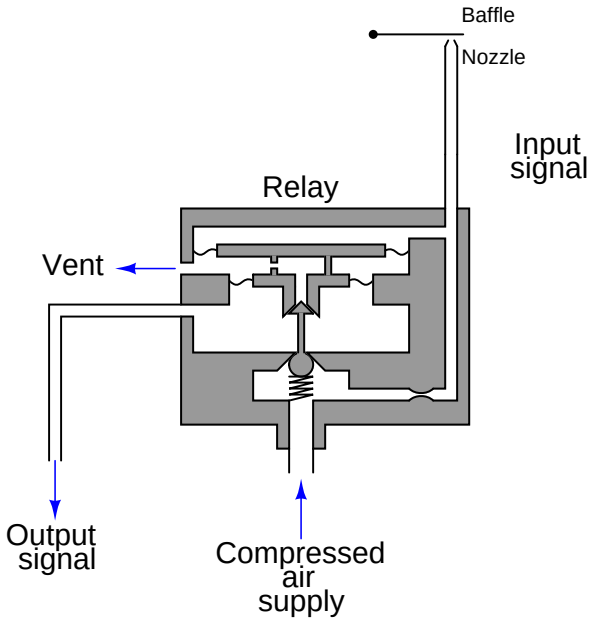
Since the relay amplifies the nozzle's backpressure, the force-balancing bellows responds even more aggressively than before (without the relay) to any change in baffle position. This makes the scale more sensitive, better able to sense small changes in applied mass than without an amplifying relay.

The Foxboro corporation designed a great many of their pneumatic instruments to used a very sensitive amplifying relay:



The motion of the diaphragm actuated a pair of valves: one with a cone-shaped plug and the other with a metal ball for a plug. The ball-plug allowed supply air to go to the output port, while the cone-shaped "stem valve" plug vented excess air pressure to the vent port.

The Fisher corporation used a different style of amplifying relay in some of their pneumatic instruments:

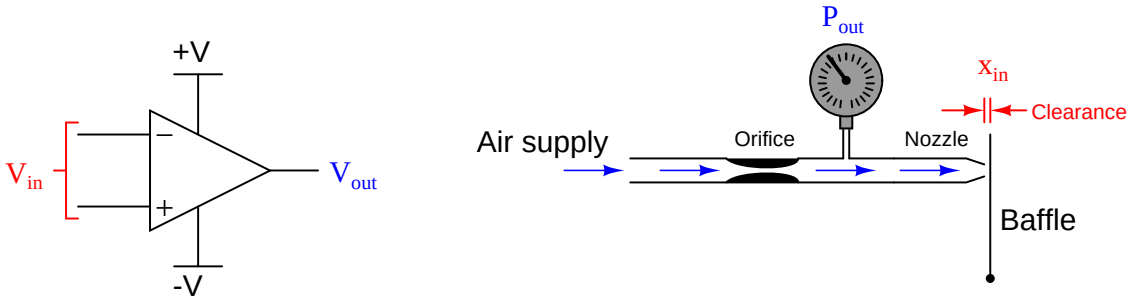


The gain of this Fisher relay was much less than that of the Foxboro relay, since output pressure in the Fisher relay was allowed to act against input pressure by exerting force on a sizable diaphragm. The movable vent seat in the Fisher relay made this design a "non-bleeding" type, meaning it possessed the ability to close both supply and vent valves at the same time, allowing it to hold an output air pressure between saturation limits without bleeding a substantial amount of compressed air to atmosphere through the vent. The Foxboro relay design, by contrast, was a "bleeding type," whose ball and stem valves could never close simultaneously, and thus would always bleed some compressed air to atmosphere so long as the output pressure remained somewhere between saturation limits.

14.4 Analogy to opamp circuits

Self-balancing pneumatic instrument mechanisms are very similar to negative-feedback operational amplifier circuits, in that negative feedback is used to generate an output signal in precise proportion to an input signal. This section compares simple operational amplifier ("opamp") circuits with analogous pneumatic mechanisms for the purpose of illustrating how negative feedback works, and learning how to generally analyze pneumatic mechanisms.

In the following illustration, we see an opamp with no feedback (open loop), next to a baffle/nozzle mechanism with no feedback (open loop):



For each system there is an input and an output. For the opamp, input and output are both electrical (voltage) signals: V_{in} is the differential voltage between the two input terminals, and V_{out} is the single-ended voltage measured between the output terminal and ground. For the baffle/nozzle, the input is the physical gap between the baffle and nozzle (x_{in}) while the output is the backpressure indicated by the pressure gauge (P_{out}).

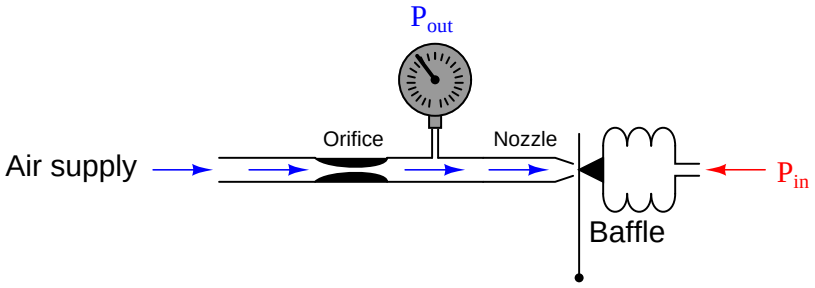
Both systems have very large gains. Operational amplifier open-loop gains typically exceed 200,000 (over 100 dB), and we have already seen how just a few thousandths of an inch of baffle motion is enough to drive the backpressure of a nozzle nearly to its limits (supply pressure and atmospheric pressure, respectively).

Gain is always defined as the ratio between output and input for a system. Mathematically, it is the quotient of output *change* and input *change*, with "change" represented by the triangular Greek capital-letter delta:

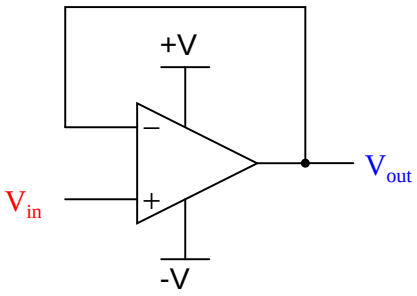
$$\text{Gain} = A = \frac{\Delta\text{Output}}{\Delta\text{Input}}$$

Normally, gain is a unitless ratio. We can easily see this for the opamp circuit, since both output and input are voltages, any unit of measurement for voltage would cancel in the quotient, leaving a unitless quantity. This is not so evident in the baffle/nozzle system, with the output represented in units of pressure and the input represented in units of distance.

If we were to add a bellows to the baffle/nozzle mechanism, we would have a system that inputs and outputs fluid pressure, allowing us to more formally define the gain of the system as a unitless ratio of $\frac{\Delta P_{out}}{\Delta P_{in}}$:



The general effect of negative feedback is to decrease the gain of a system, and also make that system's response more linear over the operating range. This is not an easy concept to grasp, however, and so we will explore the effect of adding negative feedback in detail for both systems. The simplest expression of negative feedback is a condition of 100% negative feedback, where the whole strength of the output signal gets "fed back" to the amplification system in degenerative fashion. For an opamp, this simply means connecting the output terminal directly to the inverting input terminal:



We call this "negative" or "degenerative" feedback because its effect is counteractive in nature. If the output voltage rises too high, the effect of feeding this signal to the inverting input will be to bring the output voltage back down again. Likewise, if the output voltage is too low, the inverting input will sense this and act to bring it back up again. *Self-correction* is the hallmark of any negative-feedback system.

Having connected the inverting input directly to the output of the opamp leaves us with the noninverting terminal as the sole remaining input. Thus, our input voltage signal is a ground-referenced voltage just like the output. The voltage gain of this circuit is unity (1), meaning that the output will assume whatever voltage level is present at the input, within the limits of the opamp's power supply. If we were to send a voltage signal of 5 volts to the noninverting terminal of this opamp circuit, it would output 5 volts, provided that the power supply exceeds 5 volts in potential from ground.

Let's analyze exactly why this happens. First, we will start with the equation representing the open-loop output of an opamp, as a function of its differential input voltage:

$$V_{out} = A_{OL}(V_{in(+)} - V_{in(-)})$$

As stated before, the open-loop voltage gain of an opamp is typically very large ($A_{OL} = 200{,}000$ or more!). Connecting the opamp's output to the inverting input terminal simplifies the equation: V_{out} may be substituted for $V_{in(-)}$, and $V_{in(+)}$ simply becomes V_{in} since it is now the only remaining input. Reducing the equation to the two variables of V_{out} and V_{in} and a constant (A_{OL}) allows us to solve for overall voltage gain ($\frac{V_{out}}{V_{in}}$) as a function of the opamp's internal voltage gain (A_{OL}). The following sequence of algebraic manipulations shows how this is done:

$$V_{out} = A_{OL}(V_{in} - V_{out})$$

$$V_{out} = A_{OL}V_{in} - A_{OL}V_{out}$$

$$A_{OL}V_{out} + V_{out} = A_{OL}V_{in}$$

$$V_{out}(A_{OL} + 1) = A_{OL}V_{in}$$

$$\text{Overall gain} = \frac{V_{out}}{V_{in}} = \frac{A_{OL}}{A_{OL} + 1}$$

If we assume an internal opamp gain of 200,000, the overall gain will be very nearly equal to unity (0.999995). Moreover, this near-unity gain will remain quite stable despite large changes in the opamp's internal (open-loop) gain. The following table shows the effect of major A_{OL} changes on overall voltage gain (A_V):

A_{OL} Internal gain	A_V Overall gain
100,000	0.99999
200,000	0.999995
300,000	0.999997
500,000	0.999998
1,000,000	0.999999

Note how an order of magnitude change[3] in A_{OL} (from 100,000 to 1,000,000) results is a miniscule change in overall voltage gain (from 0.99999 to 0.999999). Negative feedback clearly has a stabilizing effect on the closed-loop gain of the opamp circuit, which is the primary reason it finds such wide application in engineered systems. It was this effect that led Harold Black in the late 1920's to apply negative feedback to the design of very stable telephone amplifier circuits.

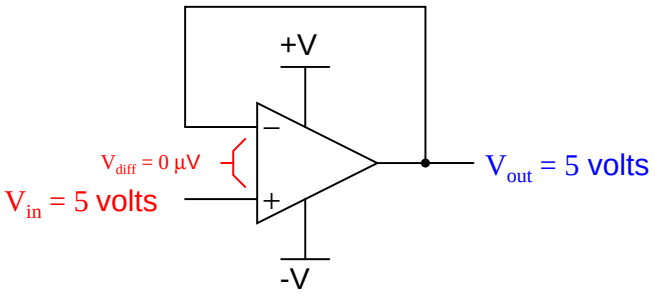
[3] An "order of magnitude" is nothing more than a ten-fold change. Do you want to sound like you're really smart and impress those around you? Just start comparing ordinary differences in terms of orders of magnitude. "Hey dude, that last snowboarder's jump was an *order of magnitude* higher than the one before!" "Whoa, that's some big air . . ." Just don't make the mistake of using decibels in the same way ("Whoa dude, that last jump was at least 10 dB higher than the one before!") – you don't want people to think you're a nerd.

If we subject our negative feedback opamp circuit to a constant input voltage of exactly 5 volts, we may expand the table to show the effect of changing open-loop gain on the output voltage, and also the differential voltage appearing between the opamp's two input terminals:

A_{OL} Internal gain	A_V Overall gain	V_{out} Output voltage	$V_{in(+)} - V_{in(-)}$ Differential input voltage
100,000	0.99999	4.99995	0.00005
200,000	0.999995	4.999975	0.000025
300,000	0.999997	4.99998	0.00002
500,000	0.999998	4.99999	0.00001
1,000,000	0.999999	4.999995	0.000005

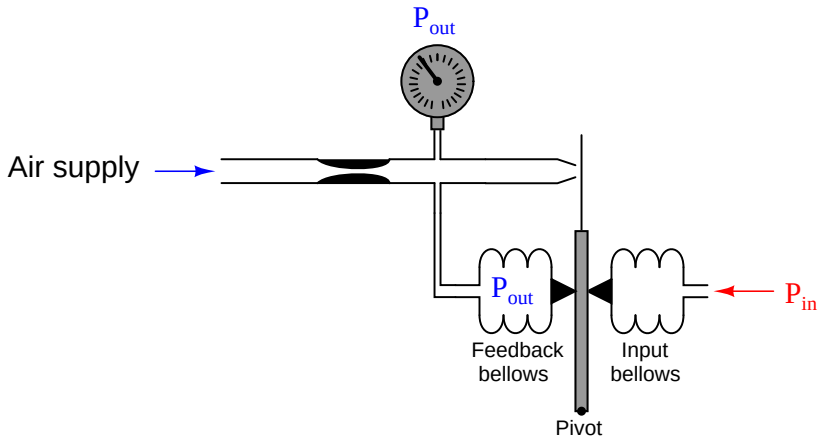
With such extremely high open-loop voltage gains, it hardly requires any difference in voltage between the two input terminals to generate the necessary output voltage to balance the input. Thus, $V_{out} = V_{in}$ for all practical purposes.

One of the "simplifying assumptions" electronics technicians and engineers make when analyzing opamp circuits is that the differential input voltage in any negative feedback circuit is zero. As we see in the above table, this assumption is very nearly true. Following this assumption to its logical consequence allows us to predict the output voltage of any negative feedback opamp circuit quite simply. For example:



If we simply assume there will be no difference of voltage between the two input terminals of the opamp with negative feedback in effect, we may conclude that the output voltage is exactly equal to the input voltage, since that is what *must* happen in order for the two input terminals to see equal potentials.

Now let us apply similar techniques to the analysis of a pneumatic baffle/nozzle mechanism. Suppose we arrange a pair of identical bellows in opposition to one another on a force beam, so any difference in force output by the two bellows will push the baffle either closer to the nozzle or further away from it:



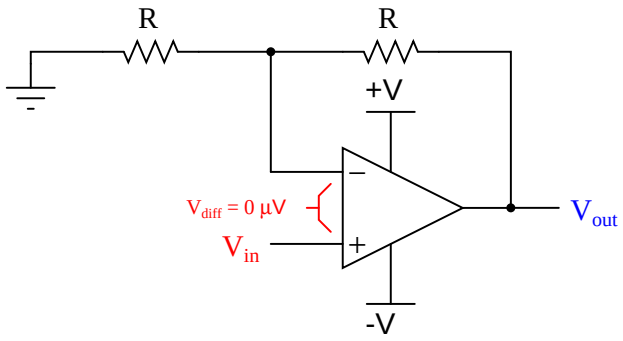
It should be clear that the left-hand bellows, which experiences the same pressure (P_{out}) as the pressure gauge, introduces negative feedback into the system. If the output pressure happens to rise too high, the baffle will be pushed away from the nozzle by the force of the feedback bellows, causing backpressure to decrease and stabilize. Likewise, if the output pressure happens to go too low, the baffle will move closer to the nozzle and cause the backpressure to rise again. Once again we see the defining characteristic of negative feedback in action: its self-correcting nature works to *counteract* any change in output conditions, such that the output pressure precisely tracks the input pressure.

As we have seen already, the baffle/nozzle is exceptionally sensitive to motion. Only a few thousandths of an inch of motion is sufficient to saturate the nozzle backpressure at either extreme (supply air pressure or zero, depending on which direction the baffle moves). This is analogous to the differential inputs of an operational amplifier, which only need to see a few microvolts of potential difference to saturate the amplifier's output.

Introducing negative feedback to the opamp led to a condition where the differential input voltage was held to (nearly) zero. In fact, this potential is so small that we safely considered it zero for the purpose of more easily analyzing the output response of the system. *We may make the exact same "simplifying assumption" for the pneumatic mechanism:* we will assume the baffle/nozzle gap remains constant in order to more easily determine the output pressure response to an input pressure.

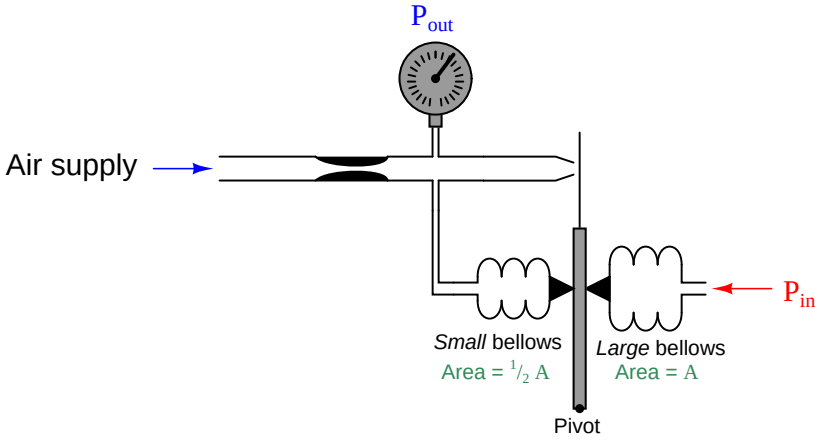
If we simply assume the baffle/nozzle gap cannot change so long as negative feedback is actively working, we may conclude that the output pressure is exactly equal to the input pressure for the pneumatic system shown, since that is what *must* happen in order for the two pressures to generate exactly opposing forces through two identical bellows so the baffle will not move from its original position.

The analytical technique of assuming perfect balance in a negative feedback system works just as well for more complicated systems. Consider the following opamp circuit:



Here, negative feedback occurs through a voltage divider from the output terminal to the inverting input terminal, such that only one-half of the output voltage gets "fed back" degeneratively. If we follow our simplifying assumption that perfect balance (zero difference of voltage) will be achieved between the two opamp input terminals due to the balancing action of negative feedback, we are led to the conclusion that V_{out} must be exactly *twice* the magnitude of V_{in}. In other words, the output voltage must increase to twice the value of the input voltage in order for the divided feedback signal to exactly equal the input signal. Thus, feeding back half the output voltage yields an overall voltage gain of two.

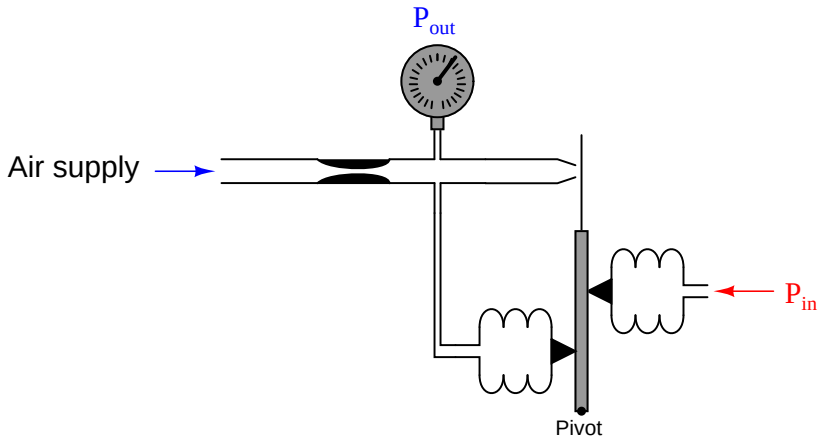
If we make the same (analogous) change to the pneumatic system, we see the same effect:



Here, the feedback bellows has been made smaller (exactly half the surface area of the input bellows). This results in half the amount of force applied to the force beam for the same amount of pressure. If we follow our simplifying assumption that perfect balance (zero baffle motion) will be achieved due to the balancing action of negative feedback, we are led to the conclusion that P_{out} must be exactly *twice* the magnitude of P_{in}. In other words, the output pressure must increase to twice the value of the input pressure in order for the divided feedback force to exactly equal the

input force and prevent the baffle from moving. Thus, our pneumatic mechanism has a pressure gain of two, just like the opamp circuit with divided feedback had a voltage gain of two.

We could have achieved the same effect by moving the feedback bellows to a lower position on the force beam instead of changing its surface area:



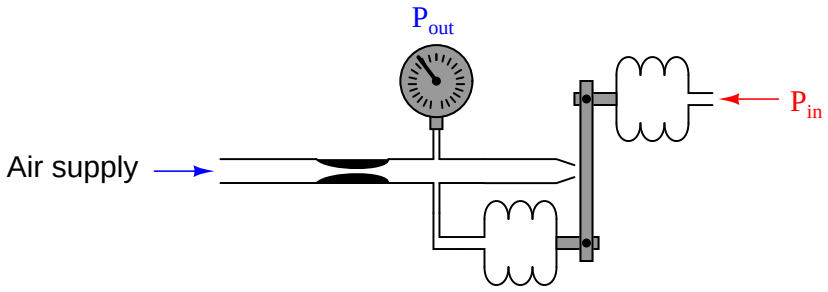
This arrangement effectively reduces the feedback force by placing the feedback bellows at a mechanical disadvantage to the input bellows. If the distance between the feedback bellows tip and the force beam pivot is exactly half the distance between the input bellows tip and the force beam pivot, the effective force ratio will be one-half. The result of this "divided" feedback force is that the output pressure must rise to *twice* the value of the input pressure, since the output pressure is at a mechanical disadvantage to the input. Once again, we see a balancing mechanism with a gain of two.

It should be noted that this mechanism, which achieves different gains by applying different lever ratios, is more properly classified as a *moment-balance* system rather than a pure *force-balance* system. In all the previous pneumatic examples where one bellows directly opposed another bellows (both bellows sharing the same centerline), the balancing action was direct force against direct force. Here, in this leverage mechanism, the balance is not a matter of one bellows force countered directly by another bellows force, but rather one bellows' *moment*[4] countered by another bellows' *moment*.

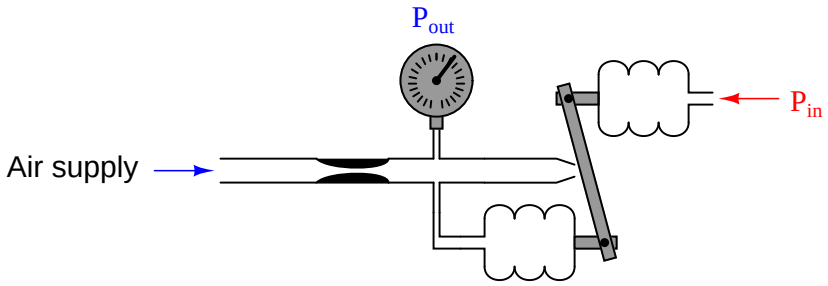
Pneumatic instruments built such that bellows' forces directly oppose one another in the same line of action to constrain the motion of a beam are known as "force balance" systems. Instruments built such that bellows' forces oppose one another through different lever lengths (such as in the last system) are technically known as "moment balance" systems, referencing the *moment arm lengths* through which the bellows' forces act to balance each other. However, one will often find that "moment balance" instruments are commonly referred to as "force balance" because the two principles are so similar.

[4]In physics, the word *moment* refers to the product of force times lever length. By the same token, we could classify this pneumatic mechanism as a *torque-balance* system.

An entirely different classification of pneumatic instrument is known as *motion balance*. The same "simplifying assumption" of zero baffle/nozzle gap motion holds true for the analysis of these mechanisms as well:

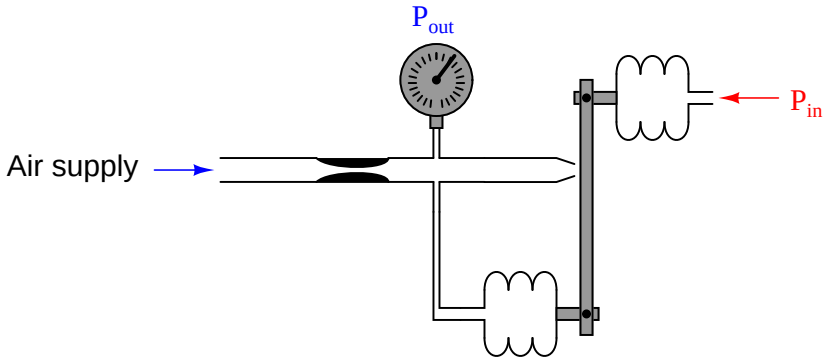


In this mechanism there is no fixed pivot for the beam. Instead, the beam hangs between the ends of two bellows units, affixed by pivoting links. As input pressure increases, the input bellows expands outward, attempting to push the beam closer to the nozzle. However, if we follow our assumption that negative feedback holds the nozzle gap constant, we see that the feedback bellows must expand the same amount, and thus (if it has the same area and spring characteristics as the input bellows) the output pressure must equal the input pressure:



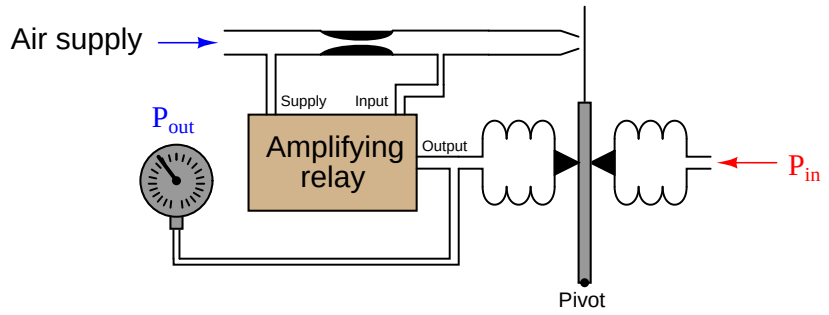
We call this a *motion* balance system instead of a *force* balance system because we see two motions canceling each other out to maintain a constant nozzle gap instead of two forces canceling each other out to maintain a constant nozzle gap.

The gain of a motion-balance pneumatic instrument may be changed by altering the bellows-to-nozzle distance such that one of the two bellows has more effect than the other. For instance, this system has a gain of 2, since the feedback bellows must move twice as far as the input bellows in order to maintain a constant nozzle gap:



Force-balance (and moment-balance) instruments are generally considered more accurate than motion-balance instruments because motion-balance instruments rely on the pressure elements (bellows, diaphragms, or bourdon tubes) possessing predictable spring characteristics. Since pressure must accurately translate to motion in a motion-balance system, there must be a predictable relationship between pressure and motion in order for the instrument to maintain accuracy. If anything happens to affect this pressure/motion relationship such as metal fatigue or temperature change, the instrument's calibration will drift. Since there is negligible motion in a force-balance system, pressure element spring characteristics are irrelevant to the operation of these devices, and their calibrations remain more stable over time.

Both force- and motion-balance pneumatic instruments are usually equipped with an *amplifying relay* between the nozzle backpressure chamber and the feedback bellows. The purpose of an amplifying relay in a self-balancing pneumatic system is the same as the purpose of providing an operational amplifier with an extremely high open-loop voltage gain: the more internal gain the system has, the closer to ideal the "balancing" effect will be. In other words, our "simplifying assumption" of zero baffle/nozzle gap change will be closer to the truth in a system where the nozzle pressure gets amplified before going to the feedback bellows:



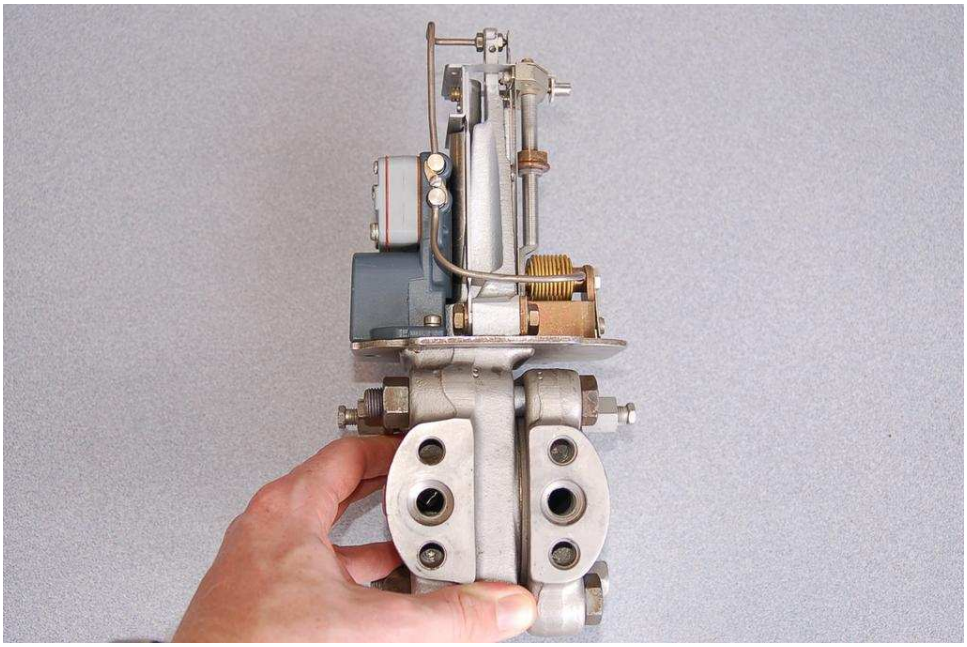
Thus, adding a relay to a self-balancing pneumatic system is analogous to increasing the open-loop voltage gain of an opamp (A_{OL}) by several-fold: it makes the overall gain *closer to ideal*. The overall gain of the system, though, is dictated by the ratio of bellows leverage on the force beam, just like the overall gain of a negative-feedback opamp circuit is dictated by the feedback network and *not* by the opamp's internal (open-loop) voltage gain.

14.5 Analysis of practical pneumatic instruments

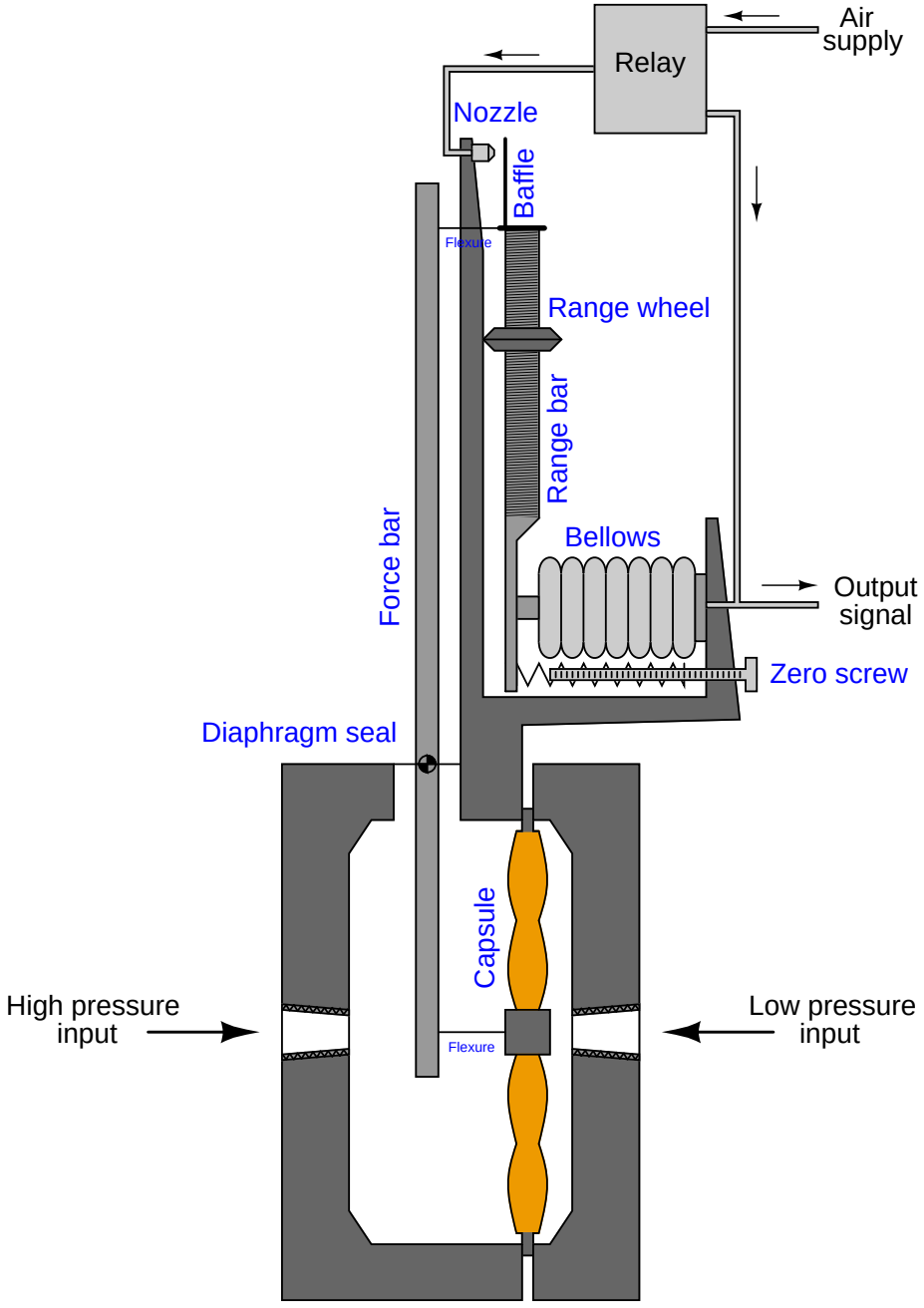
To better understand the design and operation of self-balancing pneumatic mechanisms, it is helpful to examine the workings of some actual instruments. In this section, we will explore three different pneumatic instruments: the Foxboro model 13A differential pressure transmitter, the Foxboro model E69 I/P (electro-pneumatic) transducer, the Fisher model 546 I/P (electro-pneumatic) transducer, and the Fisher-Rosemount model 846 I/P (electro-pneumatic) transducer.

14.5.1 Foxboro model 13A differential pressure transmitter

Perhaps one of the most popular pneumatic industrial instruments ever manufactured is the Foxboro model 13 differential pressure transmitter. A photograph of one with the cover removed is shown here:



A functional illustration of this instrument identifies its major components:



Part of the reason for this instrument's popularity is the extreme utility of differential pressure transmitters in general. A "DP cell" may be used to measure pressure, vacuum, pressure differential,

liquid level, liquid or gas flow, and even liquid density. A reason for this *particular* differential transmitter's popularity is excellent design: the Foxboro model 13 transmitter is rugged, easy to calibrate, and quite accurate.

Like so many pneumatic instruments, the model 13 transmitter uses the *force-balance* (more precisely, the *moment-balance*) principle whereby any shift in position is sensed by a detector (the baffle/nozzle assembly) and immediately corrected through negative feedback to restore equilibrium. As a result, the output air pressure signal becomes an analogue of the differential process fluid pressure sensed by the diaphragm capsule. In the following photograph you can see my index finger pointing to the baffle/nozzle mechanism at the top of the transmitter:

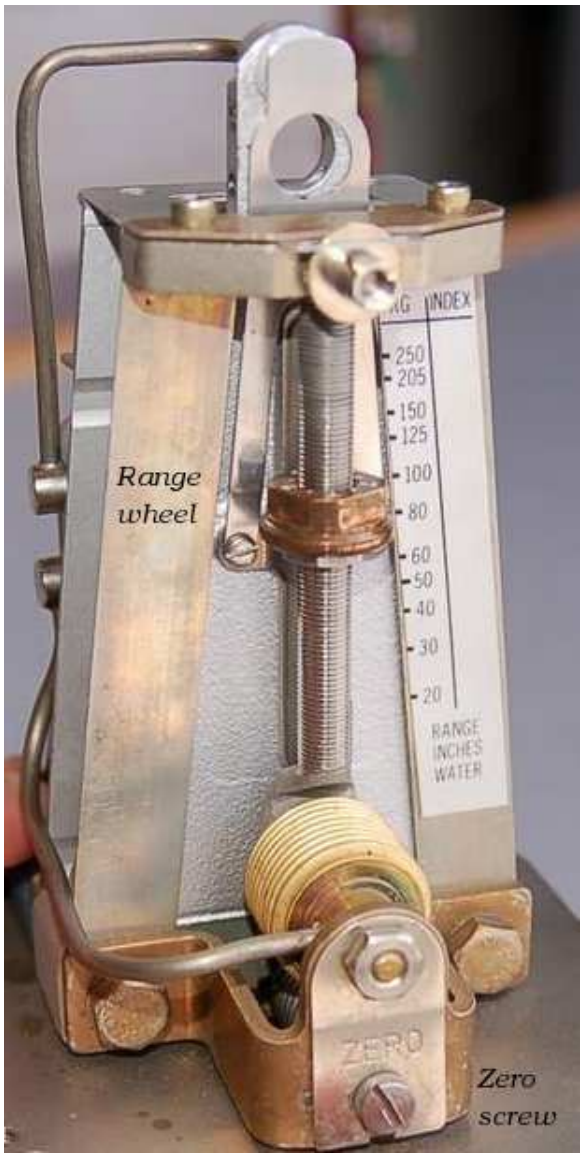
Let's analyze the behavior of this transmitter step-by-step as it senses an increasing pressure on the "High pressure" input port. As the pressure here increases, the large diaphragm capsule is forced to the right. The same effect would occur if the pressure on the "Low pressure" input port were to decrease. This is a *differential* pressure transmitter, so what it responds to is changes in pressure *difference* between the two input ports.

This resultant motion of the capsule tugs on the thin flexure connecting it to the force bar. The force bar pivots at the fulcrum (where the small diaphragm seal is located) in a counter-clockwise rotation, tugging the flexure at the top of the force bar. This motion causes the range bar to also pivot at its fulcrum (the sharp-edged "range wheel"), moving the baffle closer to the nozzle.

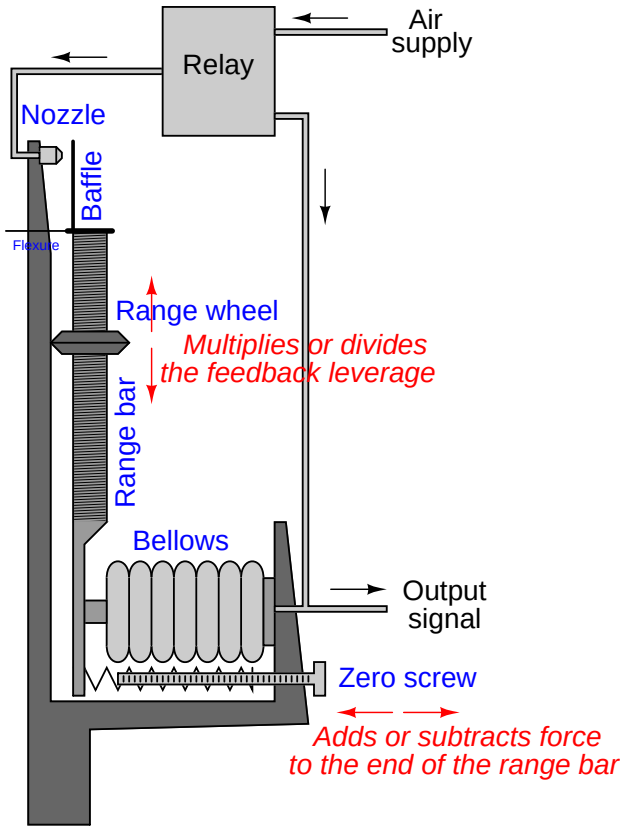
As the baffle approaches the nozzle, air flow through the nozzle becomes more restricted, accumulating backpressure in the nozzle. This backpressure increase is greatly amplified in the relay, which sends an increasing air pressure signal both to the output line and to the bellows at the bottom of the range bar. Increasing pneumatic pressure in the bellows causes it to push harder on the bottom of the range bar, counterbalancing the initial motion and returning the range bar (and force bar) to their near-original positions.

Calibration of this instrument is accomplished through two adjustments: the zero screw and the range wheel. The zero screw simply adds tension to the bottom of the range bar, pulling it in

such a direction as to collapse the bellows as the zero screw is turned clockwise. This action pushes the baffle closer to the nozzle and tends to increase air pressure to the bellows as the system seeks equilibrium. If a technician turns the range wheel, the lever ratio of the range bar changes, affecting the ratio of force bar force to bellows force. The following photograph shows the range bar and range wheel of the instrument:



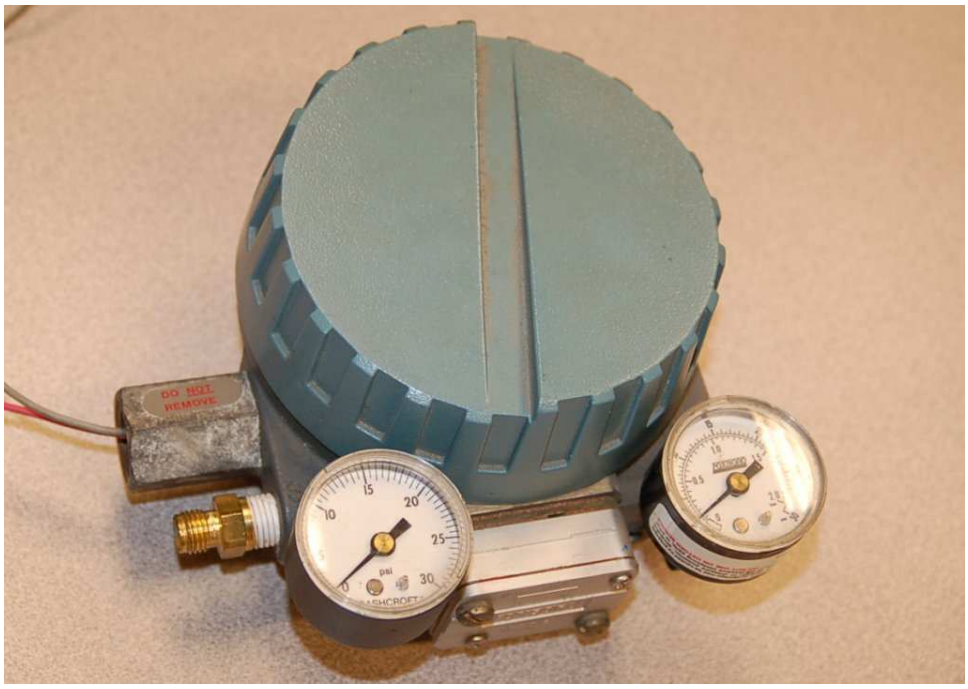
As in all instruments, the zero adjustment works by *adding or subtracting* a quantity, while the span adjustment works by *multiplying or dividing* a quantity. In the Foxboro model 13 pneumatic transmitter, the quantity in question is force, since this is a force-balance mechanism. The zero screw adds or subtracts force to the mechanical system by tensioning a spring, while the range wheel multiplies or divides force in the system by changing the mechanical advantage (force ratio) of a lever.



14.5.2 Foxboro model E69 "I/P" electro-pneumatic transducer

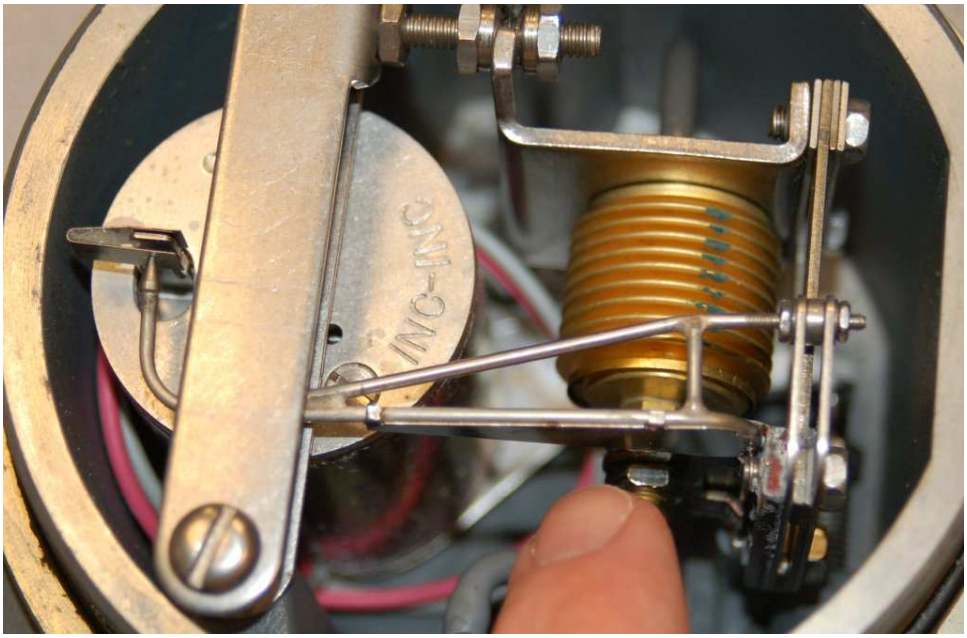
The purpose of any "I/P" transducer is to convert an electrical signal into a corresponding pneumatic signal. In most cases, this means an input of 4-20 mA DC and an output of 3-15 PSI, but alternative ranges do exist.

An example of an I/P transducer manufactured by Foxboro is the model E69, shown here:



Two pressure gauges indicate supply and output pressure, respectively. Wires convey the 4-20 mA electrical signal into the coil unit inside the transducer.

A view with the cover removed shows the balancing mechanism used to generate a pneumatic pressure signal from the electric current input. The baffle/nozzle may be seen at the left of the mechanism, the nozzle located at the end of a bent tube, facing the flat baffle on the surface of the circular coil unit:



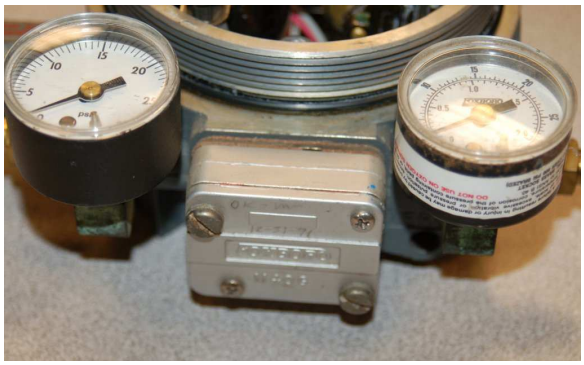
As electric current passes through the coil, it produces a magnetic field which reacts against a permanent magnet's field to generate a torque. This torque causes rotation against the restraint of a spring, with the baffle connected to the rotating assembly. Thus, the baffle moves like the needle of an analog electric meter movement in response to current: the more current through the coil, the more the coil assembly moves (and the baffle moves with it).

The nozzle faces this baffle, so when the baffle begins to move toward the nozzle, backpressure within the nozzle rises. This rising pressure is amplified by the relay, with the output pressure applied to a bellows. As the bellows expands, it draws the nozzle away from the advancing baffle, achieving balance by matching one motion (the baffle's) with another motion (the nozzle's).

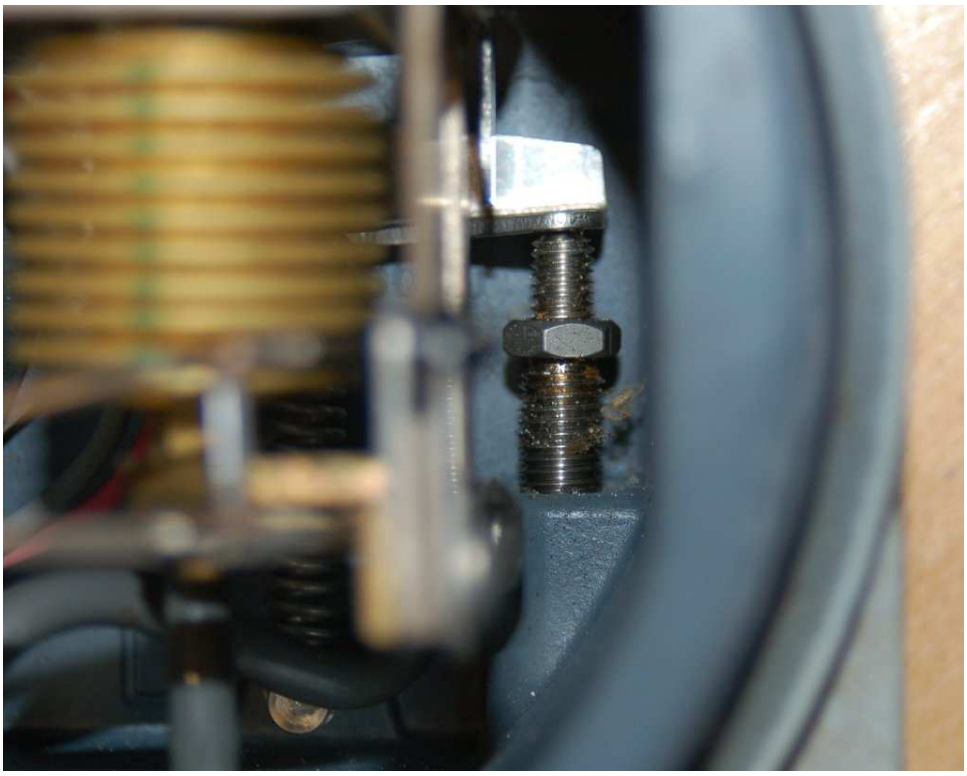
A closer view shows the baffle and nozzle in detail:

Thus, the self-balancing mechanism of the Foxboro model E69 transducer uses the *motion-balance* principle rather than the force-balance principle as applied in the Foxboro model 13 differential pressure transmitter. Instead of an input force precisely balancing an output force, constraining motion to a negligible degree, this mechanism allows the input freely move, matching that motion with a corresponding output motion to maintain a near-constant baffle/nozzle gap.

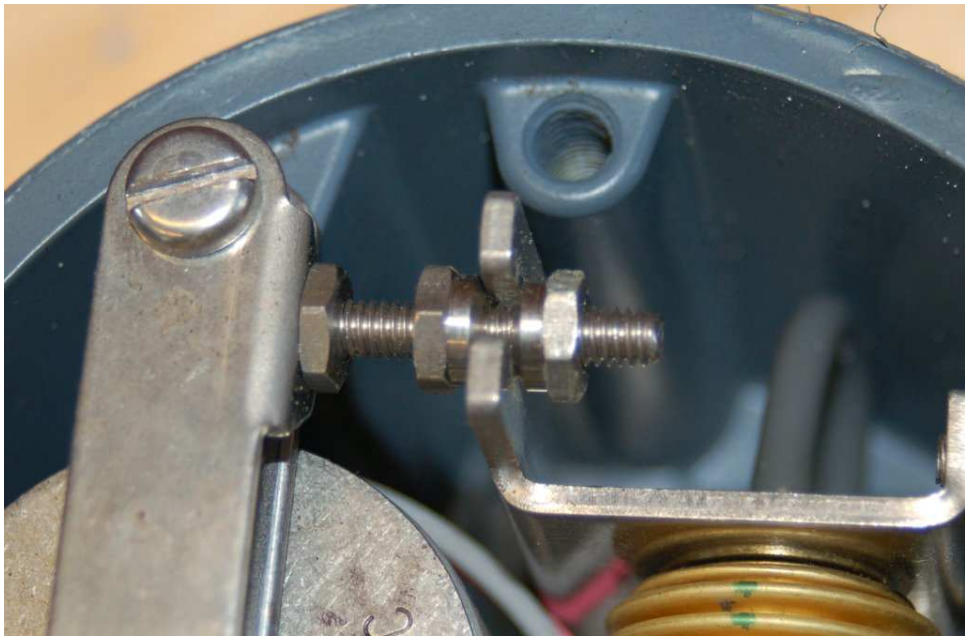
Interestingly the model E69 transducer employs the same pneumatic amplifying relay used in virtually every Foxboro pneumatic instrument:



As in all instruments, the zero adjustment works by *adding or subtracting* a quantity, while the span adjustment works by *multiplying or dividing* a quantity. In the Foxboro model E69 transducer, the quantity in question is motion, since this is a motion-balance mechanism. The zero adjustment adds or subtracts motion by offsetting the position of the nozzle closer to or further away from baffle. A close-up photograph of the zero adjustment screw shows it pressing against a tab to rotate the mounting baseplate upon which the coil unit is fixed. Rotating this baseplate add or subtracts angular displacement to/from the baffle's motion:

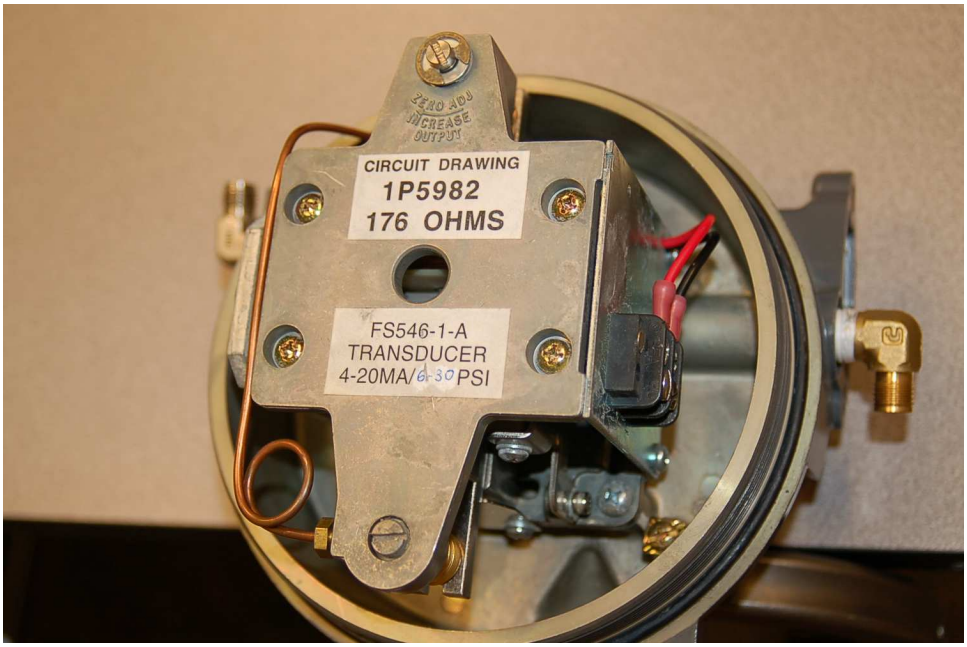


The span adjustment consists of changing the position of the nozzle relative to the baffle's center of rotation, so that a given amount of rotation equates to a different amount of balancing motion required of the nozzle. This adjustment consists of a pair of nuts locking the base of the bellows unit at a fixed distance from the baffle's center of rotation. Changing this distance alters the effective radius of the baffle as it swings around its center, therefore altering the gain (or span) of the motion-balance system:

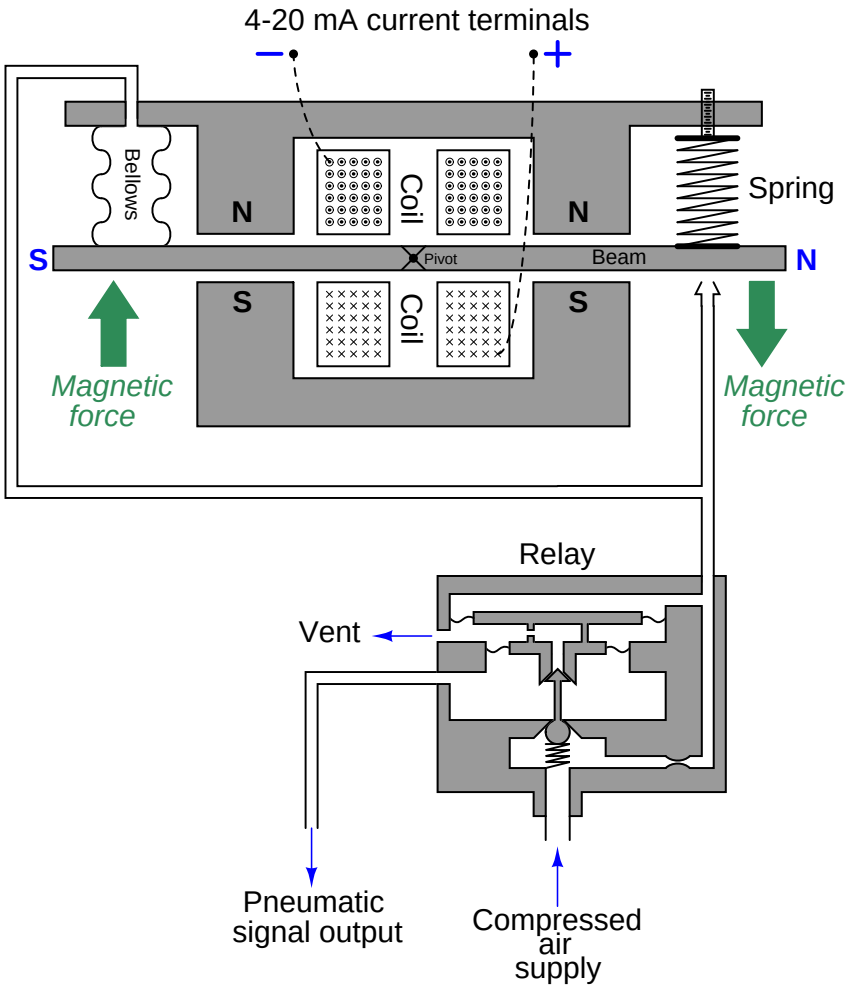


14.5.3 Fisher model 546 "I/P" electro-pneumatic transducer

The Fisher model 546 I/P transducer performs the same signal-conversion function (mA into PSI) as the Foxboro model E69, but it does so differently. The following photograph shows the internal mechanism of the model 546 transducer with its cover removed:



This particular instrument's construction tends to obscure its function, so I will use an illustrative diagram to describe its operation:



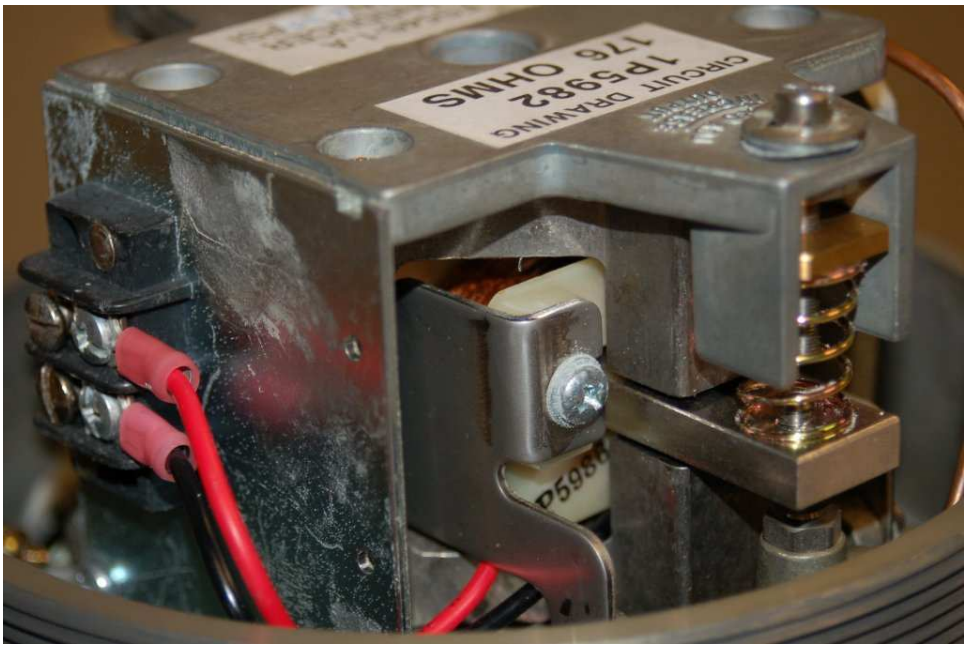
The heart of this mechanism is an iron beam, located between the poles of a permanent magnet assembly, and centered within an electromagnet coil (solenoid). Current passing through the electromagnet coil imparts magnetic poles to the ends of the beam. Following the arrow head/tail convention shown in the coil windings (the dots versus X marks) representing conventional flow vectors pointing out of the page (top) and going into the page (bottom) for the coil wrapped around the beam, the right-hand rule tells us that the beam will magnetize with the right-hand side being "North" and the left-hand side being "South." This will torque the beam clockwise around its pivot point (fulcrum), pushing the right-hand side down toward the nozzle.

Any advance of the beam toward the nozzle will increase nozzle backpressure, which is then fed to the balancing bellows at the other end of the beam. That bellows provides a restoring force to the beam to return it (nearly) to its original position. The phenomenon of an input force being

counter-acted by a balancing force to ensure minimum motion is the defining characteristic of a *force-balance* system. This is the same basic principle applied in the Foxboro model 13 differential pressure transmitter: an input force countered by an output force.

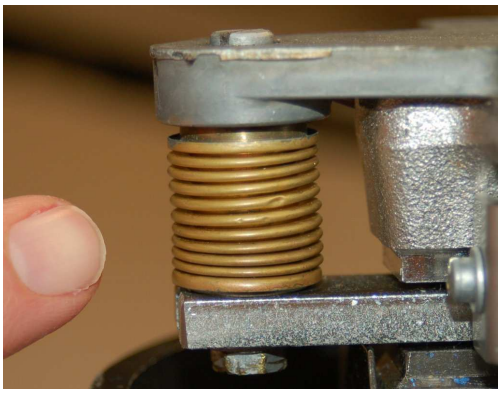
If you examine the diagram carefully, you will notice that this instrument's amplifying relay is not located within the force-balance feedback loop. The nozzle's backpressure is *directly* fed back to the balancing bellows with no amplification at all. A relay does exist, but its purpose is to provide a modest (approximately 2:1) pressure gain to raise the nozzle backpressure to standard levels (3-15 PSI, or 6-30 PSI).

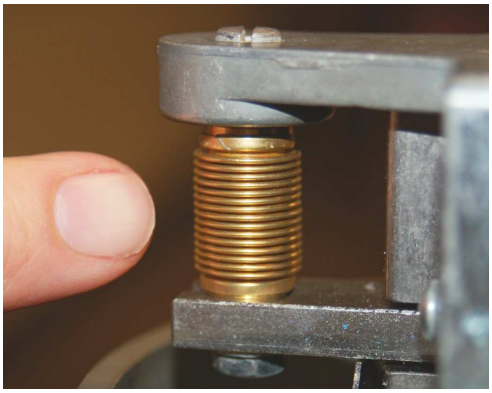
The next photograph shows the solenoid coil, force beam, and nozzle. If you look closely, you can see the copper-colored windings of the coil buried within the mechanism. The zero-adjustment spring is located above the beam, centered with the nozzle (below the beam):



Fisher manufactured these I/P transducers with two different pneumatic ranges: 3-15 PSI and 6-30 PSI. The mechanical difference between the two models was the size of feedback bellows used in each. In order to achieve the greater pressure range (6-30 PSI), a *smaller* feedback bellows was used. This may seem backward at first, but it makes perfect sense if you mentally follow the operation of the force-balance mechanism. In order to generate a greater air pressure for a given electric current through the coil, we must place the air pressure at a mechanical *disadvantage* to force it to rise higher than it ordinarily would in achieving balance. One way to do this is to decrease the effective area of the bellows, so that it takes a greater air pressure to generate the same amount of balancing force on the beam.

A 3-15 PSI bellows (left) is contrasted against a 6-30 PSI bellows (right) in this pair of photographs:





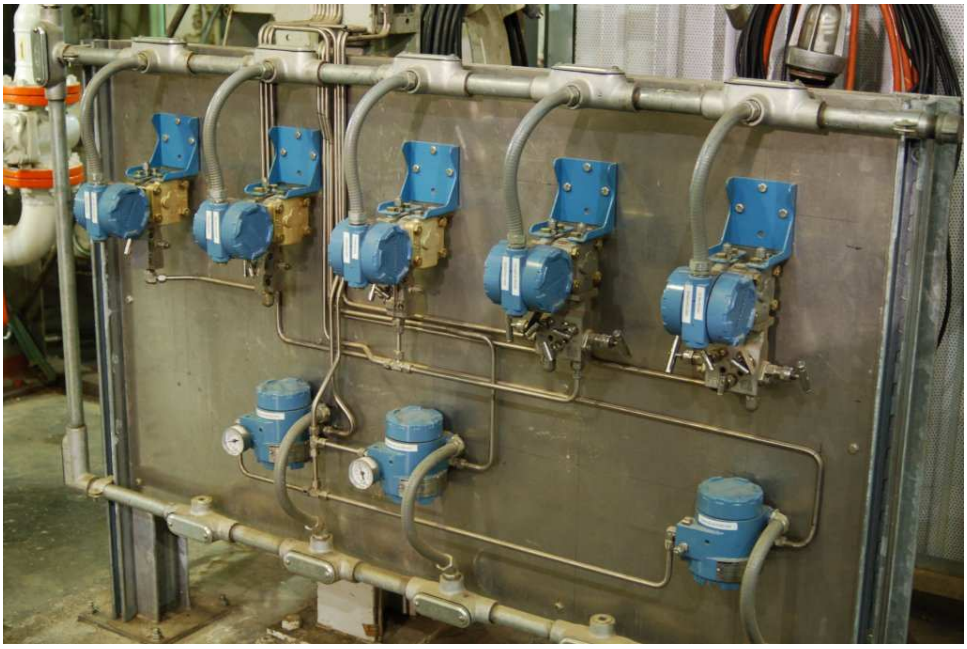
The span adjustment for this I/P transducer is achieved by varying the permanent-magnetic field strength acting against the beam's electro-magnetic field. Adjustment occurs through the use of a magnetic *shunt*: an iron plate moved closer to or further away from the permanent magnets, providing an alternate (shunt, or bypass) path for magnetic flux away from the force beam. Moving the shunt further away from the magnets strengthens the magnetic field "seen" by the beam, resulting in a multiplication of force on the beam and therefore a multiplication of output pressure. Moving the shunt closer to the magnets weakens the magnetic field "seen" by the beam, thereby dividing the reaction force and also the output pressure.

A view of the mechanism's other side reveals the magnetic shunt plate, complete with an instructional arrow showing the correct direction to turn the adjustment screw to increase output span:

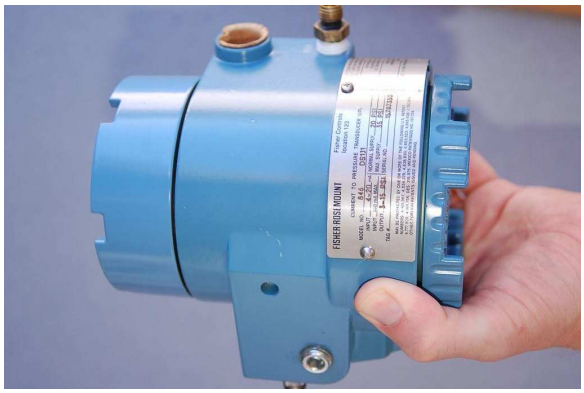
14.5.4 Fisher-Rosemount model 846 "I/P" electro-pneumatic transducer

The Fisher-Rosemount model 846 is a more modern I/P transducer than either the Foxboro model E69 or the Fisher model 546. It employs neither the force-balance nor the motion-balance principle in its operation, which makes it unique to analyze. This I/P unit is also unique in that it features a modular design allowing very convenient replacement of internal components when in service.

This next photograph shows three model 846 I/P transducers attached to a metal panel, below a set of five Rosemount model 1151 pressure transmitters:

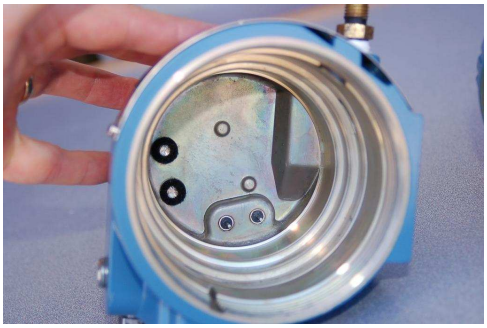


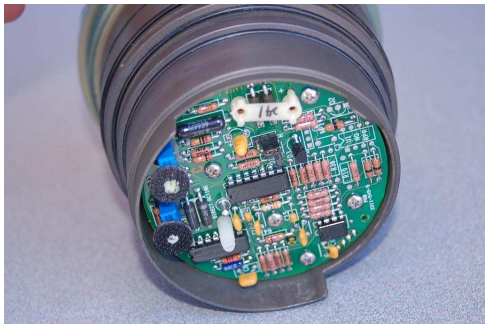
A closer photograph reveals the unit in more detail:



When one of the end-covers is unscrewed, the internal workings of the I/P may be removed as a single module. Both the removed module and the housing are shown in this photograph:

Shown separately, you can see where the module's current input terminals connect with matching pins in the housing. Even the zero and span adjustment potentiometers on the module circuit board are equipped with Velcro (hook and loop) pads, matching with pads attached to calibration screws on the housing. This simple yet effective mechanical coupling allows screws located on the exterior housing to adjust resistances on the module's circuit board for zero and span calibration, yet without exposing those delicate potentiometers to ambient weather conditions:



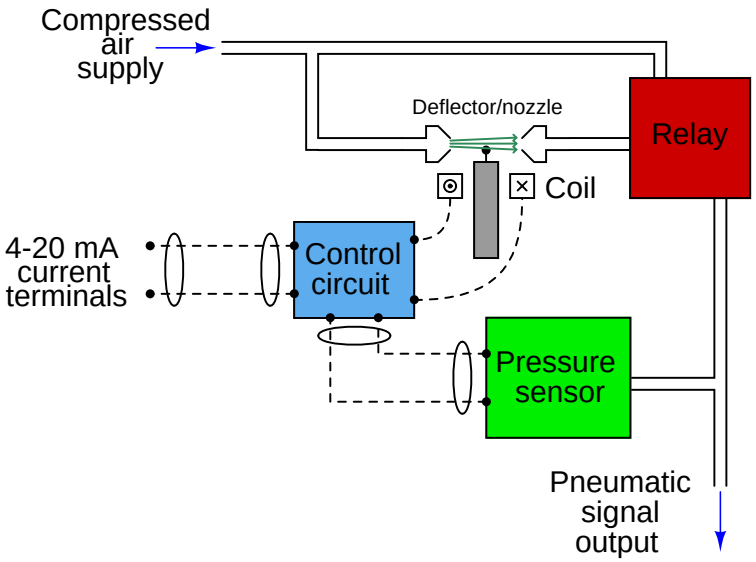


Pneumatic (air) connections are made to the housing through standard 1/4 inch female NPT pipe threads. Compressed air is passed to the module (and from the module back out to the housing) through ports, sealed from each other by O-rings[5] located on the module.

The primary benefit of this modular design is ease of maintenance in the field. If a module fails for any reason, it may be very quickly removed and replaced, with no disconnection and re-connection of signal wires or pneumatic tubes necessary.

[5] It is quite easy to dislodge these small-section, large-diameter O-rings from their respective grooves during re-assembly of the unit. Be very careful when inserting the module back into the housing!

As mentioned before, the feedback mechanism for this particular I/P transducer employs neither the force-balance nor the motion-balance principle. Rather, the negative feedback and balancing of this unit is done electronically rather than mechanically. The following diagram shows how this works:



An electronic pressure sensor continuously monitors the output pressure, with its signal being electronically compared to the input (4-20 mA) signal by the control circuit to check for equivalence. If the output does not match the input, the control circuit drives the deflector motor with more or less current as needed, to deflect the air jet more or less as it exits one nozzle and is intercepted by the other to stimulate the pneumatic amplifying relay. Thus, we see the "balancing" internal to this I/P is done electronically rather than mechanically as it was in the other I/P relays (Foxboro model E69, Fisher model 546) explored in this section.

Electronic components are less likely to drift in their calibration, and are less susceptible to the effects of mechanical vibration and mounting orientation, than mechanical balancing components.

14.6 Proper care and feeding of pneumatic instruments

Perhaps the most important rule to obey when using pneumatic instruments is to *maintain clean and dry instrument air.* Compressed air containing dirt, rust, oil, water, or other contaminants will cause operational problems for pneumatic instruments. First and foremost is the concern that tiny orifices and nozzles inside the pneumatic mechanisms will clog over time. Clogged orifices tend to result in decreased output pressure, while clogged nozzles tend to result in increased output pressure. In either case, the "first aid" repair is to pass a welding torch tip cleaner through the plugged hole to break loose the residue or debris plugging it.

Moisture in compressed air tends to corrode metal parts inside pneumatic mechanisms. This corrosion may break loose to form debris that plugs orifices and nozzles, or it may simply eat through thin diaphragms and bellows until air leaks develop. Grossly excessive moisture will cause erratic operation as "plugs" of liquid travel through thin tubes, orifices, and nozzles designed only for air passage.

A common mistake made when installing pneumatic instruments is to connect them to a general-service ("utility") compressed air supply instead of a dedicated instrument-service compressed air system. Utility air systems are designed to supply air tools and large air-powered actuators with pneumatic power. These high-flow compressed air systems are often seeded with antifreeze and/or lubricating chemicals to prolong the operating life of the piping and air-consuming devices, but the same liquids will wreak havoc on sensitive instrumentation. Instrument air supplies should be sourced by their own dedicated air compressor(s), complete with automatic air-dryer equipment, and distributed through stainless steel, copper, or plastic tubing (never black iron or galvanized iron pipe!).

The worst example of moisture in an instrument air system I have ever witnessed is an event that happened at an oil refinery where I worked as an instrument technician. Someone on the operations staff decided they would use 100 PSI instrument air to purge a process pipe filled with acid. Unfortunately, the acid pressure in the process pipe exceeded 100 PSI, and as a result acid flushed backward into the instrument air system. Within days most of the pneumatic instruments in that section of the refinery failed due to accelerated corrosion of metal components within the instruments. The total failure of multiple instruments over such a short time could have easily resulted in a disaster, but fortunately the crisis was minimal. Once the first couple of faulty instruments were disassembled after removal, the cause of failure became evident and the technicians took action to flush the lines of acid before too many more instruments suffered the same fate.

Pneumatic instruments must be fed compressed air of the proper pressure as well. Just like electronic circuits which require power supply voltages within specified limits, pneumatic instruments do not operate well if their air supply pressure is too low or too high. If the supply pressure is too low, the instrument cannot generate a full-scale output signal. If the supply pressure is too high, internal failure may result from ruptured diaphragms, seals, or bellows. Many pneumatic instruments are equipped with their own local pressure regulators directly attached to ensure each instrument receives the correct pressure despite pressure fluctuations in the supply line.

Another "killer" of pneumatic instruments is mechanical vibration. These are precision mechanical devices, so they do not generally respond well to repeated shaking. At the very least, calibration adjustments may loosen and shift, causing the instrument's accuracy to suffer. At worst, actual failure may result from component breakage[6].

[6] Having said this, pneumatic instruments can be remarkably rugged devices. I once worked on a field-mounted

14.7 Advantages and disadvantages of pneumatic instruments

The disadvantages of pneumatic instruments are painfully evident to anyone familiar with both pneumatic and electronic instruments. Sensitivity to vibration, changes in temperature, mounting position, and the like affect calibration accuracy to a far greater degree for pneumatic instruments than electronic instruments. Compressed air is an expensive utility – much more expensive per equivalent watt-hour than electricity – making the operational cost of pneumatic instruments far greater than electronic. The installed cost of pneumatic instruments can be quite high as well, given the need for special (stainless steel, copper, or tough plastic) tubes to carry supply air and pneumatic signals to distant locations. The volume of air tubes used to convey pneumatic signals over distances acts as a low-pass filter, naturally damping the instrument's response and thereby reducing its ability to respond quickly to changing process conditions. Pneumatic instruments cannot be made "smart" like electronic instruments, either. With all these disadvantages, one might wonder why pneumatic instruments are still used at all in modern industry.

Part of the answer is legacy. For an industrial facility built decades ago, it makes little sense to replace instruments that still work just fine. The cost of labor to remove old tubing, install new conduit and wires, and configure new (expensive) electronic instruments often is not worth the benefits.

However, pneumatic instruments actually enjoy some definite technical advantages which secure their continued use in certain applications even in the 21^{st} century. One decided advantage is the *intrinsic safety* of pneumatic field instruments. Instruments that do not run on electricity cannot generate electrical sparks. This is of utmost importance in "classified" industrial environments where explosive gases, liquids, dusts, and powders exist. Pneumatic instruments are also self-purging. Their continual bleeding of compressed air from vent ports in pneumatic relays and nozzles acts as a natural clean-air purge for the inside of the instrument, preventing the intrusion of dust and vapor from the outside with a slight positive pressure inside the instrument case. It is not uncommon to find a field-mounted pneumatic instrument encrusted with corrosion and filth on the outside, but factory-clean on the inside due to this continual purge of clean air. Pneumatic instruments mounted inside larger enclosures with other devices tend to protect them all by providing a positive-pressure air purge for the entire enclosure.

Some pneumatic instruments can also function in high-temperature and high-radiation environments that would damage electronic instruments. Although it is often possible to "harden" electronic field instruments to such harsh conditions, pneumatic instruments are practically immune by nature.

An interesting feature of pneumatic instruments is that they may operate on compressed gases other than air. This is an advantage in remote natural gas installations, where the natural gas itself is sometimes used as a source of pneumatic "power" for instruments. So long as there is compressed natural gas in the pipeline to measure and to control, the instruments will operate. No air compressor or electrical power source is needed in these installations. What *is* needed, however,

pneumatic controller attached to the same support as a badly cavitating control valve. The vibrations of the control valve transferred to the controller through the support, causing the baffle to hammer repeatedly against the nozzle until *the nozzle's tip had been worn down to a flattened shape*. Remarkably, the only indication of this problem was the fact the controller was having some difficulty maintaining setpoint. Other than that, it seemed to operate adequately! I doubt any electronic device would have fared as well, unless completely "potted" in epoxy.

is good filtering equipment to prevent contaminants in the natural gas (dirt, debris, liquids) from causing problems within the sensitive instrument mechanisms.

References

Patrick, Dale R. and Patrick, Steven R., *Pneumatic Instrumentation*, Delmar Publishers, Inc., Albany, NY, 1993.

Chapter 15

Digital data acquisition and networks

The advent of digital electronic circuitry has brought a steady stream of technological progress to industrial instrumentation. From early applications of digital computing in the 1960's to the first distributed control systems (DCS) in the 1970's to the "smart" transmitter revolution of the 1980's, digital technology has expanded on the capabilities and information-sharing ability of measuring and control instruments. It is the purpose of this chapter to give a general overview of digital technology as it applies to data acquisition (measuring and recording process data) and digital communication, highlighting some of the more common standards used in industry. Subsequent chapters will be devoted to more in-depth discussions of specific digital instrumentation standards.

One of the greatest advantages of digital technology over analog is the ability to communicate vast amounts of data over a limited number of data channels. In the world of 4-20 mA signaling (or 3-15 PSI signaling, for that matter!) each pair of wires could communicate only *one* variable. In the world of digital networks, one pair of wires can communicate a nearly limitless number of variables, the only limit being the *speed* of that data communication[1].

This one-signal-per-channel limit of 4-20 mA analog signals represents a technological "bottleneck" restricting data transfer between instruments and control systems. While it certainly is possible to devote a dedicated wire pair to each and every variable in an instrument system, this is very expensive to do. It is particularly cumbersome for instruments generating multiple variables of measurement, such as Coriolis flowmeters which simultaneously measure process fluid mass flow rate, fluid density, and fluid temperature; or "smart" valve positioners which continuously measure the stem position, actuator pressure(s), supply pressure, and temperature of a control valve. The data-rich capabilities of digital field instruments demands a digital form of communication to overcome the "bottleneck" of analog 4-20 mA signals.

Rosemount's HART standard was a valiant attempt to provide the "best of both worlds" in industrial instrumentation. With HART digital signals superimposed on 4-20 mA analog signals, one could retain the simplicity and certainty of analog signaling while enjoying the multi-variable

[1] The technical term for the "speed limit" of any data communications channel is *bandwidth*, usually expressed as a frequency (in Hertz).

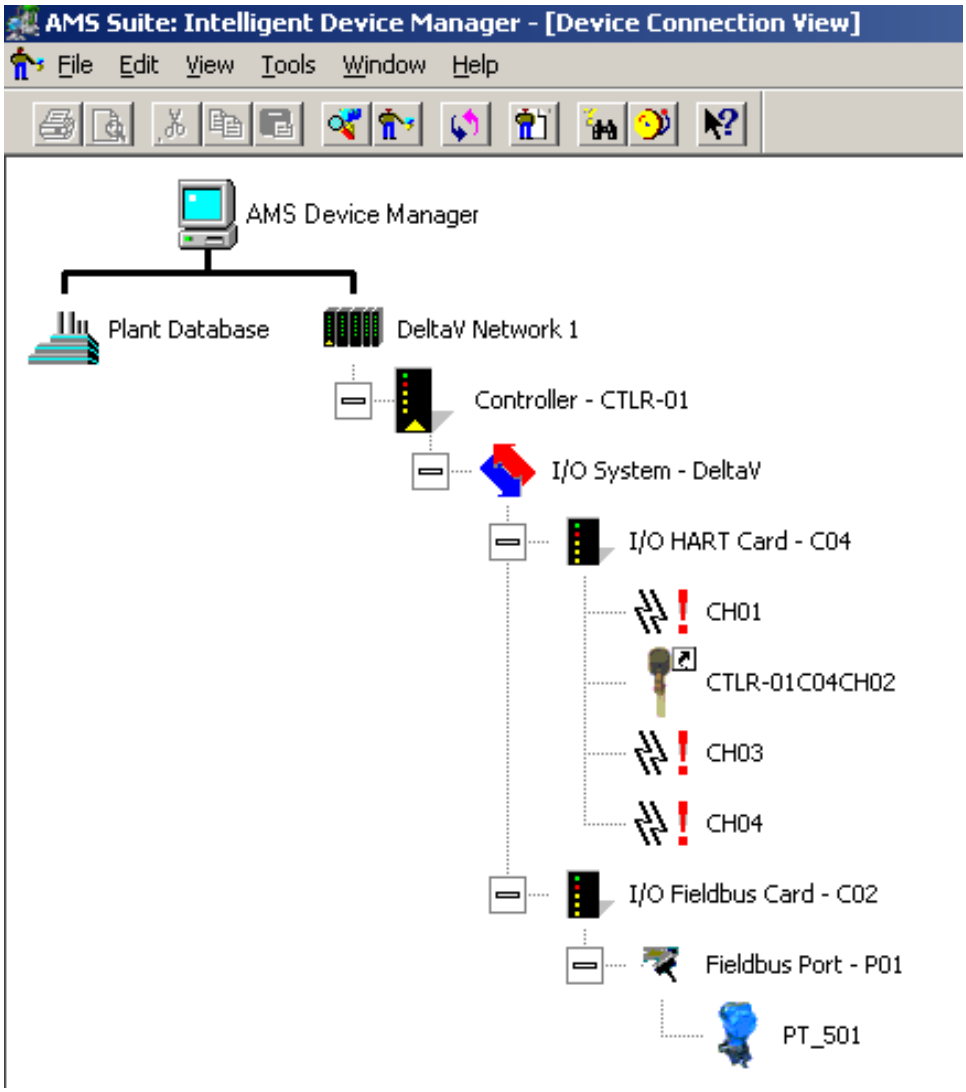
communication benefits that digital had to offer. However, HART communication is rather slow by any standard, restricting its use to maintenance (range changes, diagnostic data polling) and process control for slow processes only.

There exist many different digital communication standards (generally called "fieldbuses") designed to interconnect industrial instruments. An incomplete list is shown here:

- HART
- Modbus
- FOUNDATION Fieldbus
- Profibus PA
- Profibus DP
- Profibus FMS
- AS-I
- CANbus
- ControlNET
- DeviceNet
- BACnet

The utility of digital "fieldbus" instruments becomes apparent through the host system these instruments are connected to (typically a *distributed control system*, or *DCS*). Fieldbus-aware host systems usually have means to provide instrument information (including diagnostics) in very easy-to-navigate formats.

For example, the following screenshot shows the field instrument devices connected to a small-scale DCS used in an educational lab. Each instrument appears as an icon, which may be explored further simply by pointing-and-clicking with the mouse[2]:



[2]The host system in this case is an Emerson DeltaV DCS, and the device manager software is Emerson AMS.

Another application of digital communication technology to industrial measurement and control is what is generally referred to as a *SCADA* ("Supervisory Control And Data Acquisition") system. A SCADA system might be thought of as a distributed control system (DCS) spread over a geographically large area, such as across the span of a city or even across national borders. Typical applications of SCADA technology include:

- Electric power generation and distribution (power line, substation) systems
- Water and wastewater treatment and distribution (water line, pumping stations) systems
- Gas and oil exploration and distribution (pipeline) systems
- Large-scale agricultural (irrigation, harvesting) systems
- Storage tank monitoring systems

Process data in a SCADA system is sensed by various measurement devices (transmitters), converted to digital form by a device called an *RTU* ("Remote Terminal Unit"), and communicated to one or more *MTUs* ("Master Terminal Units") at a central location where human operators may monitor the data and make command decisions.

If the flow of information is one-way (simplex, from measurement devices to human operators), the system is more properly referred to as a *telemetry* system rather than a SCADA system. "SCADA" implies two-way (duplex) information flow, where human operators not only monitor process data but also issue commands back to the remote terminal units to effect change.

The saying "necessity is the mother of invention" certainly holds true for the development of remote telemetry and SCADA systems. The need for remote monitoring and control of electric power distribution systems led to the development of "power line carrier" analog telemetry systems as far back in time as the 1940's. These systems superimposed high-frequency (50 kHz to 150 kHz) "carrier" signals on low-frequency (50 Hz and 60 Hz) power line conductors to communicate such basic information as human voice (like a telephone network, only dedicated for power system operators), power flow (wattmeter, MVAR meter) monitoring, and protective relay (automatic trip) controls. These telemetry systems were among the first to enjoy the benefits of digital technology in the 1960's. Large-scale electric power systems simply cannot be operated safely and effectively without remote data monitoring and control, and this operational necessity pushed technological development of telemetry and SCADA systems beyond their small-scale (industrial manufacturing) counterparts.

Whether it is a "smart" temperature transmitter, a panel-mounted process controller with Ethernet communication capability, a variable-speed electric motor drive with Modbus signaling, a large-scale DCS controlling an oil refinery, or a SCADA system monitoring a power distribution system spanning across national borders, digital measurement and communication is an essential part of modern measurement and control systems. This chapter focuses on some of the basic principles of digital data formatting and communication, referencing practical applications wherever possible.

15.1 Digitization of analog quantities

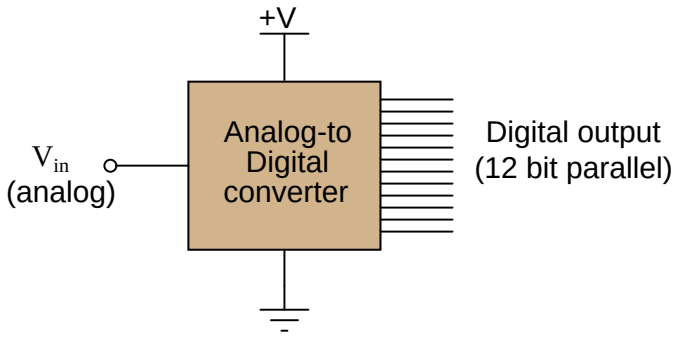
Process measurements are usually of an analog nature: the temperature of a furnace, the rate of fluid flow through a pipe, the pressure of a fluid, etc. These are all analog quantities: infinitely variable, not discrete. Although some process measurements may be discrete (e.g. counting the number of units passed by on a conveyor belt), the majority of measurements in the industrial world are analog.

In order for any digital device to successfully interface with an analog signal, that signal must be *digitized* by means of an *analog-to-digital converter* or *ADC*. This section will not endeavor to explore the intricate details of ADC circuitry, but merely to discuss ADC performance in the context of process measurements.

Many of the concerns discussed in this section are relevant to circuits converting digital values into analog signals as well. These *digital-to-analog converters*, or *DAC*s, are generally used to produce the analog drive signals required of final control elements (e.g. the output of a digital PID controller driving a 4-20 mA analog signal to a control valve positioner).

15.1.1 Resolution

In its simplest form, an ADC is an electronic circuit receiving an analog voltage signal input and generating a multi-bit binary (digital) output. Perhaps the most obvious measure of ADC performance, then, is how many bits of output are provided:



The ADC shown in the above illustration is a 12-bit unit. This means its digital output ranges from `000000000000` to `111111111111` (`000` hexadecimal to `FFF` hexadecimal, or 0 decimal to 4095 decimal). Although the ADC shown outputs its digital data in *parallel* form (with separate terminals for the 12 individual bits), many modern ADC chips are designed for *serial* data output, where a single terminal generates a sequential series of bits timed to the pulse of a clock signal.

Supposing this 12-bit ADC has an analog input voltage range of 0 to 10 volts, how do we relate any given digital number value to a voltage value, or visa-versa? The key here is to understand that the 12-bit *resolution* of this ADC means it has 2^{12}, or 4096, discrete output states. The 10 volt DC input range is therefore divided up into $2^{12} - 1$, or 4095, discrete increments:

$$\text{Analog resolution} = \frac{\text{Analog span}}{2^n - 1}$$

Where,

n = Number of binary bits in the output "word"

For our hypothetical 0-10 VDC, 12-bit converter, the analog resolution is 2.442 millivolts. Thus, for any analog signal between 0 mV and 2.442 mV, the ADC's output should be zero (binary `000000000000`); for any analog signal between 2.442 mV and 4.884 mV, the ADC's output should be one (binary `000000000001`); and so on.

The digital output value from an industrial ADC is commonly referred to as a *count*. The word "count" is used in this context as a unit of measurement. For instance, if we subjected our 12-bit ADC to a full-scale input signal of 10 VDC, we would expect to see a full-scale digital output (binary `111111111111`) of 4095 "counts." Since most ADC circuits are designed to be linear, the mathematical relationship between input voltage and digital output "counts" is a simple proportionality:

$$\frac{V_{in}}{V_{fullscale}} = \frac{\text{Counts}}{2^n - 1}$$

We may use this formula to generate a partial table of input and output values for our 0-10 VDC, 12-bit ADC:

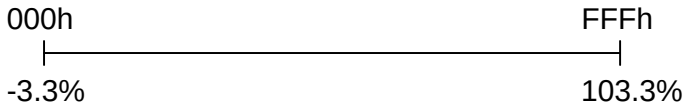
V_{in}	Counts (decimal)	Counts (hex)
0 V	0	`000`
2.46 mV	1	`001`
3.85 V	1576	`628`
4.59 V	1879	`757`
6.11 V	2502	`9C6`
9.998 V	4094	`FFE`
10 V	4095	`FFF`

In order to calculate a digital count value from a given input voltage, simply divide that voltage value by the full-scale voltage, then multiply by the full-scale count value and round down to the nearest whole number. For any given voltage value input to the ADC, there is exactly one corresponding output "count" value. The converse cannot be said, however: for any given output "count" value, there is actually a small range of possible input voltages (that range being the analog resolution of the ADC, in this case 2.442 mV).

To illustrate, let us take one of the table entries as an example: an analog input of 6.11 volts should yield a digital output of (precisely) 2502 counts. However, a digital output of 2502 counts could represent any analog input voltage ranging between 6.10989 volts and 6.11233 volts. This uncertainty is inherent to the process of "digitizing" an analog signal: by using a discrete quantity to represent something infinitely variable, some detail is inevitably lost. This uncertainty is referred to as *quantization error*: the (potential) error resulting from "quantizing" (digitizing) an inherently analog quantity into a discrete representation.

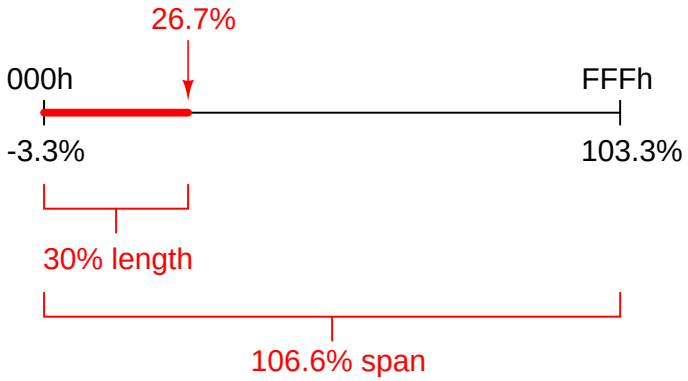
Quantization error may be reduced (but never eliminated) by using an ADC with greater resolution. A 14-bit ADC operating over the same 0-10 VDC analog input range would have approximately one-quarter the uncertainty of the 12-bit ADC (0.610 mV instead of 2.442 mV). A 16-bit ADC's uncertainty would only be (approximately) one-sixteenth that of the 12-bit ADC. The number of bits chosen for any particular ADC application is therefore a function of how precise the digitization must be.

Often you will encounter digital instruments where the digital "count" scale maps to a live-zero analog range. For example, the Siemens model 353 process controller represents process variable, setpoint, and output ("valve") percentages on a scale of -3.3% to 103.3% with a 12-bit ADC count. For this controller, a digital count of 0 represents an analog signal of -3.3%, and a digital count value of FFF hexadecimal represents 103.3%. We may show the relationship between these two scales in graphical form, like a number line:



Converting a digital count value to its respective analog percentage (or visa-versa) follows the same procedure used to convert values between any two representative scales where one of the scales has a "live zero:" take the given value and convert to a percentage of span (subtracting any live zero before dividing by the span), then calculate the other value based on that percentage of span (adding any live zero after multiplying by the span).

For example, to determine what the representative digital count value would be for an analog signal having a percentage value of 26.7% on this scale, all we need to do is determine how much of the maximum digital count value this is:

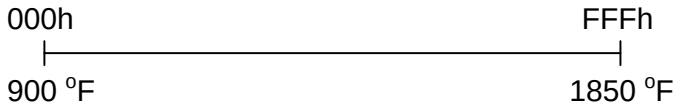


Here, the analog signal value of 26.7% is 30% away from the scale's starting point of -3.3%. Compared to the scale's span of 106.6%, this is a value of:

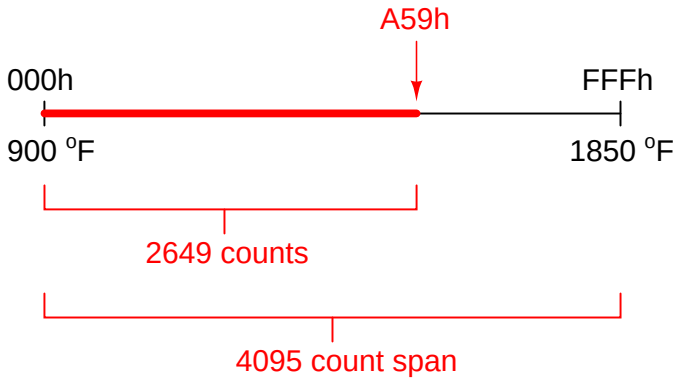
$$\frac{26.7 - (-3.3)}{106.6} = \frac{30}{106.6} = 0.2814$$

Since we know a 12-bit binary count goes from 0 to 4095, and our current 26.7% analog signal value places us at 28.14% of the 4095 full-scale count, the count value for this analog signal should be 1152, or 480 hexadecimal (480h).

Similarly, if we knew the range of this 12-bit ADC in actual process engineering units, we could translate between ADC counts and the process value by the same method. Suppose we used one of these same controllers to display the temperature of a furnace, where the lower- and upper-range values were 900 deg F and 1850 deg F, respectively. We could relate the ADC count to the temperature scale using the same "number line" format as the previous example:



Suppose the ADC count for a certain furnace temperature was A59 hexadecimal (A59h), equal to 2649 in decimal form:



To convert this count value into a temperature, first we determine its percentage of span:

$$\frac{2649}{4095} = 0.6469$$

Next, we calculate the how much of the temperature scale's span this equates to:

$$(0.6469)(1850 - 900) = 614.5$$

Thus, this temperature is 614.5 degrees hotter than the bottom of the scale (at 900 deg F). Adding the live zero value of 900 degrees F, we arrive at a furnace temperature of 1514.5 degrees F.

15.1.2 Sampling rate

The next major performance metric for analog signal digitization is how often the analog signal gets converted into digital form. Each time an ADC circuit "samples" its analog input signal, the resulting digital number is fixed until the next sample. This is analogous to monitoring a continuously moving process by taking a series of still-photographs. Any changes happening to the analog signal between sampling events are not detected by the converter, and therefore are not represented in the digital data coming from the converter.

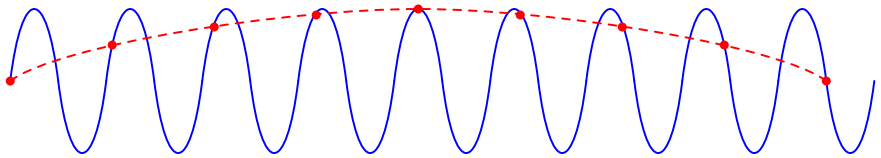
It stands to reason, then, that the sampling rate of any ADC must be at least as often as significant changes are expected to take place in the analog measurement. According to the *Nyquist Sampling Theorem*, the absolute minimum sample rate necessary to capture an analog waveform is twice the waveform's fundamental frequency. More realistic is to have the ADC sample the waveform *ten times* or more per cycle.

In general electronics work, for example with the design of electronic test equipment such as digital multimeters (DMMs) and digital storage oscilloscopes (DSOs), sampling rates must be rather fast. Modern digital oscilloscopes may have sampling rates in the *billions* of samples per second, to allow for the successful digitization of radio-frequency analog signals.

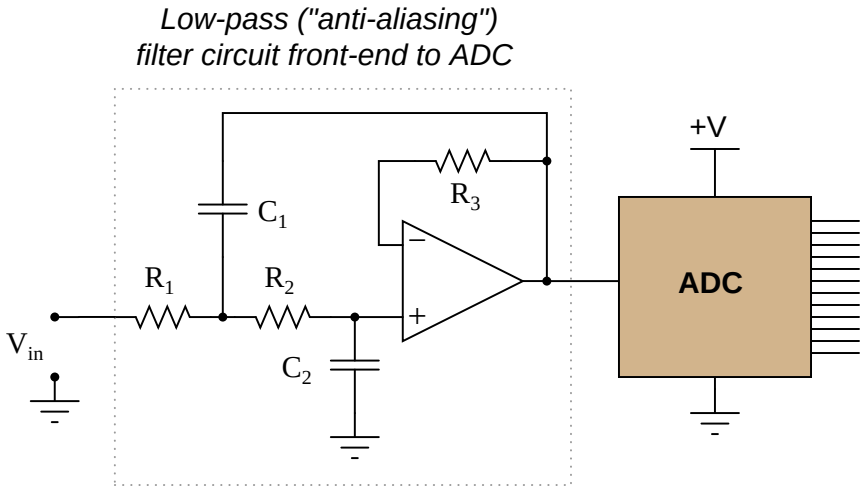
Industrial process measurements are far more forgiving than measurements commonly performed on an electronic technician's workbench, thankfully. The temperature of a large furnace may be adequately sampled at a rate of only once per minute, if need be. Even "fast" feedback processes such as liquid flow and pressure control may be controlled with reasonable stability by digital systems sampling just a few times per second.

A sampling rate that is too slow (infrequent) may detrimentally affect an instrumentation in more than one way. First, the time between samples is *dead time* to the system: time during which the digital system will be completely unresponsive to any changes in process measurement. Excessive dead time in an alarm system means an unnecessary time delay between the alarm event and the alarm signal. Excessive dead time in a feedback control loop leads to oscillation and instability. Another detrimental effect of low sampling rate is something called *aliasing*: a condition where the digital system "thinks" the frequency of an analog signal is far lower than it actually is.

A dramatic example of aliasing is shown in the following illustration, where a sinusoidal signal (colored blue) is sampled at intervals slightly slower than once per cycle (samples marked by red dots). The result (the red, dashed curve) is what appears to be a much lower-frequency signal as seen by the digital system, which only "sees" the values represented by the red dots:



The troubling nature of aliasing is that it causes the ADC to report a *completely incorrect*, yet *completely plausible* signal. One simple way to avoid aliasing in an ADC circuit is to place an analog low-pass filter circuit before the ADC's input, preventing any analog signals with frequencies beyond the Nyquist limit to pass through to the ADC. Such a "front-end" circuit is called an *anti-aliasing filter*:



Aliasing may still occur within digital systems, though, if one portion of a system "samples" the digital output of another portion at a substantially lower frequency. An example of this might be the rate at which a digital control system (such as a DCS) polls a process variable value collected by a digital sensor network (such as a network of radio-linked process transmitters, or digital fieldbus transmitters). If the DCS polling rate is sufficiently slow compared to the frequency of the signal reported by the digital transmitters, aliasing may result. The best guard against such potential troubles is to synchronize the sampling rates throughout the system.

15.2 Digital data communication theory

One of the great benefits of digital technology is the ability to *communicate* vast amounts of information over networks. This very textbook you are reading was transmitted in digital form over the electronic network we call the *internet*: a feat nearly impossible with any sort of analog electronic technology. The main benefit of digital data communication in industrial control is simple: no longer must we dedicate a single pair of wires to each and every variable we wish to measure and control in a facility as is necessary with analog (4-20 mA) signaling. With digital signaling, a single pair of wires or coaxial cable is able to convey a theoretically unlimited number of data points.

This benefit comes at a price, though: in order to communicate multiple variables (data points) over a single channel (wire pair), we must transmit and receive those signals one at a time. This means a digital communications system will necessarily exhibit some degree of *time delay* in acquiring, transmitting, receiving, and interpreting a signal. Analog systems, by contrast, are virtually instantaneous[3]. Thus, we see a contrast between analog and digital communication pitting channel capacity against speed:

Analog	**Digital**
Only one signal per channel	Many signals per channel possible
Instantaneous	Time-delayed

With modern electronic technology it is possible to build digital communication systems that are so fast, the time delays are negligible for most industrial processes, which renders the second comparison (instantaneous versus time-delayed) moot. If time is no longer an issue, the advantage that digital communication has over analog in terms of channel usage makes it the superior choice[4].

Another important advantage of digital data communication for industrial processes is increased *noise immunity*. Analog data is *continuous* by nature: a signal of 11.035 milliamps has a different meaning than a signal of 11.036 milliamps, because any measurable increment in signal represents a corresponding increment in the physical variable represented by that signal. A voltage value in a 0-5 volt digital signaling system of 0.03 volts, however, means *the exact same thing* as a voltage value of 0.04 volts: either one is still interpreted as a "0" or "low" state. *Any* amount of electrical noise imposed on an analog signal corrupts that signal to some degree. A digital signal, however, may tolerate a substantial amount of electrical noise with no corruption whatsoever.

[3] To be fair, there is such a thing as a time-multiplexed analog system for industrial data communication (I've actually worked on one such system, used to measure voltages on electrolytic "pots" in the aluminum industry, communicating the voltages across hundreds of individual pots to a central control computer).

[4] There is, of course, the issue of *reliability*. Communicating thousands of process data points over a single cable may very well represent a dramatic cost savings in terms of wire, junction boxes, and electrical conduit. However, it also means you will lose all those thousands of data points if that one cable becomes severed! Even with digital technology, there is still reason sometimes to under-utilize the bandwidth of signal cables.

Not surprisingly, though, the noise immunity enjoyed by digital signals comes with a price: a sacrifice in *resolution*. Analog signals are able to represent the smallest imaginable changes because they are continuously variable. Digital signals are limited in resolution by the number of bits in each data "word." Thus, we see another contrast between analog and digital data representation:

Analog	**Digital**
Corrupted by any amount of noise	Immune to certain (limited) amounts of noise
Unlimited resolution	Limited resolution

With modern digital electronic technology, however, the "limited resolution" problem is almost nonexistent. 16-bit converter chipsets are commonly available today for input/output (I/O) modules on digital systems, providing a resolution of 2^{16} (65,536) counts, or $\pm$ 0.00153%, which is good enough for the vast majority of industrial measurement and control applications.

This section will focus on *serial* data transmission, as opposed to *parallel*. In order to transmit digital data in parallel form, the number of wires scales directly with the number of bits in each data "word." For example, if a 16-bit ADC chip were to communicate its data to some other digital device using a parallel network, it would require a cable with 16 wires (plus a common "ground" wire) *at minimum*. Since this approach undercuts the "fewer wires" advantage that digital communications theoretically enjoys over analog communication, parallel data transmission is rarely seen in industry except for within the internal construction of a digital device (e.g. a parallel data bus inside a personal computer, or inside a PLC or DCS rack).

In serial communications systems, digital data is sent over a wire pair (or fiber optic cable, or radio channel) *one bit at a time*. A 16-bit digital "word" (two *bytes* in length) then will require a succession of 16 bits transmitted one after the other in time. How we represent each bit as an electrical signal, how we arrange those bits in time to group them into meaningful "words," and how multiple devices share access to a common communications channel, is our next subject of exploration: the technical details of serial data communication.

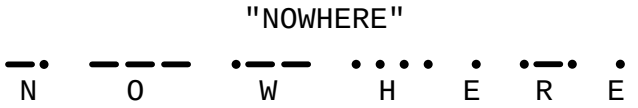
15.2.1 Serial communication principles

The task of encoding real-life data as a series of on-and-off electrical signals, and then sending those signals long distances over cables (or optical fibers, or radio waves) requires mutually-agreed *standards* for the encoding, the "packaging" of those bits, the speed at which the bits are sent, methods for multiple devices to use a common channel, and a host of other concerns. This subsection will delineate the major areas of concern where standards must be established before digital devices may communicate between each other serially. We begin with a brief exploration of some of the standard used in early *telegraph* systems.

An early form of digital communication was *Morse Code*, used to communicate alpha-numerical information as a series of "dots" and "dashes" over telegraph[5] systems. Each letter in the alphabet, and each numerical digit (0 through 9) was represented in Morse Code by a specific series of "dot" and "dash" symbols, a "dot" being a short pulse and a "dash" being a longer pulse. A similar code system called the *Continental Code* was used for early radio ("radiotelegraph") communications.

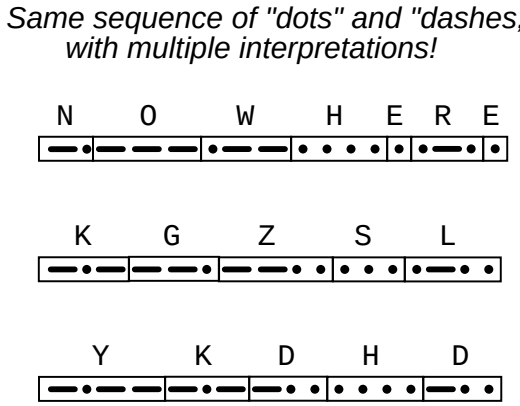
As primitive as these codes were, they encapsulated many of the basic principles we find in modern digital serial communication systems. First, a system of codes needed to exist to represent the letters and numbers needed to spell messages in English, which the general American public conversed in. Next, there needed to be some way to represent these characters individually so that one could distinguish the end of one character and the beginning of the next character.

For example, consider the Continental Code encoding for the word `NOWHERE`. By placing an extra space (a pause in time) between characters, it is easy to represent individual characters in the message:



[5] I do not expect any reader of this book to have firsthand knowledge of what a "telegraph" is, but I suspect some will have never heard of one until this point. Basically, a telegraph was a primitive electrical communication system stretching between cities using a keyswitch at the transmitting end to transmit on-and-off pulses and a "sounder" to make those pulses audible on the receiving end. Trained human operators worked these systems, one at the transmitting end (encoding English-written messages into a series of pulses) and one at the receiving end (translating those pulses into English letters).

If this space between characters were not present, it would be impossible to determine the message with certainty. By removing the spaces, we find multiple non-sensical interpretations are possible for the same string of "dots" and "dashes:"



For that matter, it is even possible to confuse the meaning of the text string "NOWHERE" when the individual characters are properly interpreted. Does the string of characters say "nowhere," or does it say "now here"?

This simple example illustrates the need for *delimiting* in serial data communication. Some means must be employed to distinguish individual groups of bits (generally called *frames* or *packets*) from one another, lest their meanings be lost. In the days when human operators sent and interpreted Morse and Continental code messages, the standard delimiter was an extra time delay (pause) between characters, and between words. This is not much different from the use of whitespace to delineate words, sentences, and paragraphs typed on a page. Sentenceswouldcertainlybeconfusingtoreadifnotforspaces!

In later years, when *teletype* machines were designed to replaced skilled Morse operators, the concept of frame delineation had to be addressed more rigorously. These machines consisted of a typewriter-style keyboard which marked either paper strips or pages with dots corresponding to a 5-bit code called the *Baudot code*. The paper strip or sheets were then read electrically and converted into a serial stream of on-and-off pulses which were then transmitted along standard telegraph circuit lines. A matching teletype machine at the receiving end would then convert the signal stream into printed characters (a telegram). Not only could unskilled operators use teletype machines, but the data rate far exceeded what the best human Morse operators could achieve[6]. However, these machines required special "start" and "stop" signals to synchronize the communication of each character, not being able to reliably interpret pauses like human operators could.

Interestingly, modern asynchronous[7] serial data communication relies on the same concept of

[6] A test message sent in 1924 between two teletype machines achieved a speed of 1920 characters per minute (32 characters per second), sending the sentence fragments "THE WESTERN ELECTRIC COMPANY", "FRESHEST EGGS AT BOTTOM MARKET PRICES", and "SHE IS HIS SISTER".

[7] "Asynchronous" refers to the transmitting and receiving devices not having to be in perfect synchronization in order for data transfer to occur. Every industrial data communications standard I have ever seen is asynchronous rather than synchronous. In synchronous serial networks, a common "clock" signal maintains transmitting and receiving devices in a constant state of synchronization, so that data packets do not have to be preceded by "start" bits or followed by "stop" bits. Synchronous data communication networks are therefore more efficient (not having

"start" and "stop" bits to synchronize the transmission of data packets. Each new packet of serial data is preceded by some form of "start" signal, then the packet is sent, and followed up by some sort of "stop" signal. The receiving device(s) synchronize to the transmitter when the "start" signal is detected, and non-precision clocks keep the transmitting and receiving devices in step with each other over the short time duration of the data packet. So long as the transmitting and receiving clocks are close enough to the same frequency, and the data packet is short enough in its number of bits, the synchronization will be good enough for each and every bit of the message to be properly interpreted at the receiving end.

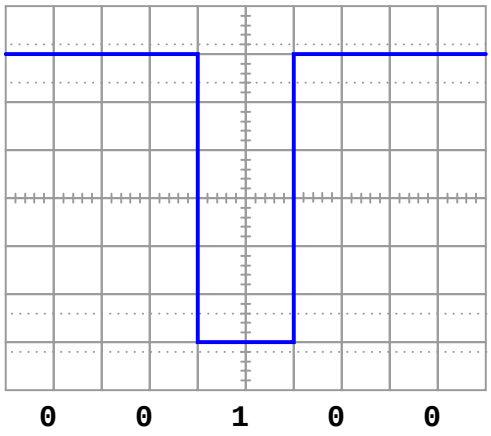
to include "extra" bits in the data stream) but also more complex. Most long-distance, heavy traffic digital networks (such as the "backbone" networks used for the Internet) are synchronous for this reason.

15.2.2 Physical encoding of bits

Telegraph systems were Boolean in nature: representing "dots" and "dashes" by one electrical state of the telegraph line, and pauses by another. When manually-actuated keyswitches were abandoned in favor of teletype machines, and Morse code abandoned in favor of the *Baudot* (5-bit) code for representing alphanumeric characters, the electrical nature of the telegraph (at least initially[8]) remained the same. The line would either be energized or not, corresponding to *marks* or *spaces* made on the teletype paper.

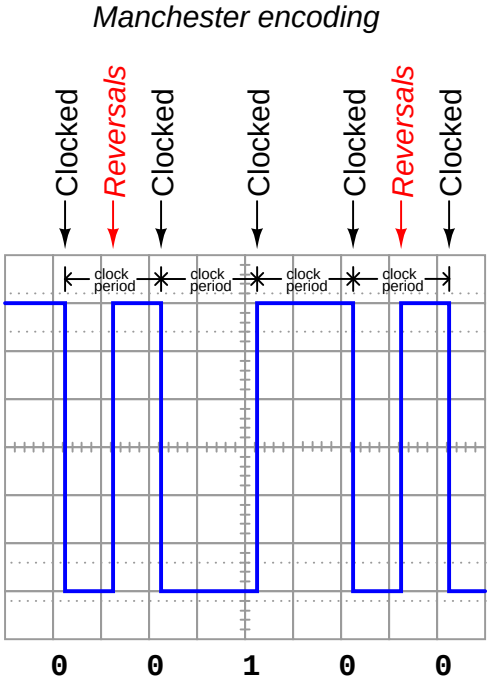
Many modern digital communication standards represent binary "1" and "0" values in exactly this way: a "1" is represented by a "mark" state and a "0" is represented by a "space" state. "Marks" and "spaces" in turn correspond to different voltage levels between the conductors of the network circuit. For example, the very common EIA/TIA-232 serial communications standard (once the most popular way of connecting peripheral devices to personal computers, formerly called RS-232) defines a "mark" (1) state as -3 volts between the data wire and ground, and a "space" (0) state as +3 volts between the data wire and ground. This is referred to as *Non-Return-to-Zero* or `NRZ` encoding:

Non-Return-to-Zero (NRZ) encoding



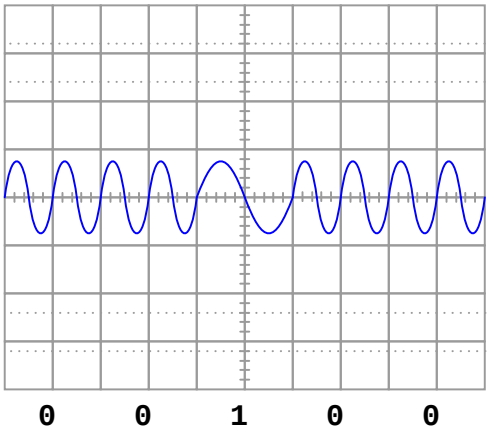
[8] Later versions of teletype systems employed audio tones instead of discrete electrical pulses so that many different channels of communication could be funneled along one telegraph line, each channel having its own unique audio tone frequency which could be filtered from other channels' tones.

This is not the only way to represent binary bits, though. An alternative method is to use an oscillating (square-wave) signal, counting *up* and *down* transitions (pulse edges) at specific times to represent 1 and 0 states. This is called *Manchester encoding*, and it is used in the 10 Mbps (10 million bits per second) version of *Ethernet* and in both the *FOUNDATION Fieldbus* "H1" and *Profibus* "PA" instrumentation network standards:



Yet another method for encoding binary 1 and 0 states is to use sine waves of different frequencies ("tone bursts"). This is referred to as *Frequency Shift Keying*, or *FSK*, and it is the method of encoding used in the HART[9] "smart" instrument communications standard.

Frequency Shift Key (FSK) encoding



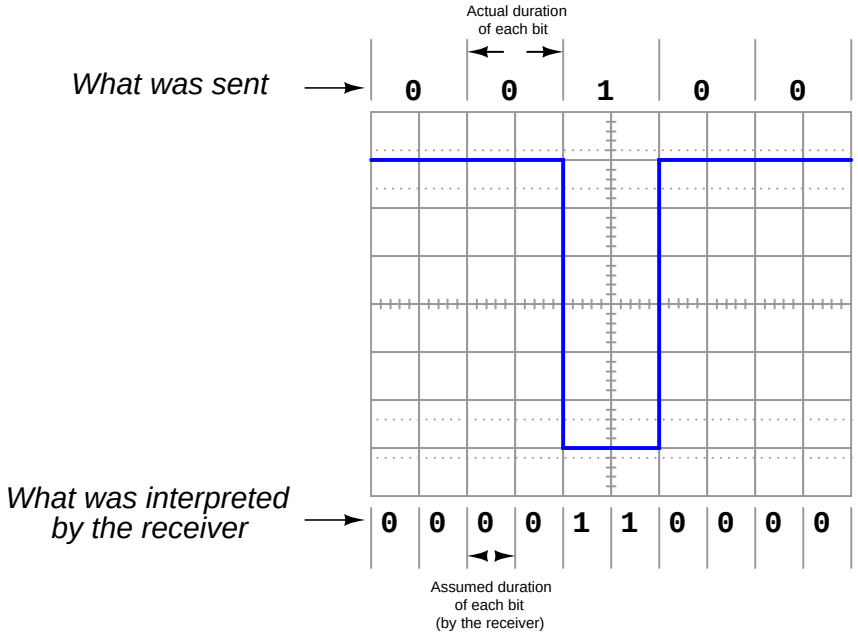
Other methods exist as well for encoding digital data along network cables, but these three are the most popular in industrial network standards.

[9] In the HART standard, a tone of 2200 Hz represents a "0" bit, while a tone of 1200 Hz represents a "1" bit.

15.2.3 Communication speed

In order to successfully communicate digital data along a network, there must not only be a standard agreed upon between transmitting and receiving devices for encoding bits (NRZ, Manchester, FSK, etc.), but there must also be a standard in place for the *speed* at which those bits will be sent. This is especially true for NRZ and FSK encoding, where the "clock" speed is not explicitly present in the signal[10].

For example, consider the confusion that could arise interpreting a NRZ signal if the transmission speed were assumed to be twice what it actually was:



Thus, one of the essential parameters in a serial data communication system is the *bit rate*, measured in *bits per second* (bps). Some communications standards have fixed bit rates, such as FOUNDATION Fieldbus H1 and Profibus PA, both standardized at exactly 31.25 kbps. Some, such as Ethernet, have a few pre-defined speeds (10 Mbps, 100 Mbps, 1 Gbps) defined by the specific transmitting and receiving hardware used. Others, such as EIA/TIA-232 may be arbitrarily set by the user at speeds ranging from 300 bps to over 115 kbps.

An older term sometimes used synonymously with bit rate is *baud rate*, however "bits per second" and "baud" are actually different things. "Baud" refers to the number of voltage (or current) alternations per second of time, whereas "bits per second" refers to the actual number of binary data bits communicated per second of time. Baud is useful when determining whether or not the bandwidth (the maximum frequency capacity) of a communications channel is sufficient for a certain communications purpose. In systems using NRZ encoding, the baud rate is equivalent[11] to the bit

[10] This is one of the advantages of Manchester encoding: it is a "self-clocking" signal.

[11] This is likely why "bit rate" and "baud rate" became intermingled in digital networking parlance: the earliest serial data networks requiring speed configuration were NRZ in nature, where "bps" and "baud" are one and the

rate: for a string of alternating bits (`010101010101`), there will be exactly one voltage transition for each bit. In systems using Manchester encoding, the worst-case[12] baud rate will be exactly *twice* the bit rate, with two transitions (one up, one down) per bit. In some clever encoding schemes, it is possible to encode multiple bits per signal transition, such that the bit rate will actually be greater than the baud rate.

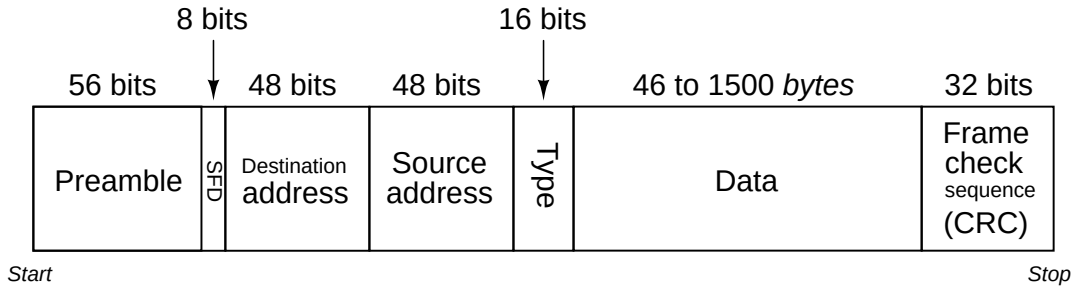
same.

[12]For Manchester encoding, "worst-case" is a sequence of identical bit states, such as `111111111111`, where the signal must make an extra (down) transition in order to be "ready" for each meaningful (up) transition representing the next "1" state.

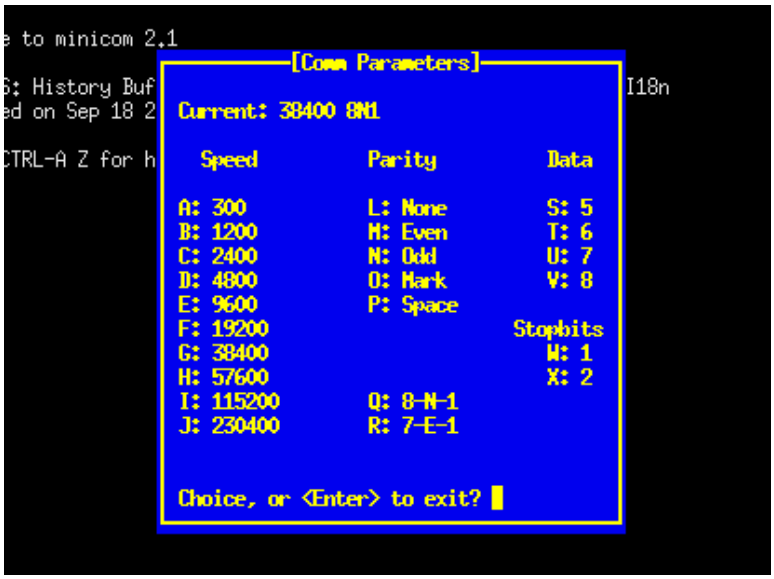
15.2.4 Data frames

As mentioned earlier in this section, serial data is usually communicated *asynchronously* in industrial networks. This means the transmitting and receiving hardware need not be in perfect synchronization to reliably send and receive digital data. In order for this to work, data must be sent in "frames" or "packets" of fixed (maximum) length, each frame preceded by a special "start" signal and concluded with a special "stop" signal. As soon as the transmitting device issues the "start" signal, the receiving device synchronizes to that start time, and runs at the pre-determined clock speed to gather the successive bits of the message until the "stop" signal is received. So long as the internal clock circuits of the transmitting and receiving devices are running at *approximately* the same speed, the devices will be synchronized closely enough to exchange a short message without any bits being lost or corrupted. There is such a thing as a *synchronous* digital network, where all transmitting and receiving devices are locked into a common clock signal so they cannot stray out of step with each other. The obvious advantage of synchronous communication is that no time need be wasted on "start" and "stop" bits, since data transfer may proceed continuously rather than in packets. However, synchronous communication systems tend to be more complex due to the need to keep all devices in perfect synchronization, and thus we see synchronous systems used for long-distance, high-traffic digital networks such as those use for Internet "backbones" and not for short-distance industrial networks.

Like bit rate, the particular scheme of start and stop bits must also be agreed upon in order for two serial devices to communicate with each other. In some networks, this scheme is fixed and cannot be altered by the user. Ethernet is an example of this, where a sequence of 64 bits (an alternating string of "1" and "0" bits ending with a "1, 1"; this is the "preamble" and "start frame delimiter" or "SFD" bit groups) is used to mark the start of a frame and another group of bits specifies the length of the frame (letting the receiver know ahead of time when the frame will end). A graphic description of the IEEE 802.3 standard for Ethernet data frames is shown here, illustrating the lengths and functions of the bits comprising an Ethernet frame:



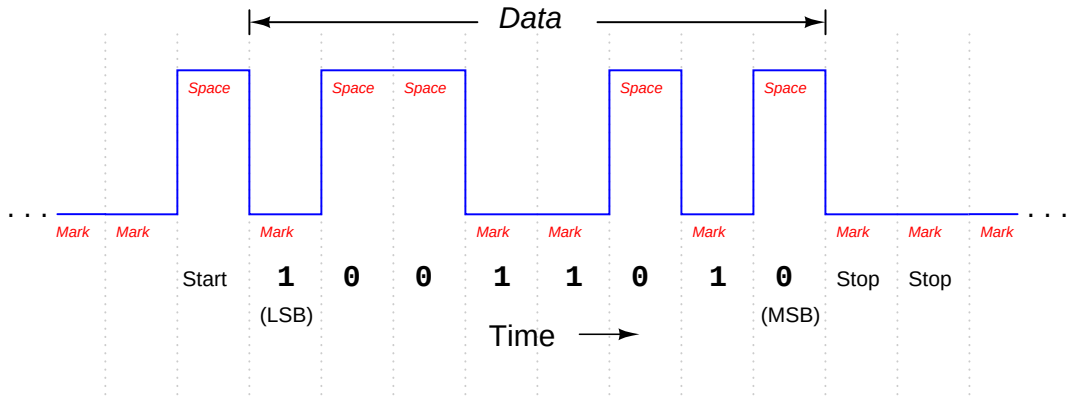
Other serial networks offer choices for the user to select regarding these parameters. One such example is EIA/TIA-232, where the user may specify not only the bit rate, but also how many bits will be used to mark the end of the data frame. It is imperative in such systems that *all* transmitting and receiving devices within a given network be configured exactly the same, so that they will all "agree" on how to send and receive data. A screenshot from a UNIX-based serial communication terminal program (called *minicom*[13]) shows these options:



In this particular screenshot, you can see the data rate options (extending from 300 bps all the way up to 230,400 bps!), the number of data bits (from 5 to 8), and the number of stop bits (1 or 2), all configurable by the user. Of course, if this program were being used for communication of data between two personal computers, *both* of those computers would need these parameters set identically in order for the communication to take place. Otherwise, the two computers would not be in agreement on speed, number of data bits, and stop bits; their respective data frames simply would not match.

[13] An equivalent program for Microsoft Windows is *Hyperterminal*. A legacy application, available for both Microsoft Windows and UNIX operating systems, is the serial communications program called *kermit*.

To give an example of an EIA/TIA-232 data frame might look like as a series of voltage states, consider this waveform communicating a string of eight bits (01011001), using NRZ encoding. Here, a single "start" marks the beginning of the data frame, while two successive "stop" bits end it. Also note how the bit sequence is transmitted "backwards," with the least-significant bit (LSB) sent first and the most-significant bit (MSB) sent last[14]:



Serial bitstream for the digital byte 01011001, where the least-significant bit (LSB) is sent first

Interestingly, the "mark" state (corresponding to a binary bit value of "1") is the default state of the communications channel when no data is being passed. The "start" bit is actually a space (0). This is the standard encoding scheme for EIA/TIA-232, EIA/TIA-485, and some other NRZ serial communication standards.

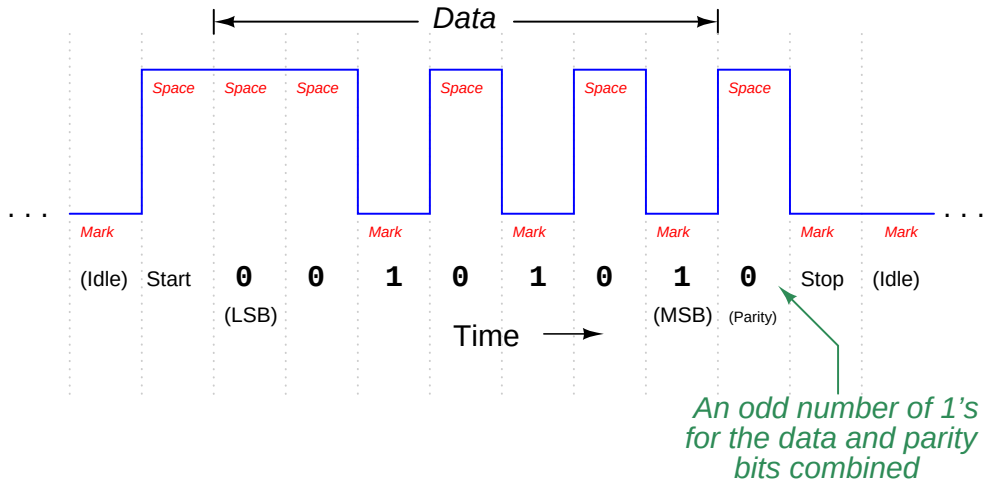
One of the options you probably noticed in the "minicom" terminal program screenshot was something called *parity*. This is a simple form of error-checking used in many serial communication standards. The basic principle is quite simple: an extra bit is added at the end of the data frame (just before the "stop" bits) to force the total number of "1" states to be either odd or even. For example, in the data stream just shown (10011010), there is an *even* number of "1" bits. If the computer sending this eight-bit data group were configured for "odd" parity, it would append an additional "1" to the end of that frame to make the total number of "1" bits odd rather than even. If the next data group were 11001110 instead (already having an odd number of "1" bits), the transmitting computer would have to attach a "0" parity bit on to the data frame in order to maintain an odd count of "1" bits.

The way this works to check for errors is for the receiving computer to count up all the "1" bits in each data frame (including the parity bit), and check to see that the total number is still odd (if the receiving computer is configured for odd parity just as the transmitting computer, which the two should *always* be in agreement). If any one bit somehow gets corrupted during transmission, the received frame will not have the correct parity, and the receiving computer will "know" something has gone wrong. Parity does not suggest *which* bit got corrupted, but it will indicate if there was a

[14] This is standard in EIA/TIA-232 communications.

single-bit[15] corruption of data, which is better than no form of error-checking at all.

The following example shows how parity-checking would work to detect a transmission error in a 7-bit data word. Suppose a digital device asynchronously transmits the character "T" using ASCII encoding ("T" = 1010100), with one start bit, one stop bit, and "odd" parity. Since the "start" bit is customarily a 0 state (space), the data transmitted in reverse order (LSB first, MSB last), the parity bit transmitted after the data's MSB, and the "stop" bit represented by a 1 state (mark), the entire frame will be the following sequence of bits: `0001010101`. Viewed on an oscilloscope display where a negative voltage represents a "mark" and a positive voltage represents a "space," the transmitted data frame will look like this:



Note how the parity bit in this particular frame is a 0, because the parity type is set for "odd," and the 7-bit data word already has an odd number of 1 bits.

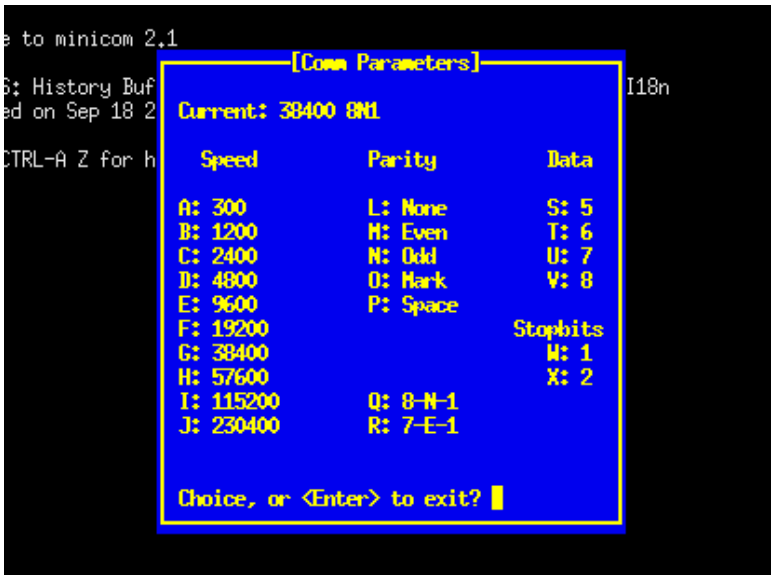
[15] It should take only a moment or two or reflection to realize that such a parity check cannot detect an *even* number of corruptions, since flipping the states of any *two* or any *four* or any *six* (or even all eight!) bits will not alter the evenness/oddness of the bit count. So, parity is admittedly an imperfect error-detection scheme. However, it is certainly better than nothing!

Now suppose this transmitted frame encounters a significant amount of electrical noise as it travels to the receiving device. If the frame reaches the receiver as shown in the next illustration, the receiver will interpret the message incorrectly:

One of the bits has been corrupted by noise, such that the frame is received as `0001000101` instead of `0001010101` as it was transmitted. When the receiver goes to count the number of 1 bits in the message (data plus parity bit, disregarding start and stop bits), it will count an even number of 1's instead of an odd number of 1's. Since the receiver is also set for "odd" parity to match the transmitter, it expects an odd number of 1's in the received message. Thus, it "knows" there is a problem somewhere in this transmission, because the received parity is not odd as it should be.

Parity-checking does not tell us *which* bit is corrupted, but it does indicate that *something* has gone wrong in the transmission. If the receiving device is programmed to take action on receipt of a non-matching parity, it may reply with a request for the transmitting device to re-send the data as many times as necessary until the parity is correct.

If we look at the "minicom" terminal screenshot again to analyze the parity options, we see there are several to choose from:



The five options for parity in this program include *None*, *Even*, *Odd*, *Mark*, and *Space*. "No" parity is self-explanatory: the transmitting computer does not attach an extra bit for parity at all, and the receiving computer does not bother to check for it. Since the inclusion of a parity bit does add to the bulk of a data frame, it has the unfortunate effect of slowing down communications (more bit "traffic" occupying the channel than would otherwise need to be), thus the option to waive parity altogether for a more compact (faster) data frame. "Even" and "Odd" parity options work as previously described, with the transmitting computer adding a parity bit to each frame to bring the total "1" bit count either to an even number or to an odd number (depending on the user's configuration), and the receiving computer checks for the same. "Mark" and "Space" are really of limited usefulness. In either of these two options, a parity bit is added, but the transmitting computer does not bother to calculate the evenness or oddness of the data bits, rather simply making the parity bit always equal to 1 ("mark") or 0 ("space") as chosen by the user. The receiving computer checks to see that the parity bit is always that value. These two options are of limited usefulness because the parity bit fails to reflect the status of the data being transmitted. The only corruption the receiving computer can detect, therefore, is a corruption of the parity bit itself!

One will often find the communications parameters of a serial network such as this displayed in "shorthand" notation as seen at the top of the "minicom" terminal display: `38400 8N1`. In this case, the terminal program is configured for a bit rate of 38400 bits per second, with a data field 8 bits long, no parity bit, and 1 stop bit. A computer configured for a bit rate of 9600 bps, with a 7-bit data field, odd parity, and 2 stop bits would be represented as `9600 7O2`.

Parity bits are not the only way to detect error, though. Some communication standards employ more sophisticated means. In the Ethernet (IEEE 802.3) standard, for example, each data frame is concluded with a *frame check sequence*, which is a collection of bits mathematically calculated by the transmitting device based on the content of the data. The algorithm is called a *cyclic redundancy*

check, or *CRC*, and is similar to the concept of "checksum" used by computers to check the integrity of data stored in hard disks and other "permanent" media. Like a parity algorithm, the CRC algorithm runs through a mathematical process whereby all the bits in the data field are counted, and a number is generated to reflect the statuses of those bits. The receiving computer takes the received data field and performs the exact same mathematical algorithm, generating its own CRC value. If any of the data's bits become corrupted during transmission, the two CRC values will not match, and the receiving computer will know *something* has gone wrong.

Like parity, the CRC algorithm is not perfect. There exists a chance that just the right combination of errors may occur in transmission causing the CRC values at both ends to match even though the data is not identical, but this is highly unlikely (calculated to be one chance in 10^{14}). It is certainly better than having no error detection ability at all.

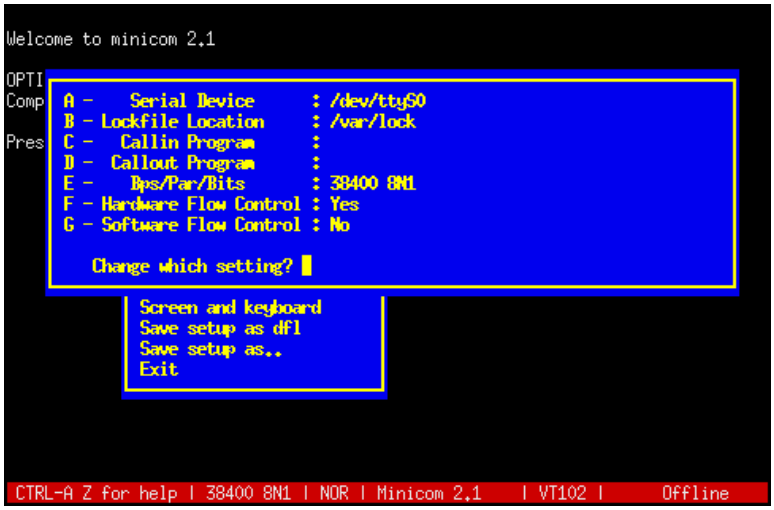
If the communications software in the receiving computer is configured to take action on a detection of error, it may return a "request for re-transmission" to the transmitting computer, so the corrupted message may be re-sent. This is analogous to a human being hearing a garbled transmission in a telephone conversation, and subsequently requesting the other person repeat what they just said.

Another option often found in serial data communications settings is something called *flow control*, not to be confused with the actual control of fluid through a pipe. In the context of digital communications, "flow control" refers to the ability of a receiving device to request a reduction in speed or even a complete cessation of data transmission in case of congestion. An example common to early personal computers was that of a mechanical printer receiving print data from a computer. While the computer may be able to transmit data to be printed at a very rapid pace, the printer is limited by the speed of its mechanism. In order to make the printing process go more smoothly, printers are equipped with *buffer memory* to store portions of the print job received from the transmitting computer that have not had time to print yet. However, these buffers were of finite size, and could become overwhelmed on large print jobs. So, if and when a printer detected its buffer near full capacity, it could issue a command to the computer to freeze data transmission until the printer's buffer has had some time to empty.

Flow control in serial networks may take place in either *hardware* mode or *software* mode. "Hardware" mode refers to the existence of additional connector pins and cable conductors specifically designated for such "halt" signals. "Software" mode refers to data codes communicated over the regular network channel telling the transmitting device to halt and resume. Software flow control is sometimes referred to as XON/XOFF in honor of the names given to these codes[16].

[16] The "XOFF" code told the transmitting device to halt its serial data stream to give the receiving device a chance to "catch up." In data terminal applications, the XOFF command could be issued by a user by pressing the key combination <Ctrl><S>. This would "freeze" the stream of text data sent to the terminal by the host computer. The key combination <Ctrl><Q> sent the "XON" code, enabling the host computer to resume data transmission to the terminal.

The following screen shot shows options for flow control in the "minicom" terminal program:



Here, you can see "hardware" flow control enabled and "software" flow control disabled.

15.2.5 Channel arbitration

When two or more communication devices exchange data, the directions of their communication may be classified into one of two categories: *simplex* or *duplex*. A "simplex" network is one-way communication only. A sensor outputting digital data to a remotely-located indicator over a digital network would be an example of simplex communication, where the flow of information goes from sensor to indicator, and never the other direction. A public-address (PA) system is an analog example of a simplex communication system, since audio information only goes in one direction (from the person with the microphone to the audience).

"Duplex" communication refers to two-way data exchange. Voice telephony is an analog example of two-way (duplex) communication, where either person at the end of the connection can hear the other person talking. Duplex communication may be further subdivided into *half-duplex* and *full-duplex*, referring to whether or not the two-way communication may be simultaneous. In a "full-duplex" system, both devices may transmit data to each other simultaneously because they have separate channels (separate wires, or optical fibers, or radio frequencies) for their respective transmissions. In a "half-duplex" system, only one device may transmit at any time because the devices must share a common channel. A telephone system is an example of a full-duplex system, although it may be rather difficult for the people to understand each other when they are speaking over one another. A push-to-talk radio system ("walkie-talkie") is an example of a half-duplex system, where each person must take turns talking.

Most industrial data networks are half-duplex, if only for the reason that most networks consist of more than two devices on a network segment. When more than two devices share a network, there are not enough data channels to allow *all* of the devices to simultaneously transmit and listen to each other. Thus, virtually any network supporting more than two devices will be half-duplex at best, and may even be limited to simplex operation in some cases.

In half-duplex systems, there must be some way for the respective devices to "know" when they are allowed to transmit. If multiple devices sharing one communications channel attempt to transmit simultaneously, their messages will "collide" in such a way that no device on the network will be able to interpret either message. The problem is analogous to two people simultaneously pressing the "talk" buttons on their two-way radio units: neither of the talking people can hear each other, and anyone else on the same channel hears the garbled amalgam of those two peoples' superimposed transmissions. In order to avoid this scenario in a half-duplex network, there must be some strategy to coordinate transmissions so only one device may "talk" at any given time. The problem of deciding "who" gets to "talk" at any given time is generally known as *channel arbitration*. Several strategies for addressing this problem have been developed in the data communications field, a few of which will be described in this subsection.

Master-slave

Our first method works on the principle of having only one device on the network (the "master") with permission to arbitrarily transmit data. All other devices on the network are "slaves," which may only respond in direct answer to a query from the master. If the network happens to be simplex in nature, slave devices don't even have the ability to transmit data – all they can do is "listen" and receive data from the one master device.

For example, in a half-duplex master-slave network, if one slave device has data that needs to be sent to another slave device, the first slave device must wait until it is prompted ("polled") by the master device before it is allowed to transmit that data to the network. Once the data is transmitted, any and all slave devices may receive that transmission, since they all "listen" to the same communications channel.

An example of an industrial network using master-slave channel arbitration is HART *multidrop*, where multiple HART field instruments are parallel-connected on the same wire pair, and one device (usually a dedicated computer) serves as the master node, polling the field instruments one at a time for their data.

Another example of a master-slave industrial network is a *Modbus* network connecting a programmable logic controller (PLC) to multiple variable-frequency motor drives (VFDs). The master device (the PLC) initiates all communications, with the slave devices (the motor drives) at most replying to the PLC master (and in many cases not replying at all, but merely receiving data from the PLC in simplex mode).

Master-slave arbitration is simple and efficient, but suffers from one glaring weakness: if the master device happens to fail, all communication on the network halts. This means the ability of *any* device on the network to transmit information utterly depends on the proper function of *one* device, representing a high level of dependence on that one (master) device's function.

Some master-slave networks address this problem by pre-assigning special "back-up" status to one or more slave devices. In the event that the master device fails and stops transmitting for a certain amount of time, the back-up device becomes "deputized" to act as the new master, taking over the role of the old master device by ensuring all slave devices are polled on schedule.

Token-passing

Another method of arbitrating which device gets to transmit on a channel in a half-duplex network is the *token-passing* method. Here, a special data message called the "token" serves as temporary authorization for each device to transmit. Any device in possession of the token is allowed to act as a master device, transmitting at will. After a certain amount of time, that device must relinquish the token by transmitting the token message on the network, complete with the address of the next device. When that other device receives the token message, it switches into master mode and transmits at will. The strategy is not unlike a group of people situated at a table, where only one of them at a time holds some object universally agreed to grant speaking authority to the holder.

Token-passing ensures only one device is allowed to transmit at any given time, and it also solves the problem inherent to master-slave networks of what happens when the master device fails. If one of the devices on a token-passing network fails, its silence will be detected after the last token-holding device transmits the token message to the failed device. After some pre-arranged period of time, the last token-holding device may re-transmit the token message to the *next* device after the one that failed, re-establishing the pattern of token sharing and ensuring all devices get to "speak" their turn once more.

Examples of token-passing networks include the general-purpose Token Ring network standard (IEEE 802.5) and the defunct Token Bus (IEEE 802.4). Some proprietary industrial networks such as Honeywell's TDC 3000 network (called the *Local Control Network*, or *LCN*) utilized token-passing to arbitrate access to the network.

Token-passing networks require a substantially greater amount of "intelligence" built into each network device than master-slave requires. The benefits, though, are greater reliability and a high level of bandwidth utilization. That being said, token-passing networks may suffer unique disadvantages of their own. For example, there is the question of what to do if such a network becomes severed, so that the one network is now divided into two segments. At the time of the break, only one device will possess the token, which means only one of the segments will possess any token at all. If this state of affairs holds for some time, it will mean the devices lucky enough to be in the segment that still has the token will continue communicating with each other, passing the token to one another over time as if nothing was wrong. The isolated segment, however, lacking any token at all, will remain silent even though all its devices are still in working order and the network cable connecting them together is still functional. In a case like this, the token-passing concept fares no better than a master-slave network. However, what if the designers of the token-passing network decide to program the devices to automatically generate a new token in the event of prolonged network silence, anticipating such a failure? If the network becomes severed and broken into multiple segments, the isolated segments will now generate their own tokens and resume communication between their respective devices, which is certainly better than complete silence as before. The problem now is, what happens if a technician locates the break in the network cable and re-connects it? Now, there will be *multiple* tokens on one network, and confusion will reign!

Another example of a potential token-passing weakness is to consider what would happen to such a network if the device in possession of the token failed before it had an opportunity to relinquish the token to another device. Now, the entire network will be silent, because no device possesses the token! Of course, the network designers could anticipate such a scenario and pre-program the devices to generate a new token after some amount of silence is detected, but then this raises the possibility of the previously-mentioned problem when a network becomes severed and multiple tokens arise in

an effort to maintain communication in those isolated network segments, then at some later time the network is re-connected and now multiple tokens create data collision problems.

CSMA

A completely different method of channel arbitration is where any and all devices have the ability to initiate communication if the network is silent. This is generally called *CSMA*, or "Carrier Sense Multiple Access." There are no dedicated master and slave devices with CSMA, nor do devices equally share "mastership" on a scheduled cycle as with token-passing. *Any* device on a CSMA network may "talk" whenever the network is free. This is analogous to an informal conversation between multiple people, where anyone is free to speak in a moment of silence.

Of course, such an egalitarian form of channel arbitration invites instances where two or more devices begin communicating simultaneously. This is called a *collision*, and must be addressed in some manner in order for any CSMA network to be practical.

Multiple methods exist to overcome this problem. Perhaps the most popular in terms of number of installed networks is *CSMA/CD* ("Carrier Sense Multiple Access with Collision Detection"), the strategy used in Ethernet. With CSMA/CD, all devices are not only able to sense an idle channel, but are also able to sense when they have "collided" with another transmitting device. In the event of a collision, the colliding devices both cease transmission, and set random time-delays to wait before re-transmission. The individual time delays are randomized to decrease the probability that a re-collision between the same devices will occur after the wait. This strategy is analogous to several peers in one group holding a conversation, where all people involved are equally free to begin speaking, and equally deferential to their peers if ever two or more accidently begin speaking at the same time. Occasional collisions are normal in a CSMA/CD network, and should not be taken as an indication of trouble unless their frequency becomes severe.

A different method of addressing collisions is to pre-assign to each device on the network a priority number, which determines the order of re-transmission following a collision. This is called *CSMA/BA*, or "Carrier Sense Multiple Access with Bitwise Arbitration," and it is analogous to several people of different social levels in one group holding a conversation. All are free to speak when the room is silent, but if two or more people accidently begin speaking at the same time, the person of highest "rank" is allowed to continue while the "lower-rank" person(s) must wait. This is the strategy used in DeviceNet, an industrial network based on CAN technology, one of the more popular data networks used in automotive engine control systems.

Some CSMA networks lack the luxury of collision detection, and must therefore strive to prevent collisions rather than gracefully recover from them. Wireless digital networks are an example where collision detection is not an option, since a wireless (radio) device having a single antenna and a single channel cannot "hear" any other devices' transmissions while it is transmitting, and therefore cannot detect a collision if one were to occur. A way to avoid collisions for such devices is to pre-assign each device on the network with a priority number, which determines how long each device is forced to wait after detecting a "quiet" network before it is allowed to transmit a new message. So long as no two devices on the network have the same "wait" time, there will be no collisions. This strategy is called *CSMA/CA*, or "Carrier Sense Multiple Access with Collision Avoidance," and is the technique used for WLAN networks (the IEEE 802.11 specification). A consequence of collision avoidance, though, is unequal access to the network. Those devices with higher-priority (shorter wait times) will always have an advantage in transmitting their data over devices of lower priority. The degree of disparity in network access grows as more devices occupy the network. CSMA/CA is analogous to a group of shy people talking, each person afraid to speak at the same time as another, and so each person waits a different amount of time following the conclusion of the last utterance before daring to speak. This sort of ultra-polite behavior may ensure no one accidently interrupts

another, but it also means the shiest person will hardly ever get a chance to speak.

A potential problem in any digital network, but particularly networks employing CSMA arbitration, is something known as *jabbering*. If a network device happens to fail in such a way that it ceaselessly transmits a signal on the network, none of the other CSMA devices will ever be allowed to transmit because they continuously detect a "carrier" signal from the jabbering device[17]. Some Ethernet components sport *jabber latch* protection circuits designed to detect jabber and automatically cut off the offending device from the network.

15.2.6 Code sets

After engineering methods to encode bits of data as electrical (or optical or radio) signals, standardizing the speed of those bits' broadcast, framing bits as groups complete with start and stop signals, and providing for multiple devices to share a common communications channel, there still remains the issue of how to make "1" and "0" symbols represent something other than Boolean values (on/off, true/false, mark/space, etc.). This is where *codes* become useful.

Morse and Baudot codes

In the early days of communication, Morse code was used to represent letters of the alphabet, numerals (0 through 9), and some other characters in the form of "dot" and "dash" signals. In the International Morse Code, no character requires more than six bits of data, and some (such as the common letters E and T) require only one bit.

The variable bit-length of Morse code, though very efficient[18] in terms of the total number of "dots" and "dashes" required to communicate textual messages, was difficult to automate in the form of teletype machines. In answer to this technological problem, Emile Baudot invented a different code where each and every character was five bits in length. Although this gave only 32 characters, which is not enough to represent the 26-letter English alphabet, plus all ten numerals and punctuation symbols, Baudot successfully addressed this problem by designating two of the characters as "shift" characters: one called "letters" and the other called "figures." The other 30 characters had dual (overloaded) meanings, depending on the last "shift" character issued in the serial data stream[19].

[17] I once encountered this very type of failure on the job, where a copper-to-fiber adapter on a personal computer's Ethernet port jammed the entire network by constantly spewing a meaningless stream of data. Fortunately, indicator lights on all the channels of the communications equipment clearly showed where the offending device was on the network, allowing us to take it out of service for replacement.

[18] Morse code is an example of a *self-compressing* code, already optimized in terms of minimum bit count. Fixed-field codes such as Baudot and the more modern ASCII tend to waste bandwidth, and may be "compressed" by removing redundant bits.

[19] For example, the Baudot code 11101 meant either "Q" or "1" depending on whether the last shift character was "letters" or "figures," respectively. The code 01010 meant either "R" or "4". The code 00001 meant either "T" or a "5".

EBCDIC and ASCII

A much more modern attempt at encoding characters useful for text representation was *EBCDIC*, the "Extended Binary Coded Decimal Interchange Code" invented by IBM in 1962 for use with their line of large ("mainframe") computers. In EBCDIC, each character was represented by a one-byte (eight bit) code, giving this code set 256 (2^8) unique characters. Not only did this provide enough unique characters to represent all the letters of the English alphabet (lower-case *and* capital letters separately!) and numerals 0 through 9, but it also provided a rich set of *control characters* such as "null," "delete," "carriage return," "linefeed," and others useful for controlling the action of electronic printers and other machines.

A number of EBCDIC codes were unused (or seldom used), though, which made it somewhat inefficient for large data transfers. An attempt to improve this state of affairs was *ASCII*, the "American Standard Code for Information Interchange" first developed in 1963 and then later revised in 1967, both by the American National Standards Institute (ANSI). ASCII is a seven-bit code, one bit shorter per character than EBCDIC, having only 128 unique combinations as opposed to EBCDIC's 256 unique combinations. The compromise made with ASCII versus EBCDIC was a smaller set of control characters.

IBM later created their own "extended" version of ASCII, which was eight bits per character. In this extended code set were included some non-English characters plus special graphic characters, many of which may be placed adjacently on a paper printout or on a computer console display to form larger graphic objects such as lines and boxes.

ASCII is wildly popular, even today. Nearly every digital transmission of English text in existence employs ASCII as the character encoding[20]. Nearly every text-based computer program's source code is also stored on media using ASCII encoding, where 7-bit codes represent alphanumeric characters comprising the program instructions.

[20] Including the source code for this textbook!

The basic seven-bit ASCII code is shown in this table, with the three most significant bits in different columns and the four least significant bits in different rows. For example, the ASCII representation of the upper-case letter "F" is 1000110, the ASCII representation of the equal sign (=) is 0111101, and the ASCII representation of the lower-case letter "q" is 1110001.

ASCII code set

↓ LSB / MSB →	000	001	010	011	100	101	110	111
0000	NUL	DLE	SP	0	@	P	‘	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	”	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EOT	DC4	$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	’	7	G	W	g	w
1000	BS	CAN	(	8	H	X	h	x
1001	HT	EM	)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[	k	{
1100	FF	FS	,	<	L	\	l	|
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

Unicode

There exist *many* written languages whose characters cannot and are not represented by either EBCDIC or ASCII. In in attempt to remedy this state of affairs, a new standardized code set is being developed called *Unicode*, with sixteen bits per character. This large bit field gives 65,536 possible combinations, which should be enough to represent every unique character in every written language in the entire world. In deference to existing standards, Unicode encapsulates both ASCII and EBCDIC as sub-sets within its defined character set[21].

And no, I am not going to include a table showing all the Unicode characters!

[21] To illustrate, the first 128 Unicode characters (0000 through 007F hexadecimal) are identical to ASCII's 128 characters (00 through 7F hexadecimal)

15.2.7 The OSI Reference Model

Digital data communication may be described in many ways. For example, a connection formed between two computers to exchange a text document is a multi-layered activity, involving many steps to convert human language into electrical impulses for transmission, then re-convert those electrical impulses into human language again at the receiving end. Not surprisingly, there usually exist many different ways to perform this same task: different types of networks, different encodings, different communications and presentation software, etc.

To illustrate by analogy, think of all the actions and components necessary to transport items using an automobile. In order to move furniture from an apartment to a house, for example, you would require the following:

- An appropriate vehicle
- Addresses or directions for both locations
- A driver's license and knowledge of driving rules
- Fuel for the vehicle
- Knowledge of how to safely stack furniture for transport
- Knowledge of how the furniture is to be placed in the house

These details may seem trivial to mention, as human beings familiar with the common task of moving personal belongings from one location to another, but imagine having to describe every single action and component to someone from another planet ignorant of vehicles, addresses, maps, driver's licenses, fuel, etc. One way to help describe all this complexity would be to assign different people to different layers of detail. For example, an automotive engineer could discuss the details of how engines burn fuel to do mechanical work (propelling the vehicle) while a furniture loader could describe how furniture is to be loaded and taken off the vehicle. A driving instructor could then explain all the procedures of safely driving the vehicle, while a city planner could explain the organization of streets and addresses in relation to a city map. Finally, an interior decorator could wax eloquent on the proper placement of furniture in the house. Each person would be describing a different aspect of the furniture move, each one of those aspects being important to the overall goal of moving furniture from one location to another.

Moreover, for each one of the aspects described by a specialist, there may exist several different alternatives. For example, there are many different models and designs of vehicle one might use for the job, and there may be different driving rules depending on where the two locations are for the move. Addresses and directions will *certainly* vary from city to city, and even within one city there will be alternative routes between the two locations. Finally, there is virtually no end to arrangements for furniture at the destination house, each one with its own functional and aesthetic merits.

By the same token, we may divide the act of digitally communicating data into several distinct aspects, from the physical representation of 0 and 1 bits as electrical/optical/radio signals to the final presentation of data in a form meaningful to human beings. Each of those aspects is important to the overall goal of digital data communication, and there may very well be many alternative methods (standards) for each aspect. We may represent 0 and 1 bits using NRZ (Non-Return to

Zero) encoding, Manchester encoding, FSK modulation, etc.; the signals may be electrical or they may be optical or they may even be radio waves; the options for electrical cables and connector types are many. Bits may be framed differently as they are packaged for transmission, and arbitration between devices on the network handled in a variety of different ways. How we address multiple devices on a network so messages get routed to their proper destinations is important as well.

A scheme originally intended as a formal standard, but now widely regarded as a general model to describe the portions of other standards, helps us clarify the complexity of digital communications by dividing communication functions into seven distinct "layers." Developed by the *ISO* (International Organization for Standards)[22] in 1983, the *OSI Reference Model* divides communication functions into the following categories, shown in this table with examples:

Layer 7 **Application**	This is where digital data takes on practical meaning in the context of some human or overall system function. *Examples: HTTP, FTP, Telnet, SSH*
Layer 6 **Presentation**	This is where data gets converted between different formats. *Examples: ASCII, EBCDIC, MPEG, JPG, MP3*
Layer 5 **Session**	This is where "conversations" between digital devices are opened, closed, and otherwise managed for reliable data flow. *Examples: Sockets, NetBIOS*
Layer 4 **Transport**	This is where complete data transfer is handled, ensuring all data gets put together and error-checked before use. *Examples: TCP, UDP*
Layer 3 **Network**	This is where the system determines network-wide addresses, ensuring a means for data to get from one node to another. *Examples: IP, ARP*
Layer 2 **Data link**	This is where basic data transfer methods and sequences (frames) are defined within the smallest segment(s) of a network. *Examples: CSMA/CD, Token passing, Master/Slave*
Layer 1 **Physical**	This is where data bits are equated to electrical, optical, or other signals. Other physical details such as cable and connector types are also specified here. *Examples: EIA/TIA-232, 422, 485, Bell 202*

The vast majority of digital networking standards in existence address mere portions of the 7-layer model. Any one of the various Ethernet standards, for example, applies to layers 1 and 2, but none of the higher-level layers. In other words, Ethernet is a means of encoding digital information in electronic form and packaging that data in a standard format understandable to other Ethernet devices, but it provides no functionality beyond that. Common industrial network standards such as EIA/TIA-232 and EIA/TIA-485 don't even go that far, being limited mostly to layer 1 concerns (signal voltage levels, wiring, and in some cases types of electrical connectors). By contrast, other industrial networking standards specify nothing about lower-level layers, but focus instead on high-level concerns. Modbus, for example, is concerned only with layer 7, and not with

[22] If you are thinking the acronym should be "IOS" instead of "ISO," you are thinking in terms of English. "ISO" is a non-English acronym!

any of the lower-level layers[23]. This means if two or more industrial devices on a network (such as programmable logic controllers, or PLCs) use "Modbus" to communicate with each other, it refers only to the high-level programming codes designed to poll and interpret data within those devices. The actual cable connections, electrical signals, and communication techniques used in that "Modbus" network may vary widely. Anything from EIA/TIA-232 to Ethernet to a wireless network such as WLAN may be used to actually communicate the high-level Modbus instructions between PLCs.

The following sections explore some common networking standards used for industrial instrumentation systems. The OSI Reference Model will be mentioned when and where appropriate.

15.3 EIA/TIA-232, 422, and 485 networks

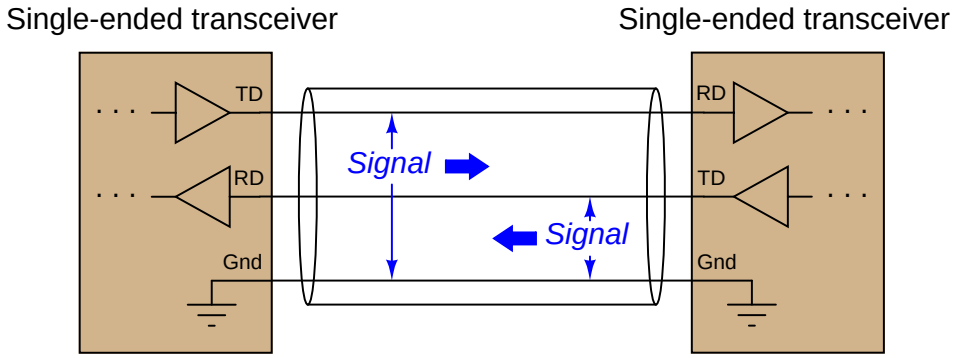
Some of the simplest types of digital communication networks found in industry are defined by the EIA (Electronic Industry Alliance) and TIA (Telecommunications Industry Alliance) groups, under the numerical labels 232, 422, and 485. This section discusses these three network types.

[23] It should be noted here that some network standards incorporating the name "Modbus" actually do specify lower-level concerns. *Modbus Plus* is a layer 2 standard, for example.

15.3.1 EIA/TIA-232

The EIA/TIA-232C standard, formerly[24] known as *RS-232*, is a standard defining details found at layer 1 of the OSI Reference Model (voltage signaling, connector types) and some details found at layer 2 of the OSI model (asynchronous transfer, "handshaking" signals between transmitting and receiving devices). In the early days of personal computers, almost every PC had either a 9-pin or a 25-pin connector (and sometimes multiple of each!) dedicated to this form of digital communication. For a while, it was *the* way peripheral devices such as keyboards, printers, modems, and mice were connected to the PC. USB (Universal Serial Bus) has now all but replaced EIA/TIA-232 for personal computers, but it still lives on in the world of industrial devices.

EIA/TIA-232 networks are point-to-point, intended to connect only two devices[25]. The signaling is *single-ended* (also known as *unbalanced*), which means the respective voltage pulses are referenced to a common "ground" conductor, a single conductor used to transfer data in each direction:

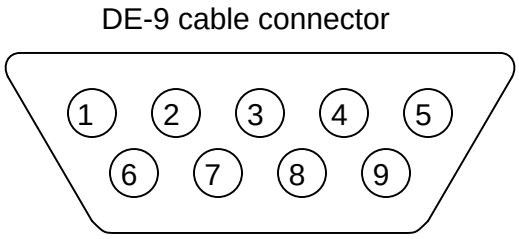


EIA/TIA-232 specifies positive and negative voltages (with respect to the common ground conductor) for its NRZ signaling: any signal more negative than -3 volts detected at the receiver is considered a "mark" (1) and any signal more positive than +3 volts detected at the receiver is considered a "space" (0). EIA/TIA-232 transmitters are supposed to generate -5 and +5 volt signals (minimum amplitude) to ensure at least 2 volts of noise margin between transmitter and receiver.

[24] The designation of "RS-232" has been used for so many years that it still persists in modern writing and manufacturers' documentation, despite the official status of the EIA/TIA label. The same is true for EIA/TIA-422 and EIA/TIA-485, which were formerly known as RS-422 and RS-485, respectively.

[25] "Daisy-chain" networks formed of more than two devices communicating via EIA/TIA-232 signals have been built, but they are rarely encountered, especially in industrial control applications.

Cable connectors are also specified in the EIA/TIA-232 standard, the most common being the DE-9[26] (nine-pin) connector. The "pinout" of a DE-9 connector for any *DTE* (Data Terminal Equipment) device at the end of an EIA/TIA-232 cable is shown here:

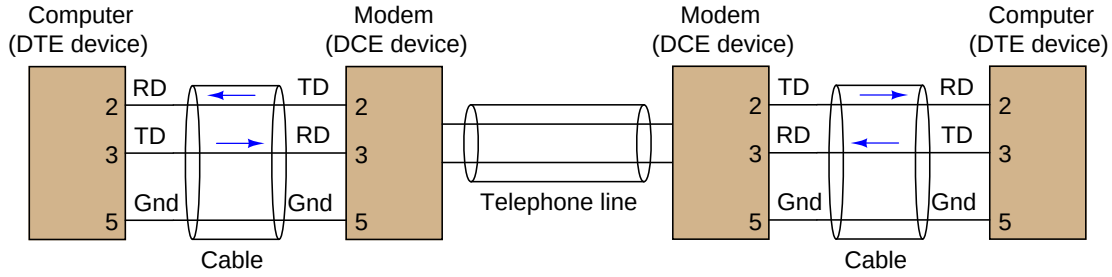


Pin number	Assignment	Abbreviation
1	Carrier Detect	CD
2	**Received Data**	**RD**
3	**Transmitted Data**	**TD**
4	Data Terminal Ready	DTR
5	**Signal Ground**	**Gnd**
6	Data Set Ready	DSR
7	Request To Send	RTS
8	Clear To Send	CTS
9	Ring Indicator	RI

Those terminals highlighted in **bold** font represent those connections absolutely essential for any EIA/TIA-232 link to function. The other terminals carry optional "handshaking" signals specified for the purpose of coordinating data transactions (these are the layer 2 details).

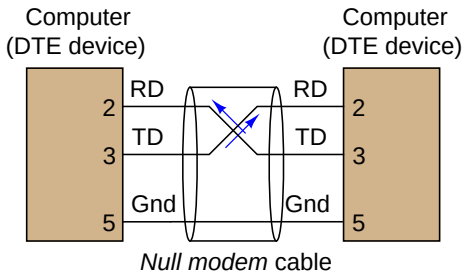
[26] Often (incorrectly) called a "DB-9" connector.

For *DCE* (Data Communications Equipment[27]) devices such as modems, which extend the EIA/TIA-232 signal path onward to other devices, the assignments of pins 2 and 3 are swapped: pin 2 is the Transmitted Data (TD) pin while pin 3 is the Received Data (RD) for a DCE device. This allows straight pin-to-pin connections between the DTE and DCE devices, so the transmit pin of the DTE device connects to the receive pin of the DCE, and visa-versa.



[27] Also known by the unwieldy acronym *DCTE* (Data Circuit Terminating Equipment). Just think of "DTE" devices as being at the very end of the line, whereas "DCE" devices are somewhere in the middle, helping to exchange serial data between DTE devices.

If one desires to directly connect two DTE devices together using EIA/TIA-232, a special cable called a *null modem* must be used, which swaps the connections between pins 2 and 3 of each device. A "null modem" connection is necessary for the transmit pin of each DTE device to connect to the receive pin of the other DTE device:



The concept of a "null modem" is not unique to EIA/TIA-232 circuits[28]. Any communications standard where the devices have separate "transmit" and "receive" channels will require a "null modem" connection with transmit and receive channels swapped to be able to communicate directly without the benefit of interconnecting DCE devices. Four-wire EIA/TIA-485 and Ethernet over twisted-pair wiring are two examples of network standards where a "null" style cable is required for two DTE devices to directly connect.

EIA/TIA-232 networks may be simple, but they tend to be rather limited both in data bit rate and distance, those two parameters being inversely related. References to the EIA/TIA-232 standard repeatedly cite a maximum data rate of 19.2 kbps at 50 feet cable rate. Experimental tests[29] suggest greater rate/distance combinations may be possible in optimum conditions (low cable capacitance, minimum noise, good grounding). Since this communications standard was developed to connect peripheral devices to computers (typically within the physical span of one room), and at modest speeds, neither of these limitations were significant to its intended application.

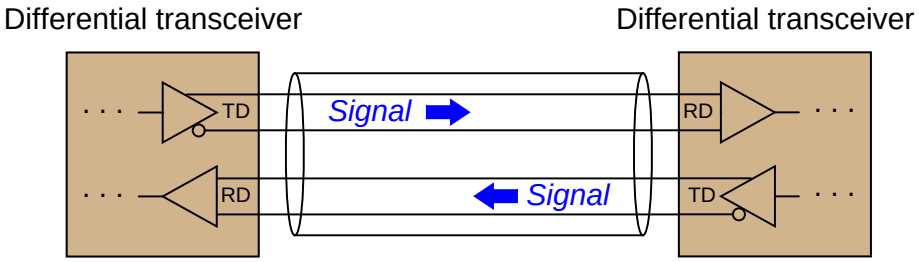
[28] In fact, the concept is not unique to digital systems at all. Try talking to someone using a telephone handset held upside-down, with the speaker near your mouth and the microphone hear your ear, and you will immediately understand the necessity of having "transmit" and "receive" channels swapped from one end of a network to the other!

[29] Once I experimented with the fastest data rate I could "push" an EIA/TIA-232 network to, using a "flat" (untwisted, unshielded pair) cable less than ten feet long, and it was 192 kbps with occasional data corruptions. Park, Mackay, and Wright, in their book *Practical Data Communications for Instrumentation and Control* document cable lengths as long as 20 meters at 115 kbps for EIA/TIA-232, and 50 *meters* (over 150 feet!) at 19.2 kbps: over three times better than the EIA/TIA-232 standard.

15.3.2 EIA/TIA-422 and EIA/TIA-485

The next two network standards[30] are less comprehensive than EIA/TIA-232, specifying only the electrical characteristics of signaling without any regard for connector types or any layer 2 (handshaking) considerations. Within these domains, the 422 and 485 standards differ significantly from 232, their designs intended to optimize both maximum cable length and maximum data rate.

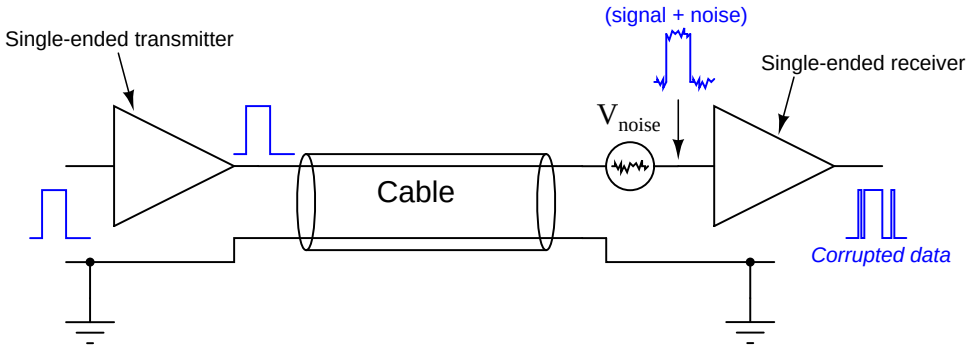
To begin with, the electrical signaling used for both EIA/TIA-422 and EIA/TIA-485 is *differential* rather than single-ended (*balanced* rather than unbalanced). This means a dedicated *pair* of wires is used for each communications channel rather than a single wire whose voltage is referenced to a common ground point as is the case with EIA/TIA-232:



Using dedicated wire pairs instead of single conductors sharing a common ground means that EIA/TIA-422 and EIA/TIA-485 networks enjoy much greater immunity to induced noise than EIA/TIA-232. Any electrical noise induced along the length of network cables tends to be fairly equal on all non-grounded conductors of that cable, but since the receivers in EIA/TIA-422 and EIA/TIA-485 networks response only to differential voltages (not common-mode voltages), induced noise is ignored.

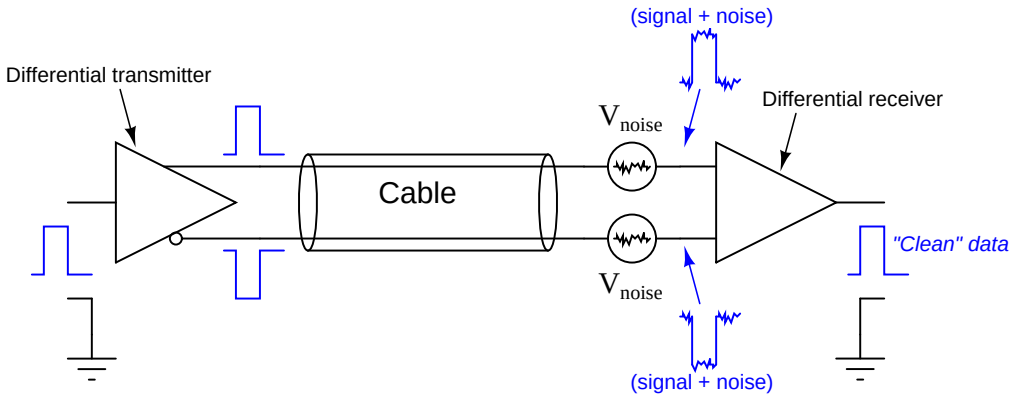
[30] Former labels for EIA/TIA-422 and EIA/TIA-485 were RS-422 and RS-485, respectively. These older labels persist even today, to the extent that some people will not recognize what you are referring to if you say "EIA/TIA-422" or "EIA/TIA-485."

The advantage differential signaling enjoys over single-ended signaling may be understood by graphical comparison. The first illustration shows how electrical noise imposed on the ungrounded conductor of a simplex communications cable becomes superimposed on the digital data signal, detected at the receiving end. Noise is modeled here as a voltage source in series along the ungrounded conductor, near the receiving end. In reality, it is more likely to be distributed along the bulk of the cable length:



If the superimposed noise voltage detected at the receiver has sufficient peak-to-peak amplitude to push the signal voltage above or below critical threshold levels, the receiver will interpret this as a change of digital state and cause corruptions in the data stream.

By contrast, any noise superimposed on ungrounded conductors in a differential signaling circuit cancel at the receiver, because the close proximity of those two conductors ensures any induced noise will be the same. Since the receiver responds only to *differential* voltage between its two inputs, this common-mode noise cancels, revealing a "clean" data signal at the end:



Both EIA/TIA-422 and EIA/TIA-485 systems use differential signaling, allowing them to operate over much longer cable lengths at much greater cable speeds than EIA/TIA-232 which is single-ended. Other high-speed network standards including Ethernet and USB (Universal Serial Bus) use differential signaling as well.

EIA/TIA-422 is a *simplex* (one-way) communications standard, whereas EIA/TIA-485 is a *duplex* (two-way) standard. Both support more than two devices on a network segment. With EIA/TIA-422, this means one transmitter and multiple receivers. With EIA/TIA-485, this may include multiple *transceivers* (devices capable of both transmitting and receiving at different times: half-duplex). Four wires are necessary to connect two such devices when full-duplex (simultaneous two-way communication) is required, and full-duplex is only practical between two devices (as shown in the previous illustration).

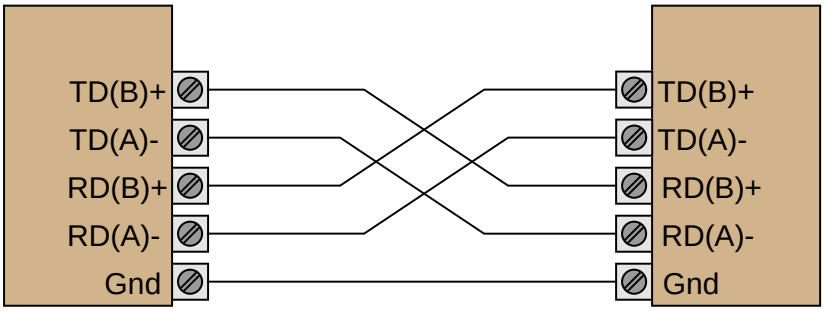
EIA/TIA-422 and EIA/TIA-485 specify positive and negative voltage differences (measured between each dedicated wire pair) for its signaling: any signal more negative than -200 millivolts is a "mark" (1) and any signal more positive than +200 millivolts is a "space" (0). These voltage thresholds are much lower than for EIA/TIA-232 ($\pm$ 3 volts) due to the noise-canceling properties of differential signaling. EIA/TIA-422 transmitters ("drivers") are supposed to generate -2 and +2 volt signals (minimum amplitude) to ensure at least 1.8 volts of noise margin between transmitter and receiver. EIA/TIA-485 drivers are allowed a smaller noise margin, with the minimum signal levels being -1.5 volts and +1.5 volts.

The maximum recommended cable length for both EIA/TIA-422 and EIA/TIA-485 networks is 1200 meters, which is greater than half a mile[31]. The maximum data rate is inversely dependent on cable length (just as it is for EIA/TIA-232), but substantially greater owing to the noise immunity of differential signaling. With the long cable lengths and higher data rates made possible by differential signaling, some applications may require *terminating resistors* to eliminate reflected signals. Experiments conducted by Texas Instruments demonstrate acceptable signal integrity at 200 kbps over a cable 100 feet long with no termination resistors. With a termination resistor at the receiver input (for simplex data transmission) in place on the same 100 foot cable, a data rate of 1 Mbps was achieved.

[31] 1200 meters is the figure commonly cited in technical literature. However, Park, Mackay, and Wright, in their book *Practical Data Communications for Instrumentation and Control* document EIA/TIA-422 and EIA/TIA-485 networks operating with cable lengths up to 5 km (over 16,000 feet!) at data rates of 1200 bps. Undoubtedly, such systems were installed with care, using high-quality cable and good wiring practices to minimize cable capacitance and noise.

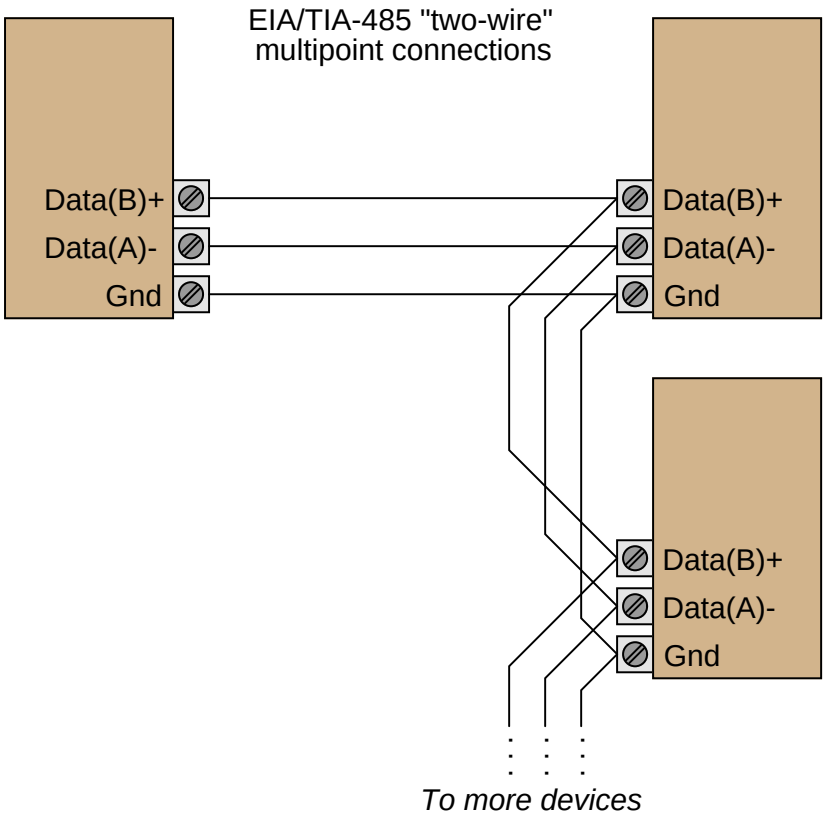
Due to the lack of standardization for cable connectors in EIA/TIA-422 and EIA/TIA-485 networks, there are no established pin numbers in certain connectors designated for the differential transmit and receive conductors. A common convention seen in industrial devices, though, are the labels "A" and "B", alternative labeled "-" and "+" or "A-" and "B+" in honor of their idle-state polarities (the "mark" or "1" state). In a 4-wire EIA/TIA-485 network, where full-duplex operation is possible, the terminals and connections will look something like this:

EIA/TIA-485 full-duplex connections



Note the use of a ground conductor connecting both devices together. Even though the data signaling is differential and therefore does not theoretically require a common ground connection (since common-mode voltage is ignored), a ground connection helps ensure the common-mode voltage does not become excessive, since *real* receiver circuits will not properly function when exposed to certain levels of common-mode voltage.

A popular connection scheme for EIA/TIA-485 half-duplex operation is where the Transmitted Data (TD) and Received Data (RD) terminal pairs are combined, so that two-way communication may occur over one pair of wires. With such devices, it is customary to label the terminals simply as "Data" (A- and B+):

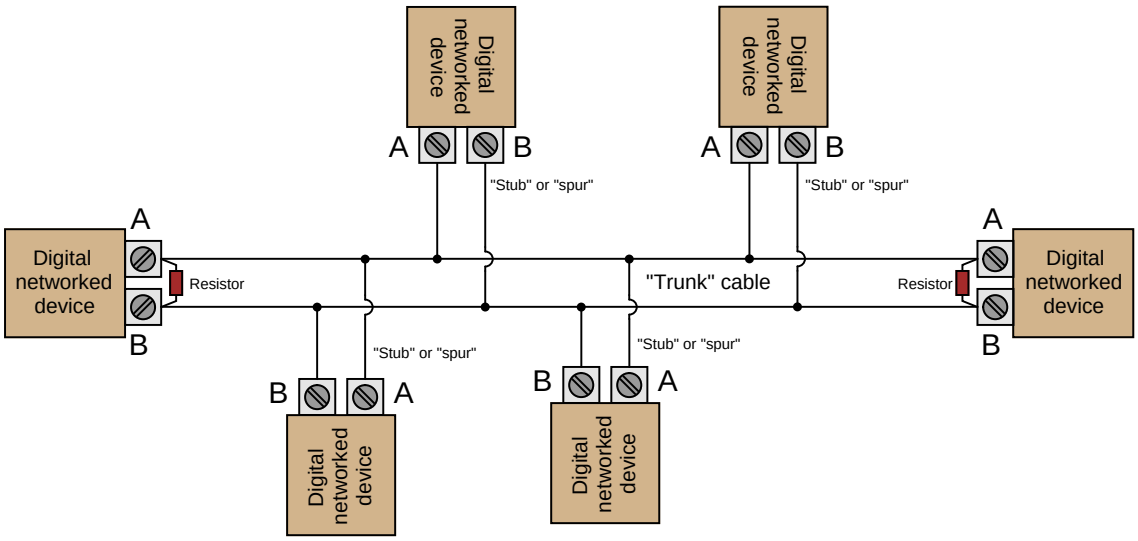


The possibility of half-duplex operation begs the question of channel arbitration and device addressing, but since the EIA/TIA-485 standard does not specify anything outside of layer 1 concerns, these matters are left to other networking standards to fulfill. In other words, EIA/TIA-485 is not a complete data communications standard, but merely serves as the layer 1 component of other standards such as Allen-Bradley's *Data Highway* (DH), Opto 22's *Optomux*, and others.

Given the potential for high-speed communication along lengthy runs of cable using EIA/TIA-422 or EIA/TIA-485, the potential necessity of terminating resistors to prevent signal "reflection" is very real. Networks operating with short cables, and/or slow data rates, may work just fine without termination resistors[32]. However, the effects of reflected signals grows more pronounced as the reflection time (time-of-flight for the signal to travel "round-trip" from one end of the cable to the other and back) approaches a substantial fraction of the bit time.

[32] In fact, a great many EIA/TIA-485 networks in industry operate "unterminated" with no problems at all.

No network should have more than two termination resistors, one at each (far) end, and care should be taken to limit the lengths of all cable "stubs" or "spurs" branching off of the main "trunk" cable:

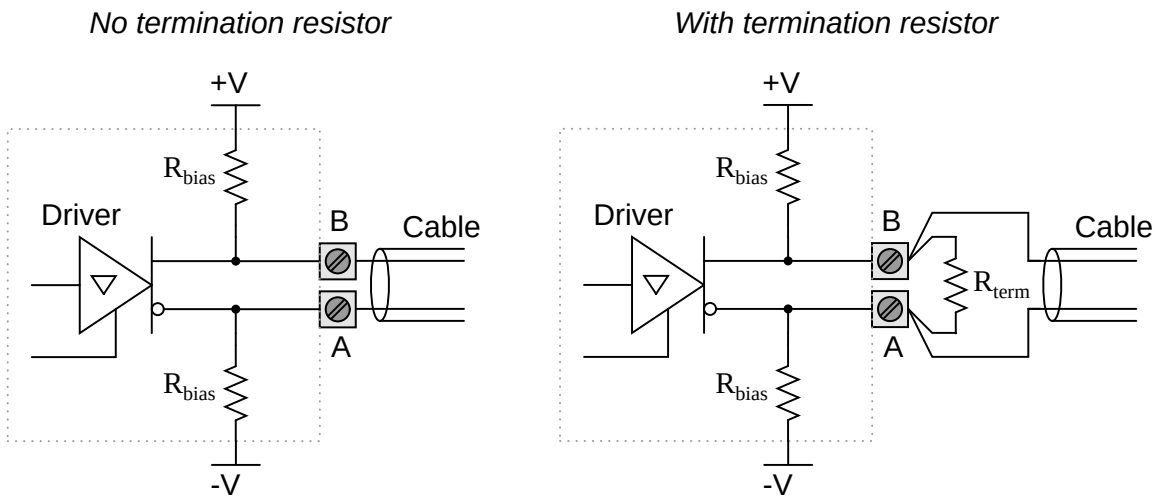


The proper value for these resistors, of course, is equality with the characteristic impedance[33] of the cable itself. A termination resistor value greater than the cable's surge impedance will still allow positive reflections of limited amplitude, while a termination resistor value less than the cable's surge impedance will still allow negative reflections of limited amplitude.

However, the inclusion of resistive loads to an EIA/TIA-422 or EIA/TIA-485 network may cause other problems. Many devices use a pair of *biasing resistors* internally to establish the "mark" state necessary for idle conditions, connecting the "A" terminal to the negative supply voltage rail through a resistor and the "B" terminal to the positive supply voltage rail through another resistor. Connecting a terminating resistor between terminals "A" and "B" will alter the voltage levels normally provided by these biasing resistors, consequently causing problems.

[33] For detailed explanation of how and why this is necessary, refer to section 5.5 beginning on page 255.

The following schematic diagram shows the equivalent circuit of an EIA/TIA-485 transceiver device, with and without a terminating resistor connected:



When the driver is in high-impedance (High-Z) mode, the "idle" state of the wire pair will be established by the bias resistors (equal to the supply voltage so long as there is no loading). However, a terminating resistor will act as a DC load to this biasing network, causing a substantial reduction of the "idle" state voltage toward 0 volts. Recall that -200 millivolts was the receiving threshold value for a "mark" state in both EIA/TIA-422 and EIA/TIA-485 standards (terminal "A" negative and terminal "B" positive). If the presence of a terminating resistor[34] reduces the idle state voltage to less than 200 millivolts absolute, the network's function may be compromised.

Thus, we see that the inclusion of any terminating resistors must be accompanied by an analysis of the devices' bias resistor networks if we are to ensure robust network operation. It is foolhardy to simply attach terminating resistors to an EIA/TIA-422 or EIA/TIA-485 network without considering their combined effect on biasing.

[34] Actually *two* terminating resistors in parallel, since one with be at each end of the cable! The actual DC biasing network will be more complicated as well if more than one device has its own set of internal bias resistors.

15.4 Ethernet networks

An engineer named Bob Metcalfe conceived the idea of Ethernet in 1973, while working for the Xerox research center in Palo Alto, California. His fundamental invention was the CSMA/CD method of channel arbitration, allowing multiple devices to share a common channel of communication while recovering gracefully from inevitable "collisions." In Metcalfe's vision, all of the "network intelligence" would be built directly into "controller" devices situated between the DTE devices (computers, terminals, printers, etc.) and a completely passive coaxial cable network. Unlike some other networks in operation at the time, Metcalfe's did not rely on additional devices to help coordinate communications between DTE devices. The coaxial cable linking DTE devices together would be completely passive and "dumb," performing no task but the conduction of broadcast signals between all devices. In that sense, it served the same purpose as the "luminiferous ether" once believed to fill empty space: conducting electromagnetic waves between separated points.

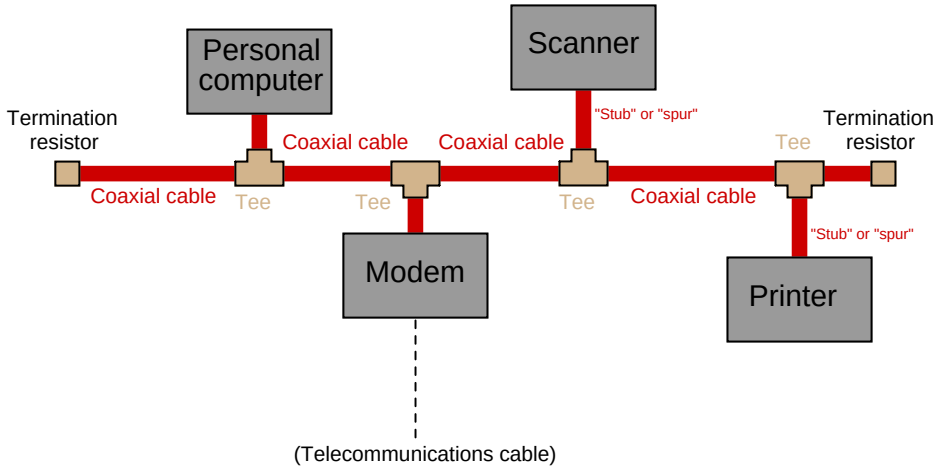
Metcalfe's original network design operated at a data rate of 2.94 Mbps, impressive for its time. By 1980, the three American computer companies DEC (Digital Equipment Corporation), Intel, and Xerox had collaborated to revise the Ethernet design to a speed of 10 Mbps, and released a standard called the *DIX Ethernet* standard (the acronym "DIX" representing the first letter of each company's name). Later, the IEEE Local and Metropolitan Networks Standards Committee codified the DIX Ethernet standard under the numeric label 802.3. At the present time there exist many "supplemental" standards underneath the basic 802.3 definition, a few of them listed here:

- 802.3a-1985 *10BASE2 "thin" Ethernet*
- 802.3d-1987 *FOIRL fiber-optic link*
- 802.3i-1990 *10BASE-T twisted-pair cable Ethernet*
- 802.3u-1995 *100BASE-T "Fast" Ethernet and Auto-Negotiation*
- 802.3x-1997 *Full-Duplex standard*
- 802.3ab-1999 *1000BASE-T "Gigabit" Ethernet over twisted-pair cable*

The IEEE 802.3 standard is limited to layers 1 and 2 of the OSI Reference Model: the "Physical" and "Data link" layers. In the physical layer (1), the various supplements describe all the different ways in which bits are electrically or optically represented, as well as permissible cable and connector types. In the data link layer (2), the IEEE standard describes how devices are addressed (each one with a unique identifier known as a *MAC address*, consisting of a 48-bit binary number usually divided into six bytes, each byte written as a two-character hexadecimal number), and also how data frames are organized for Ethernet transmissions.

15.4.1 Repeaters (hubs)

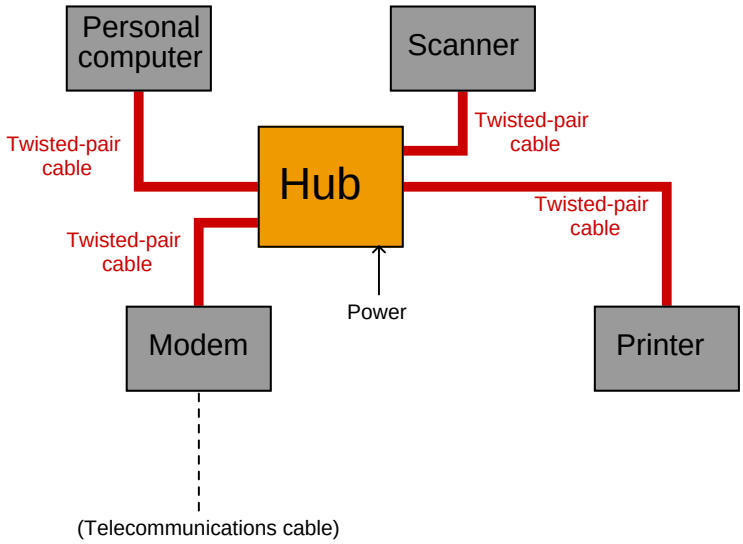
Bob Metcalfe's original design for Ethernet consisted of DTE devices connected to a common coaxial cable through the use of "tee" connectors, like this:



This cabling arrangement suffered several problems. First, it was inconvenient to run through an office building, since each DTE device needed to be coupled rather closely to the main "trunk." Short cable segments (called *stubs*, *spurs*, or *drops*) joining the main trunk line to each DTE device could not be too long, or else they would could cause multiple signal reflections to occur in the main line. Secondly, the signal strength decreased with each "tee" connector: every time the signal branched, it would lose power. Thirdly, the need for termination resistors at the far ends of the "ether" cable opened up the possibility that those terminators might fail, fall off, or be forgotten during installation or maintenance[35].

[35] These very same problems may arise in FOUNDATION Fieldbus networks, for the exact same reason: the cabling is passive (for increased reliability). This makes FOUNDATION Fieldbus instrument systems challenging to properly install for most applications (except in really simple cases where the cable route is straightforward), which in my mind is its single greatest weakness at the time of this writing (2009). I strongly suspect Ethernet's history will repeat itself in FOUNDATION Fieldbus at some later date: a system of reliable "hub" devices will be introduced so that these problems may be averted, and installations made much simpler.

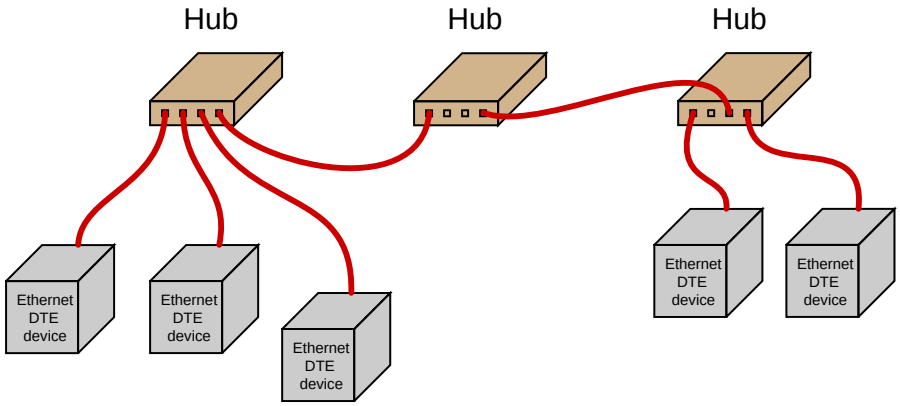
As Ethernet evolved as a practical networking standard, one of the many improvements added to its design was the concept of a *repeating hub*. A "repeater" is an active device designed to re-broadcast a signal, usually to overcome inevitable power losses incurred as that signal propagates along a cable. Repeaters are common in the telecommunications industry, where telephone, television, and computer signals must travel hundreds or thousands of miles between points of transmission and reception. A "repeating hub" is a repeater with multiple ports for many cables to plug into, where any signal entering on any cable is repeated to *all* ports on the device. Thus, a repeating hub (or simply "hub") allows multiple Ethernet devices to interconnect with no degradation in signal quality:



Not only do hubs improve system performance by boosting signals' voltage levels, but they also eliminate the need for termination resistors in the network. With a hub-based system, each and every cable terminates at either a DTE or DCE device, which is (now) designed with the proper termination resistance built-in to their internal transceiver circuitry. This means each and every Ethernet cable is automatically terminated by the proper impedance simply by plugging it in to the Ethernet port of *any* device. "Stub" or "spur" cables with their length restrictions are also a thing of the past, since no cable ever splits or branches in a hub-based network system.

Hubs are considered "layer 1" devices, because they operate purely on the physical layer of Ethernet: all they do is receive Ethernet signals and re-broadcast those signals in boosted form to all other devices plugged into the hub. As a piece of interconnecting hardware, a hub is considered a DCE (Data Communications Equipment), as opposed to the end-of-cable devices such as computers and printers which are DTEs (Data Terminal Equipment).

Repeating hubs may be connected together to form larger networks[36]:

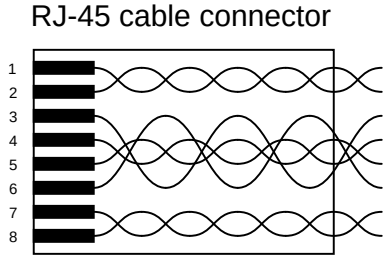


Since hubs are merely "layer 1" devices, mindlessly boosting and re-broadcasting signals received to their ports, their presence does not mitigate collisions between transmitting devices. As far as collisions between those devices is concerned, they might as well be directly connected together on a single piece of coaxial cable. One way to express this concept is to say that all portions of the network are part of the same *collision domain*. In other words, a collision happening in one portion of the network happens in *all* portions of the network.

[36]There are practical limits as to how many hubs may be "daisy-chained" together in this manner, just as there are practical limits to how long a twisted-pair cable may be (up to 100 meters). If too many hubs are cascaded, the inevitable time delays caused by the process of repeating those electrical impulses will cause problems in the network.

15.4.2 Ethernet cabling

Along with hubs came another form of Ethernet cable and connector: *unshielded, twisted pair* (UTP) wiring and *RJ-45* "flat" connectors. These cables use multiple twisted pairs of wires instead of the coaxial cable specified in Metcalfe's original Ethernet. The purpose of using twisted-wire pairs is to reduce magnetic signal coupling (for more information on this, refer to section 8.3.5 beginning on page 362).



For 10 Mbps Ethernet over UTP cable (called 10BASE-T) and for 100 Mbps Ethernet (called 100BASE-TX), only two[37] out of four available wire pairs are used:

Pin number	Assignment	Abbreviation
1	Transmit Data (+)	TD+
2	Transmit Data (-)	TD-
3	Receive Data (+)	RD+
4	(not used)	
5	(not used)	
6	Receive Data (-)	RD-
7	(not used)	
8	(not used)	

It should be noted that 1000 Mbps ("Gigabit") Ethernet over twisted-wire pairs does in fact use all four pairs in an eight-wire cable: a departure from traditional UTP Ethernet cable wiring.

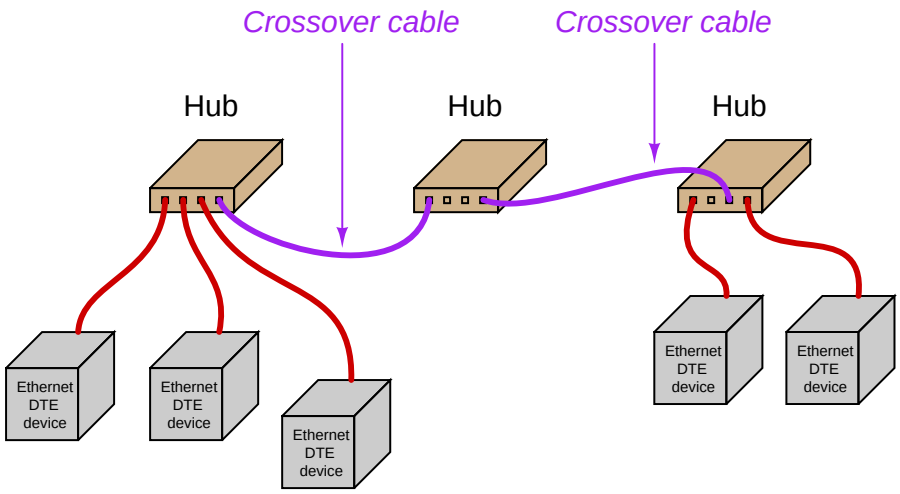
Pin number	Assignment	Abbreviation
1	Pair "A" (+)	BI_DA+
2	Pair "A" (-)	BI_DA-
3	Pair "B" (+)	BI_DB+
4	Pair "C" (+)	BI_DC+
5	Pair "C" (-)	BI_DC-
6	Pair "B" (-)	BI_DB-
7	Pair "D" (+)	BI_DD+
8	Pair "D" (-)	BI_DD-

[37] With only half the available wire pairs used in a standard 10 Mbps or 100 Mbps Ethernet cable, this opens the possibility of routing *two* Ethernet channels over a single four-pair UTP cable and RJ-45 connector. Although this is non-standard wiring, it may be a useful way to "squeeze" more use out of existing cables in certain applications. In fact, "splitter" devices are sold to allow two RJ-45-tipped cables to be plugged into a single RJ-45 socket such that one four-pair cable will then support two Ethernet pathways.

Along with UTP cables and RJ-45 connectors came a significant alteration to the basic electrical scheme of Ethernet. Metcalfe's original design used a simple coaxial cable as the "ether" connecting devices together. Such cables had only two conductors, meaning each device needed to transmit *and* receive data over the same two conductors. With UTP cable's four pairs of conductors, transmission and reception of signals is handled over different wire pairs. This means connections made between Ethernet devices must employ a "swap" between TD and RD wire pairs in order for communication to take place, so that the "receiver" circuitry of one device connects to the "transmitter" circuitry of the other, and visa-versa. This is precisely the same problem experienced inherent to EIA/TIA-232 and four-wire EIA/TIA-485 networks, where separate wire pairs for "transmit" and "receive" are different.

In a typical Ethernet system, the interconnecting hubs perform this transmit/receive swap. Hubs are considered DCE devices, while computers and other end-of-the-line devices are considered DTE devices. This means the pin assignments of DTE and DCE devices must be different in order to ensure the transmit/receive pin swap necessary for straight-through cables to work. This also means if someone ever wishes to directly connect two Ethernet DTE devices together without the benefit of a hub in between, a special *crossover* cable must be used for the connection, identical in function to the *null modem* cable used to connect two EIA/TIA-232 DTE devices together.

Furthermore, the same problem exists when multiple hubs are connected to form larger networks. Since each hub is a DCE device, a straight-through cable connecting two hubs together will pass transmitted signals from one hub directly to the "transmit" pins of the other hub, not the "receive" pins as it needs to. Consequently, a "crossover" cable should be used to connect two Ethernet hubs together in order to avoid this problem:



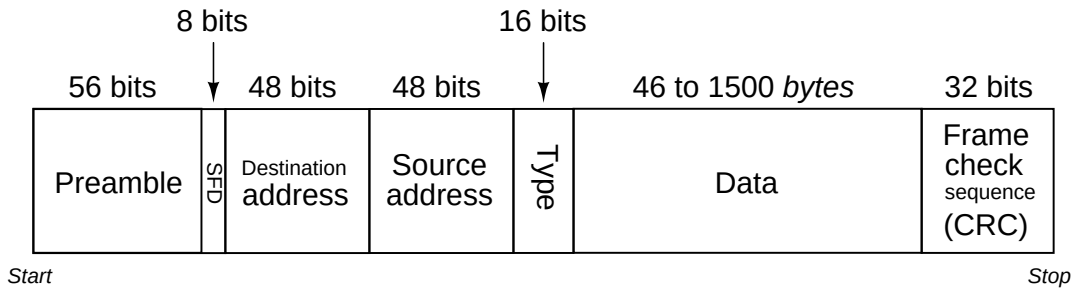
Some early Ethernet hubs provided a different solution to the "crossover" problem, and that was a crossover *switch* built into the hub, allowing a person to manually switch the transmit and receive wire pairs with the push of a button. In this next photograph of a four-port Ethernet hub, you can see the "Normal/Uplink" pushbutton on the right-hand side of the front panel, controlling the furthest-right port of the hub. This switch is supposed to be placed in the "Normal" position if the device plugged into that port is a DTE device, and placed in the "Uplink" position if the device is a DCE device (e.g. another hub):

Note the LED indicator lights by each port on the hub. One LED indicates whether or not the cable is active (when a powered Ethernet DTE device is plugged into that port of the hub), while the other LED indicates traffic on the cable (by blinking). These LEDs are very helpful for identifying a crossover problem. This hub even has an LED indicating the occurrence of collisions (the "Col" LED just below the main power LED), giving simple visual indication of collision frequency.

Some modern hubs use auto-sensing technology to perform any necessary transmit/receive pin swaps, rendering crossover cables and crossover pushbuttons unnecessary for hub-to-hub connections. 1000BASE-T ("Gigabit" Ethernet) hubs have this as a standard feature.

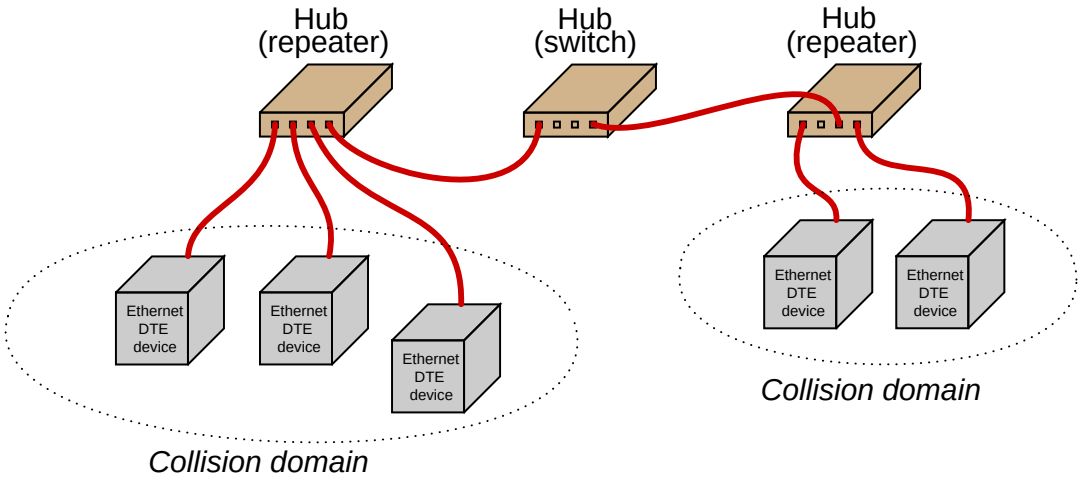
15.4.3 Switching hubs

The next evolutionary step in Ethernet network connections is the introduction of a *switching hub*, or simply *switch*. A "switch" looks exactly like a repeating hub, but it contains intelligence to route transmitted signals only to specific ports, rather than broadcasting every received data frame to all ports. What enables this to happen is the information contained in each Ethernet frame transmitted by DTE devices:



Note that part of the frame includes both a source address and a destination address. These refer to the 48-bit "MAC" addresses uniquely identifying each and every Ethernet device. A switching hub "learns" the identities of all devices plugged into each of its ports by remembering the "source" addresses received through those ports. When a switch receives an Ethernet frame with a destination address it recognizes as residing on one of its ports, it *only* repeats that frame to that specific port, and not to the other ports. This reduces the amount of "traffic" seen at the other ports, and also avoids unnecessary collisions because frames only get sent to their intended destinations. If a switch receives a data frame with an unrecognized destination address, it defaults to basic "hub" behavior by broadcasting that frame to all ports.

The presence of a switching hub in a larger network has the effect of dividing that network into separate collision domains, so that a collision occurring in one domain does not "spill over" into another domain where it would delay communication between those devices:



Of course, collisions between these two domains may still occur, for instance if a device in the first domain tries to transmit to a device in the second domain at the exact same time that a device in the second domain attempts to transmit to a device in the first.

With this added intelligence, switching hubs are considered "layer 2" devices, since they operate not just at the physical layer of electrical impulses, but also at the next layer of device addressing. Since switching hubs add benefit beyond repeating hubs without any drawbacks[38], most people elect to use switches whenever possible.

[38] Even the cost difference is slight.

15.5 Internet Protocol (IP)

I remember first learning about the world-wide Internet, and wondering what it actually *looked like*. The first vision entering my mind when people told me about a computer network spanning nearly all of the United States and many other parts of the world was that of a thick cable strung along telephone poles and buried underground, with a big sign on it saying "Internet." I also remember well the shock of learning that although the Internet made use of several high-capacity networks (called *backbones*) connecting large data centers in different cities, the real "magic" of the Internet did not reside in any particular cable or link. Instead, what made the Internet so widespread and accessible was actually a *protocol* allowing for the free exchange of data along and between disparate systems. This "protocol" allowed digital data to be packaged in such a way that it could be sent along nearly any kind of communications link (from copper wires to fiber-optic to radio waves) – and indeed along multiple pathways between the same two points – while arriving at the destination intact. Thus, the Internet was akin to a random patchwork of existing communications pathways pressed into coordinated service by the sharing of a common "language." In this section, we will investigate the protocol at the heart of the Internet, appropriately called *Internet Protocol*, or *IP*.

Physical network standards such as Ethernet only define aspects relevant to lower layers of the OSI Reference Model. While these details are essential for communication to occur, they are not enough on their own to support a wide-spread communications system. For this reason, network standards such as EIA/TIA-485 and Ethernet almost always comprise the lower layer(s) of a more complex communications protocol capable of managing higher-order addresses, message integrity, "sessions" between computers, and a host of other details.

Internet Protocol (IP) manages network addresses and data handling over a much larger physical domain than Ethernet is able to. The basic principle of IP is that large messages are broken down into *packets* transmitted individually and received individually (then reassembled at the receiver to form the original, complete message). An analogy for this process might be an author with a printed paper manuscript for a book, who needs to get her manuscript to a print shop across town. Unfortunately, the mail service in this town cannot handle the bulky manuscript in one piece, so the author divides the manuscript into bundles of 10 pages each and mails each of these bundles to the print shop, with instructions in each envelope on how to re-assemble the bundles into the complete book. The individual bundles may not make it to the print shop on the same day, or even in the correct order, but the instructions contained within each one make it possible for the people at the print shop to reassemble the entire manuscript once all the bundles have arrived.

This strategy for transmitting large digital messages is at the heart of the Internet: data sent from one computer to another over the internet is first broken down into packets, which are then sent and routed over a variety of pathways to their destination. The receiving computer then reassembles the packets into the original form. This "fragmentation" of data may seem unnecessary, but it actually provides a great deal of flexibility in how data is routed from one point to another.

15.5.1 IP addresses

IP is a "layer 3" technology, being concerned with network-wide addresses for routing information between two different locations. IP is not concerned with the details of communication along any particular wire or fiber-optic cable. It is not "aware" of how bits are represented electrically, or what kind of connectors are used to couple cables together. IP is only concerned with "networks" in the broad sense of the word, as abstract collections of computers that *somehow* (it doesn't care exactly how) are connected to each other.

Networking equipment (DCE) designed to pay attention to IP addresses for routing purposes are called, not surprisingly, *routers*. Their purpose is to direct packets to their appropriate destinations in the shortest amount of time.

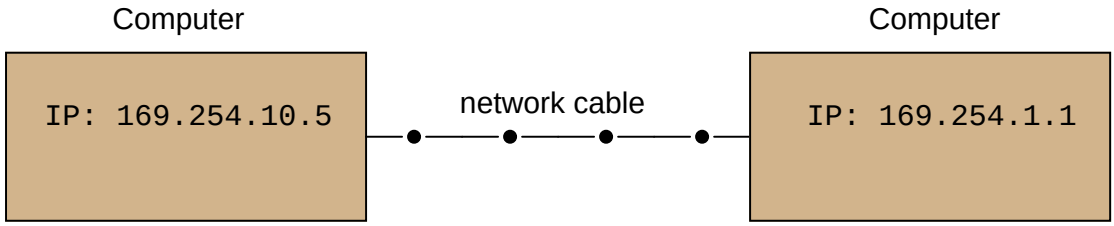
In order for the Internet Protocol to specify where packets are coming from and where they are going to, each source and destination must bear its own *IP address*. IP version 4 (IPv4) uses 32-bit addresses, usually expressed as four octets written using decimal numbers. For example:

IP address `00000000 00000000 00000000 00000000` is written as `0.0.0.0`

IP address `11111111 11111111 11111111 11111111` is written as `255.255.255.255`

IP address `10101001 11111010 00101101 00000011` is written as `169.250.45.3`

In order for two inter-connected computers to exchange data using Internet Protocol, each one must have a unique IP address:



At first, this may seem redundant. Doesn't each and every Ethernet device already have its own unique "MAC address" 48 bits in length to distinguish it from every other Ethernet device in existence? If so, why add *another* set of identifying addresses to the system?

This is true – Ethernet devices are already uniquely addressed – but those MAC addresses serve different purposes than IP addresses. Recall that Ethernet is a standard only at layers 1 and 2, and is not "aware" of any higher-level concerns. Ethernet MAC addresses are useful to switching hubs and other Ethernet DCE devices tasked with management of Ethernet data frames, but those MAC addresses – unique as they may be – have little relevance in the greater picture of IP where we must fragment and reassemble messages over very large-scale networks. More importantly, the reason we need IP addresses is to be able to use interconnecting networks other than Ethernet. For example, two computers may be connected to each other with a simple EIA/TIA-232 cable (or even using radio transceiver units for a "wireless" connection) instead of Ethernet, but still use Internet Protocol to break up large messages and reassemble them at the receiving end[39]. By having

[39] In fact, this is precisely the state of affairs if you use a *dial-up* telephone connection to link your personal computer

its own dedicated addressing scheme, IP ensures computers may be able to disassemble data into packets, send those packets, receive those packets, then re-assemble the packets into the original data *regardless of the physical interconnection details, channel arbitration methods, or anything else in between.* In a sense, IP is the "glue" that holds disparate networks together, and makes something like a world-wide Internet possible when so many different network types exist to connect digital devices together.

A helpful analogy is to think of Ethernet MAC addresses like Social Security numbers for United States citizens. Each US citizen should have their own Social Security number, unique to all living persons. This number is used for many purposes, including identification on Federal tax documents, to help route specific information (such as income records and Social Security payments) to the proper people. Despite the uniqueness of these numbers, though, people still need separate mailing addresses in order to receive mail through the postal service and other package distribution agencies. The mailing address serves a different purpose than the Social Security "address" each US citizen possesses. Furthermore, the existence of separate mailing addresses ensures even non-citizens living in the United States (e.g. foreign students, ambassadors, etc.) who have no Social Security numbers still have a way to send and receive mail.

Given the addressing purpose of Internet Protocol (to designate addresses over an extremely large collection of digital communication devices), addresses must be chosen with care. IP version 4 uses a 32-bit field to designate addresses, limiting its address capacity to 2^{32} unique addresses. As large as this number is, it is not enough to uniquely identify all Internet-capable devices worldwide. The inventors of IP did not dream their Internet would grow to the proportions it has today. Let this be a lesson to all those involved with computers: the future will *always* be bigger than you think! A variety of clever techniques has been developed to deal with this shortage of IP addresses. One of them is to dynamically assign addresses to Internet-connected computers *only when they are turned on.* This is how most personal Internet connections work: when you boot up your personal computer to connect to the Internet, your service provider assigns you a temporary IP address through a protocol called DHCP (Dynamic Host Configuration Protocol). Your provider then forces you to relinquish this temporary IP address when you shut down your computer, so someone else may use it for theirs.

The *Internet Corporation for Assigned Names and Numbers*, or *ICANN*, is the organization responsible[40] for assigning IP addresses to Internet users worldwide (among other tasks). This group has designated certain IP address ranges specific to internal (i.e. *Local Area Network*, or *LAN*) network devices, which shall never be used "publicly" to address devices on the world-wide Internet. These specially-designated "private" LAN address ranges are as follows:

`10.0.0.0` to `10.255.255.255`

`172.16.0.0` to `172.31.255.255`

`192.168.0.0` to `192.168.255.255`

with the Internet. If you use dial-up, your PC may not use Ethernet at all to make the connection to your telephone provider's network, but rather it might uses EIA/TIA-232 or USB to a modem (modulator/demodulator) device, which turns those bits into modulated waveforms transmittable over a voice-quality analog telephone line.

[40] Prior to ICANN's formation in 1999, the *Internet Assigned Numbers Authority*, or *IANA* was responsible for these functions. This effort was headed by a man named Jon Postel, who died in 1998.

Additionally, all computers have their own special *loopback* IP address, used to send IP message packets to itself for certain purposes (including diagnostics): `127.0.0.1`. This IP address is completely *virtual*, not associated with any network hardware at all[41]. Therefore, the `ping` command executed on any computer should *always* be able to detect address 127.0.0.1, regardless of the status or even existence of actual network hardware (cards or interfaces) on that computer. Failure of the `ping` command to detect the loopback address is a sign that the computer's operating system is not configured to use Internet Protocol.

A computer's loopback address may have uses other than diagnostic. Some computer applications are network-oriented by nature, and rely on IP addresses even if the application is performing some local function rather than a function between computers on an actual network. The *X-windows* graphic-user interface (GUI) system popularly used on UNIX operating systems is an example of this, referencing the loopback address to form a connection between client and server applications running on the same computer.

15.5.2 Subnetworks and subnet masks

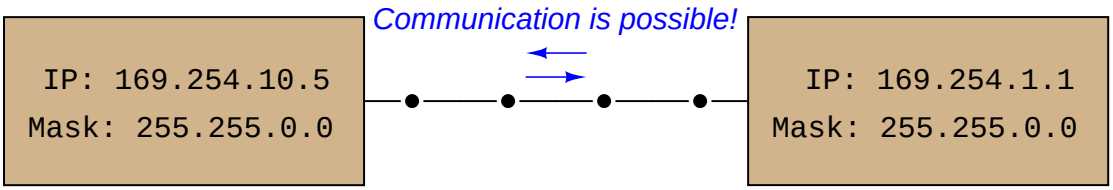
IP (version 4) addresses are used in conjunction with something called *subnet masks*[42] to divide IP networks into "subnetworks." A "subnetwork" is a range of IP-addressed devices allowed to communicate with each other. You may think of the subnet mask to be a sort of "filter" used to identify IP addresses not belonging to the proper range.

The subnet mask works as a bitwise filter, identifying those bits in the IP address defining the subnetwork. For example, if the subnet mask on a computer is set to `255.0.0.0` (binary `11111111 00000000 00000000 00000000`), it means the first 8 bits of the IP address define the subnetwork, and thus the computer is only allowed to communicate with another computer belonging to the same subnetwork (i.e. having the same first octet in its IP address).

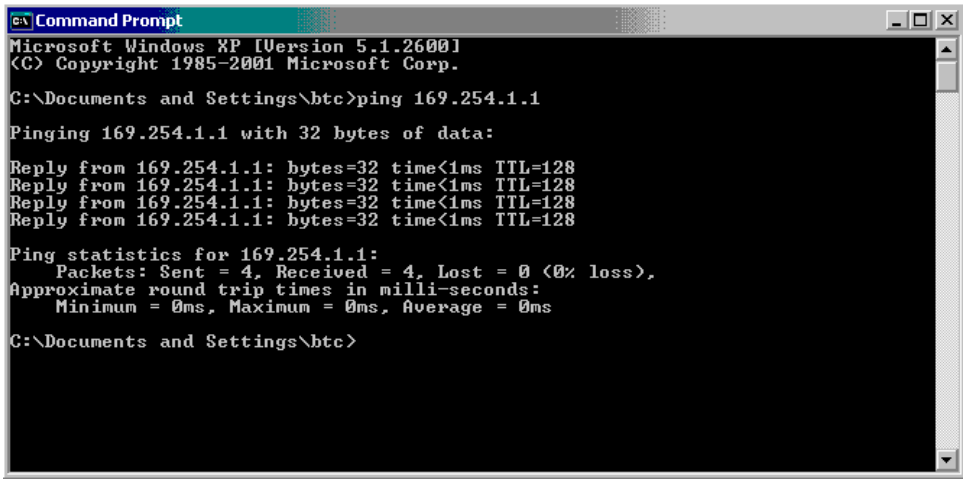
[41] The term "loopback" refers to an old trick used by network technicians to diagnose suspect serial port connections on a computer. Using a short piece of copper wire (or even a paperclip) to "jumper" pins 2 and 3 on an EIA/TIA-232 serial port, any serial data transmitted (out of pin 3) would be immediately received (in pin 2), allowing the serial data to "loop back" to the computer where it could be read. This simple test, if passed, would prove the computer's low-level communication software and hardware was working properly and that any networking problems must lie elsewhere.

[42] Also called "netmasks" or simply "masks."

A set of examples showing two interconnected computers with differing IP addresses (and in some cases, different masks) illustrates how this works[43]. In the first example, two computers with IP addresses differing in the last two octets are able to communicate because their subnets are the same (`169.254`):



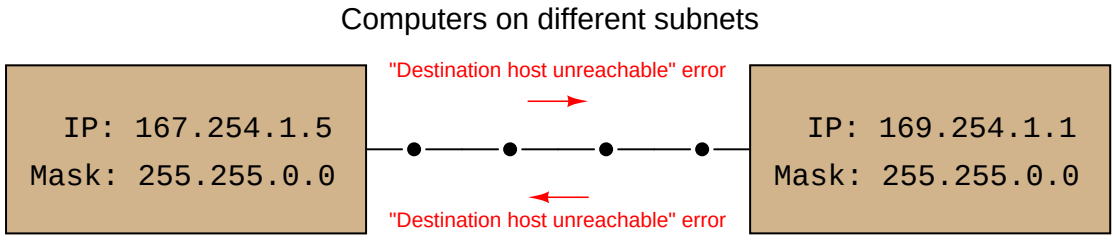
We may check to see the IP addresses and subnet masks are correct by using a command-line program called `ping`, available on nearly all computer systems. A screenshot of `ping` being used on a personal computer running the Microsoft Windows XP operating system is shown here:



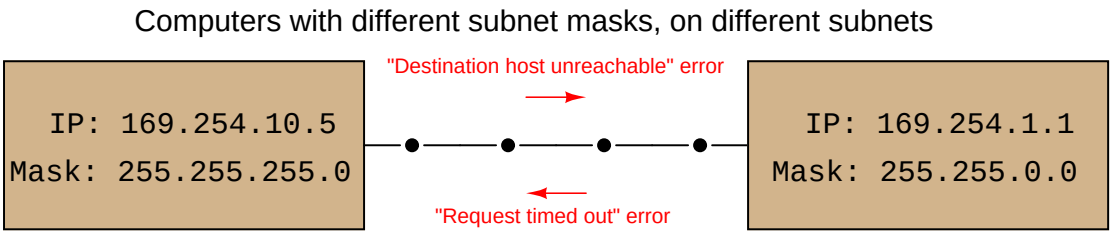
The `ping` utility works by sending a very short digital message to the specified IP address, requesting a reply from that computer. There are usually multiple attempts, with four being shown in this particular example. In fact, it is common among networking professionals to use the word "ping" as a verb, as in "I tried to ping that computer, but it gave no response."

[43] These are real test cases I performed between two computers connected on a 10 Mbps Ethernet network. The error messages are those generated by the `ping` utility when communication was attempted between mis-matched computers.

In the next example, we see two computers with the same mask value, but with different address values in the octets designated by their masks. In other words, these two computers belong to different subnets: one to `167.254` and the other to `169.254`, and as a result they are not allowed to communicate with each other using Internet Protocol. The resulting error messages generated by the `ping` utility are shown in this diagram:



In the last example, we see two computers having different mask values as well as different IP addresses. The subnet of the left-hand computer is `169.254.10` while the subnet of the right-hand computer is `169.254`:



The computer on the left may only communicate with IP addresses matching in the first three octets (`169.254.10`). Seeing that the destination address for the second computer does not match in its third octet, `ping` returns a "Destination host unreachable" error message when executed from the left-hand computer.

When the computer on the right attempts to communicate with ("ping") the computer on the left, it is allowed to transmit to that computer because its mask only screens for agreement in the first two octets (`169.254`), which happen to match. However, the computer on the left is not allowed to transmit to the computer on the right because of its more restrictive subnet, and so `ping` running on the right-hand computer returns a "Request timed out" error message because it never receives a reply from the left-hand computer to any of its queries.

With just two computers connected by a single cable, the concept of subnetworks and masks seems useless, and indeed it is on this small of a scale. However, "subnetting" is a useful technique for managing high traffic loads on large networked systems using IP addresses, and so it is commonly seen in many local area networks (LANs) such as those found at industry and commercial sites.

Another use of `ping` is to search for unknown IP addresses on a known subnet. This may be done by "pinging" to the *broadcast address* for that subnet: an IP address formed by the known subnet numbers, followed by all binary 1's filling the unknown bit spaces. For example, you could use `ping` to search for devices on the subnet `156.71` (subnet mask `255.255.0.0`) by using the following command:

```
ping 156.71.255.255
```

15.5.3 IP version 6

The next version of IP (version 6, or IPv6) uses 128-bit addresses, giving 2^{128} address possibilities (in excess of 3.4×10^{38}), in stark contrast to IPv4's paltry 2^{32} address space. To put this enormous quantity into perspective, there are enough IPv6 addresses to designate nearly 57 *billion* of them for each and every gram of the Earth's mass[44]. While IPv4 addresses are typically written as four octets in decimal form (e.g. 169.254.10.5), this notation would be very cumbersome for writing IPv6 addresses. Thus, IPv6 addresses are written as a set of eight hexadecimal numbers (up to four characters per number) separated by colons, such as `4ffd:522:c441:d2:93b2:f5a:8:101f`. The phase-in of IPv6 to replace IPv4 has already started for certain portions of the Internet, but the full transition to IPv6 is expected to take many years. The IPv6 "loopback" virtual address for computers is `0:0:0:0:0:0:0:1`, or more simply written as `::1`.

15.5.4 DNS

The acronym *DNS* actually stands for two related things: *Domain Name System* and *Domain Name Server*. The first meaning of "DNS" refers to the system of exchanging numerical IP addresses with alphanumeric *Uniform Resource Locators* (*URLs*) which are easier for human beings to remember. When you use web browser software to navigate to a web site on the Internet, you have the option of entering the URL *name* of that site (e.g. `www.google.com`) or a numerical IP address (e.g. `75.125.53.104`). Special computers connected to the Internet called *Domain Name Servers*, and *Domain Name Resolvers* (*DNRs*) use the *Address Resolution Protocol* (*ARP*) to convert your target web site name to its actual IP address so that a connection may be made between that computer and yours.

ICANN, the same organization responsible for allotting IP addresses, also maintains databases for all registered domain names.

[44] According to Douglas Giancoli's *Physics for Scientists and Engineers* textbook, the mass of the Earth is 5.98×10^{24} kg, or 5.98×10^{27} grams. Dividing 2^{128} (the number of unique IPv6 addresses) by the Earth's mass in grams yields the number of available IPv6 address per gram of Earth mass. Furthermore, if we assume a grain of sand has a mass of about 1 milligram, and that the Earth is modeled as a very large collection of sand grains (not quite the truth, but good enough for a dramatic illustration!), we arrive at 57 *million* IPv6 addresses per grain of sand on Earth.

15.5.5 Command-line diagnostic utilities

In addition to `ping`, another utility program useful for troubleshooting network connections from a computer's command line interface is `ipconfig`. When executed, `ipconfig` returns a listing of all available (configured and operating) network interfaces on that computer:

```
C:\Documents and Settings\btc>ipconfig

Windows IP Configuration

Ethernet adapter Local Area Connection:

        Connection-specific DNS Suffix  . :
        IP Address. . . . . . . . . . . . : 169.254.1.2
        Subnet Mask . . . . . . . . . . . : 255.255.0.0
        Default Gateway . . . . . . . . . :

Ethernet adapter Wireless Network Connection:

        Media State . . . . . . . . . . . : Media disconnected

C:\Documents and Settings\btc>
```

The equivalent command for UNIX operating systems is `ifconfig`, shown in this screenshot:

```
root@Renegade2:/home# ifconfig
eth0      Link encap:Ethernet  HWaddr 00:13:20:08:ec:e6
          inet addr:192.168.0.64  Bcast:192.168.0.255  Mask:255.255.255.0
          inet6 addr: fe80::213:20ff:fe08:ece6/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:170901 errors:0 dropped:0 overruns:0 frame:0
          TX packets:107550 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:212178154 (212.1 MB)  TX bytes:14005068 (14.0 MB)

eth1      Link encap:Ethernet  HWaddr 00:0e:35:a2:1b:7f
          inet6 addr: fe80::20e:35ff:fea2:1b7f/64 Scope:Link
          UP BROADCAST MULTICAST  MTU:1500  Metric:1
          RX packets:441 errors:0 dropped:0 overruns:0 frame:0
          TX packets:570 errors:0 dropped:6 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:0 (0.0 B)  TX bytes:0 (0.0 B)
          Interrupt:18 Base address:0x2000 Memory:48005000-48005fff

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          inet6 addr: ::1/128 Scope:Host
          UP LOOPBACK RUNNING  MTU:16436  Metric:1
          RX packets:8 errors:0 dropped:0 overruns:0 frame:0
          TX packets:8 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:516 (516.0 B)  TX bytes:516 (516.0 B)

root@Renegade2:/home#
```

Some of the more interesting details contained in the output from `ifconfig` are the IPv6 addresses (in addition to IPv4 addresses), and details on the "loopback" address (IPv4 `127.0.0.1` or IPv6 `::1`).

A utility intended to reveal the DNS name of a computer given its IP address, or visa versa, is `nslookup`. The same command works on Microsoft Windows and UNIX operating systems alike. Here, we see the UNIX version used to identify four IP addresses of the popular Google search engine web site:

```
root@Renegade2:/home# nslookup www.google.com
Server:         192.168.0.1
Address:        192.168.0.1#53

Non-authoritative answer:
www.google.com  canonical name = www.l.google.com.
Name:   www.l.google.com
Address: 74.125.53.103
Name:   www.l.google.com
Address: 74.125.53.147
Name:   www.l.google.com
Address: 74.125.53.99
Name:   www.l.google.com
Address: 74.125.53.104
```

Another utility used to explore network connections is `traceroute` (spelled `tracert` on Microsoft Windows operating systems). This utility sends a test packet to the designated destination address, returning information on all the "hops" the IP packet takes between computers along the network to reach its destination and the amount of time taken to make the trip. Execution of `traceroute` on a UNIX computer and `tracert` on a Microsoft Windows computer are shown here:

```
root@Renegade2:/home# traceroute www.google.com
traceroute to www.google.com (74.125.53.147), 30 hops max, 40 byte packets
 1  home (192.168.0.1)  4.763 ms  4.803 ms  4.784 ms
 2  tukw-dsl-gw22-214.tukw.qwest.net (63.231.10.214)  57.927 ms  59.993 ms  61.916 ms
 3  tukw-agw1.inet.qwest.net (71.217.184.169)  63.924 ms  65.905 ms  67.882 ms
 4  sea-core-01.inet.qwest.net (67.14.1.194)  71.775 ms  71.784 ms  73.609 ms
 5  sea-brdr-01.inet.qwest.net (205.171.26.54)  75.642 ms  77.442 ms  79.421 ms
 6  63.146.26.198 (63.146.26.198)  81.438 ms  67.052 ms  68.856 ms
 7  sl-gw20-sea-0-0-0.sprintlink.net (144.232.6.8)  70.633 ms  56.617 ms  60.219 ms
 8  sl-googl13-199181-0.sprintlink.net (144.224.13.138)  62.133 ms  64.301 ms  66.162
 9  209.85.249.32 (209.85.249.32)  68.140 ms 209.85.249.34 (209.85.249.34)  70.028 ms
10  216.239.46.204 (216.239.46.204)  79.865 ms  81.739 ms  85.587 ms
11  64.233.174.121 (64.233.174.121)  249.193 ms 64.233.174.129 (64.233.174.129)  113.
12  72.14.232.70 (72.14.232.70)  93.458 ms 72.14.232.10 (72.14.232.10)  99.574 ms 72.
13  72.14.232.6 (72.14.232.6)  68.876 ms 72.14.232.2 (72.14.232.2)  70.251 ms pw-in-f
root@Renegade2:/home#
```

```
U:\>tracert www.google.com

Tracing route to www.l.google.com [209.85.173.99]
over a maximum of 30 hops:

  1  4294964928 ms  4294964927 ms  4294964927 ms  134.39.250.1
  2  4294964927 ms  4294964927 ms  4294964927 ms  bellingham-2691.ctc.edu [192.6
4.1.105]
  3  4294964931 ms  4294964931 ms  4294964930 ms  ge-0-1-0--941.seawescar1.infra
.wa-k20.net [68.179.207.210]
  4  4294964931 ms  4294964931 ms  4294964930 ms  ge-3-0-3--0.seawescore1.infra.
wa-k20.net [68.179.203.26]
  5  4294964931 ms  4294964931 ms  4294964931 ms  ge-2-2-0--311.iccr-sttlwa01-02
.infra.pnw-gigapop.net [209.124.188.182]
  6  4294964931 ms  4294964930 ms  4294964931 ms  pnwgp-cust.tr01-sttlwa01.trans
itrail.net [137.164.131.186]
  7  4294964931 ms  4294964931 ms  4294964931 ms  te4-3--301.tr01-sttlwa01.trans
itrail.net [137.164.131.185]
  8  4294964931 ms  4294964931 ms  4294964931 ms  137.164.130.158
  9  4294964931 ms  4294964931 ms  4294964931 ms  209.85.249.32
 10  4294964935 ms  4294964938 ms  4294964938 ms  216.239.46.208
 11  4294964936 ms  4294964937 ms  4294964937 ms  64.233.174.127
 12  4294964937 ms  4294964936 ms  4294964937 ms  209.85.251.149
 13  4294964943 ms  4294964937 ms  4294964941 ms  209.85.251.145
 14  4294964941 ms  4294964944 ms  4294964937 ms  mh-in-f99.google.com [209.85.1
73.99]

Trace complete.

U:\>_
```

15.6 Transmission Control Protocol (TCP) and User Datagram Protocol (UDP)

At the next OSI Reference Model layer (layer 4) is a set of protocols specifying virtual "ports" at transmitting and receiving devices through which data is communicated. The purpose of these virtual ports is to manage multiple types of data transactions to and from the same IP address, such as in the case of a personal computer accessing a web page (using HTTP) and sending an email message (using SMTP) at the same time. An analogy to help understand the role of ports is to think of multiple packages delivered to different people at a common address such as a business office. The mailing address for the office is analogous to the IP address of a computer exchanging data over a network: it is how other computers on the network "find" that computer. The persons' names or department numbers written on the different packages are analogous to virtual ports on the computer: "places" where specific messages are directed once they arrive at the common address.

Transmission Control Protocol (*TCP*) and *User Datagram Protocol* (*UDP*) are two methods used to manage "ports" on a DTE device, with TCP being the more complex (and robust) of the two. Both TCP and UDP must rely on IP addressing to specify which devices send and receive data, which is why you will often see these protocols listed in conjunction with IP (e.g. TCP/IP and UDP/IP). TCP and UDP are both useless on their own: a protocol specifying port locations without an IP address would be as meaningless as a package placed in the general mail system with just a name or a department number but no street address. Conversely, Internet Protocol anticipates the presence of a higher-level protocol such as TCP or UDP by reserving a portion of its "datagram" (packet) bit space for a "protocol" field[45] to specify which high-level protocol generated the data in the IP packet.

TCP is a complex protocol specifying not only which virtual "ports" will be used at the sending and receiving devices, but also how packet transmission will be guaranteed. A *segment* of data sent via TCP will be error-checked to guard against corruption, marked with special bit flags if of an "urgent" priority, and otherwise marked and labeled in such a way that a high degree of communication reliability is assured. A simplified analogy for TCP's actions might be *certified mail* in the United States postal system, where certain extra steps are taken to ensure delivery and receipt of a parcel.

UDP is a much simpler protocol, lacking many of the data-integrity features of TCP. It is quite common to see UDP applied in industrial settings, where communication takes place over much smaller networks than the world-wide Internet. Another reason UDP is more common in industrial applications is that it is easier to implement in the "embedded" computer hardware at the heart of many industrial devices. The TCP algorithm requires greater computational power and memory capacity than the UDP algorithm, and so it is much easier to engineer a single-chip computer (i.e. microcontroller) to implement UDP than it would be to implement TCP.

[45] Both IPv4 and IPv6 reserve eight bits for this purpose.

Using another utility program on a personal computer called `netstat` (available for both Microsoft Windows and UNIX operating systems) to check active connections[46], we see the various IP addresses and their respective port numbers (shown by the digits following the colon after the IP address) as a list, organized by TCP connections and UDP connections:

```
C:\Documents and Settings\btc>netstat -an

Active Connections

  Proto  Local Address          Foreign Address        State
  TCP    0.0.0.0:135            0.0.0.0:0              LISTENING
  TCP    0.0.0.0:445            0.0.0.0:0              LISTENING
  TCP    0.0.0.0:2869           0.0.0.0:0              LISTENING
  TCP    127.0.0.1:1025         0.0.0.0:0              LISTENING
  TCP    127.0.0.1:5152         0.0.0.0:0              LISTENING
  TCP    169.254.1.2:23         169.254.1.1:1116       ESTABLISHED
  TCP    169.254.1.2:139        0.0.0.0:0              LISTENING
  UDP    0.0.0.0:445            *:*
  UDP    0.0.0.0:500            *:*
  UDP    0.0.0.0:1062           *:*
  UDP    0.0.0.0:4500           *:*
  UDP    0.0.0.0:7725           *:*
  UDP    127.0.0.1:123          *:*
  UDP    127.0.0.1:1063         *:*
  UDP    127.0.0.1:1066         *:*
  UDP    127.0.0.1:1900         *:*
  UDP    169.254.1.2:123        *:*
  UDP    169.254.1.2:137        *:*
  UDP    169.254.1.2:138        *:*
  UDP    169.254.1.2:1900       *:*

C:\Documents and Settings\btc>
```

Many different port numbers have been standardized for different applications at OSI Reference Model layers above 4 (above that of TCP or UDP). Port 25, for example, is always used for SMTP (Simple Mail Transfer Protocol) applications. Port 80 is used by HTTP (HyperText Transport Protocol), a layer-7 protocol used to view Internet "web" pages. Port 107 is used by TELNET applications, a protocol whose purpose it is to establish command-line connections between computers for remote administrative work. Port 22 is used by SSH, a protocol similar to TELNET but with significantly enhanced security. Port 502 is designated for use with Modbus messages communicated over TCP/IP.

[46] In this particular case, I typed `netstat -an` to specify *all* (`a`) ports with *numerical* (`n`) IP addresses and port numbers shown.

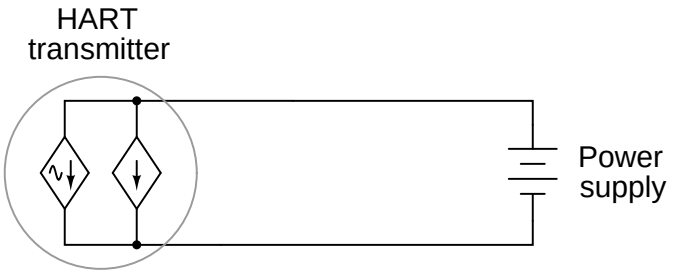
15.7 The HART digital/analog hybrid standard

A technological advance introduced in the late 1980's was *HART*, an acronym standing for **H**ighway **A**ddressable **R**emote **T**ransmitter. The purpose of the HART standard was to create a way for instruments to digitally communicate with one another over the same two wires used to convey a 4-20 mA analog instrument signal. In other words, HART is a *hybrid* communication standard, with one variable (channel) of information communicated by the analog value of a 4-20 mA DC signal, and another channel for digital communication whereby many other variables could be communicated using pulses of current to represent binary bit values of 0 and 1. Those digital current pulses are *superimposed* upon the analog DC current signal, such that the same two wires carry both analog and digital data simultaneously.

The HART standard was developed with existing installations in mind. The medium for digital communication had to be robust enough to travel over twisted-pair cables of very long length and unknown characteristic impedance. This meant that the data communication rate for the digital data had to be very slow, even by 1980's standards. The HART standard is concerned only with layers 1 (FSK modulation, $\pm$ 0.5 mA signaling), 2 (Master-slave arbitration, data frame organization), and 7 (specific commands to read and write device data) of the OSI Reference model. Layers 3 through 6 are irrelevant to the HART standard.

Digital data is encoded in HART using the Bell 202 modem standard: two audio-frequency "tones" (1200 Hz and 2200 Hz) are used to represent the binary states of "1" and "0," respectively, transmitted at a rate of 1200 bits per second. This is known as *frequency-shift keying*, or *FSK*. The physical representation of these two frequencies is an AC current of 1 mA peak-to-peak superimposed on the 4-20 mA DC signal. Thus, when a HART-compatible device "talks" digitally on a two-wire loop circuit, it produces tone bursts of AC current at 1.2 kHz and 2.2kHz. The receiving HART device "listens" for these AC current frequencies and interprets them as binary bits.

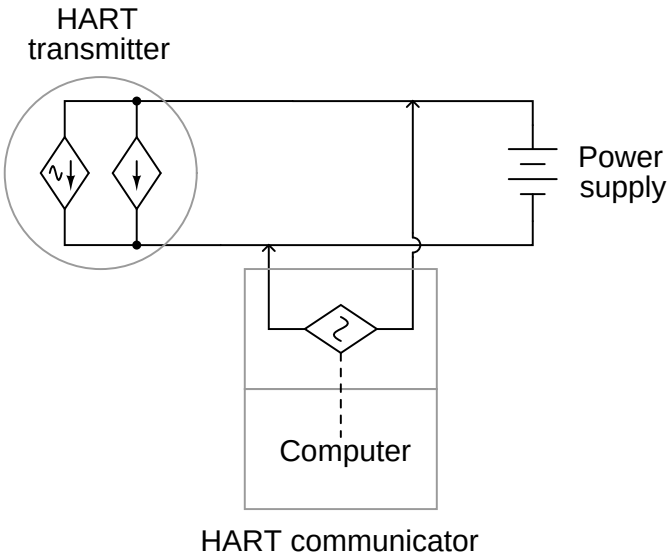
An important consideration in HART current loops is that the total loop resistance (precision resistor values plus wire resistance) must fall within a certain range: 250 ohms to 1100 ohms. Most 4-20 mA loops (containing a single 250 ohm resistor for converting 4-20 mA to 1-5 V) measure in at just over 250 ohms total resistance, and work quite well with HART. Even loops containing two 250 ohm precision resistors meet this requirement. Where technicians often encounter problems is when they set up a loop-powered HART transmitter on the test bench with a lab-style power supply and *no* 250 ohm resistor anywhere in the circuit:



The HART transmitter may be modeled as two parallel current sources: one DC and one AC. The DC current source provides the 4-20 mA regulation necessary to represent the process measurement as an analog current value. The AC current source turns on and off as necessary to "inject" the 1

mA P-P audio-frequency HART signal along the two wires. Inside the transmitter is also a HART modem for interpreting AC voltage tones as HART data packets. Thus, data transmission takes place through the AC current source, and data reception takes place through a voltage-sensitive modem, all inside the transmitter, all "talking" along the same two wires that carry the DC 4-20 mA signal.

For ease of connection in the field, HART devices are designed to be connected in parallel with each other. This eliminates the need to break the loop and interrupt the DC current signal every time we wish to connect a HART communicator device to communicate with the transmitter. A typical HART communicator may be modeled as an AC voltage source[47] (along with another HART voltage-sensitive modem for receiving HART data). Connected in parallel with the HART transmitter, the complete circuit looks something like this:



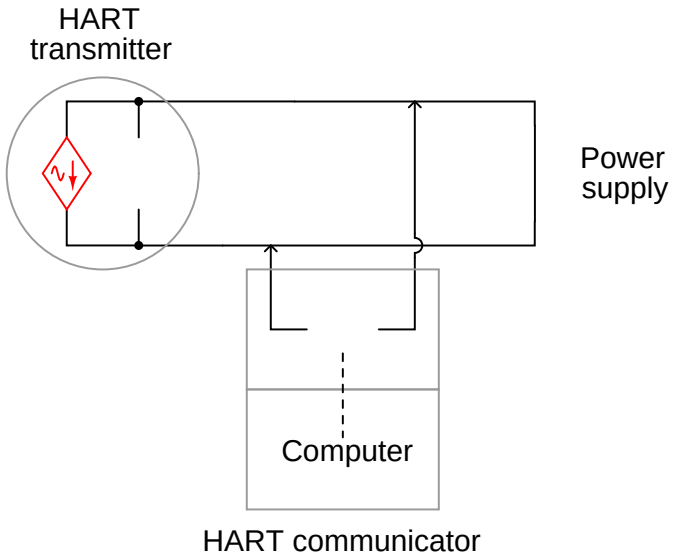
[47] The HART standard specifies "master" devices in a HART network transmit AC voltage signals, while "slave" devices transmit AC current signals.

The actual hand-held communicator may look like one of these devices:

With all these sources in the same circuit, it is advisable to use the *Superposition Theorem* for analysis. This involves "turning off" all but one source at a time to see what the effect is for each source, then superimposing the results to see what all the sources do when all are working simultaneously.

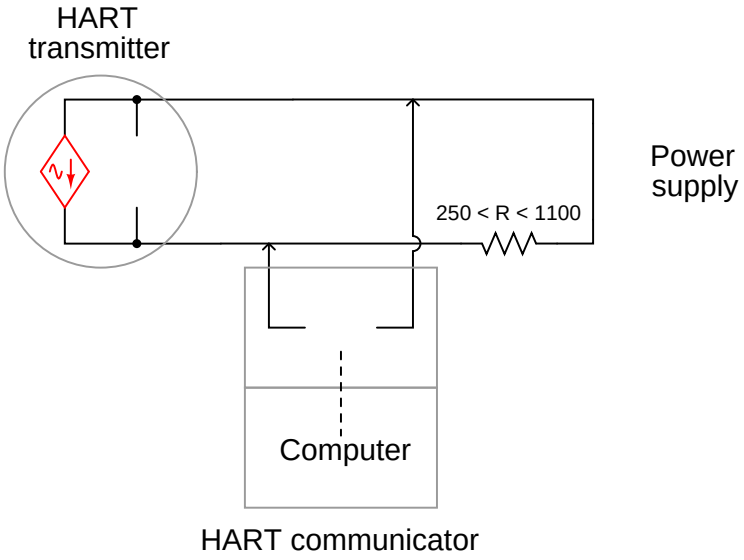
We really only need to consider the effects of either AC source to see what the problem is in this circuit with no loop resistance. Consider the situation where the transmitter is sending HART data to the communicator. The AC current source inside the transmitter will be active, injecting its 1 mA P-P audio-frequency signal onto the two wires of the circuit. The AC voltage source in the communicator will disconnect itself from the network, allowing the communicator to "listen" to the transmitter's data.

To apply the Superposition Theorem, we replace all the other sources with their own equivalent internal resistances (voltage sources become "shorts," and current sources become "opens"):



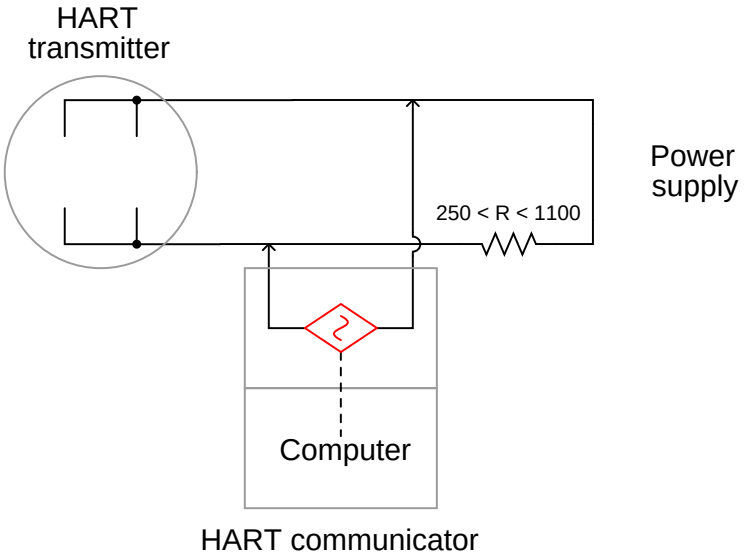
The HART communicator is "listening" for those audio tone signals sent by the transmitter's AC source, but it "hears" nothing because the DC power supply's equivalent short-circuit prevents any significant AC voltage from developing across the two wires. This is what happens when there is no loop resistance: no HART device is able to receive data sent by any other HART device.

The solution to this dilemma is to install a resistance of at least 250 ohms but not greater than 1100 ohms between the DC power source and all other HART devices, like this:



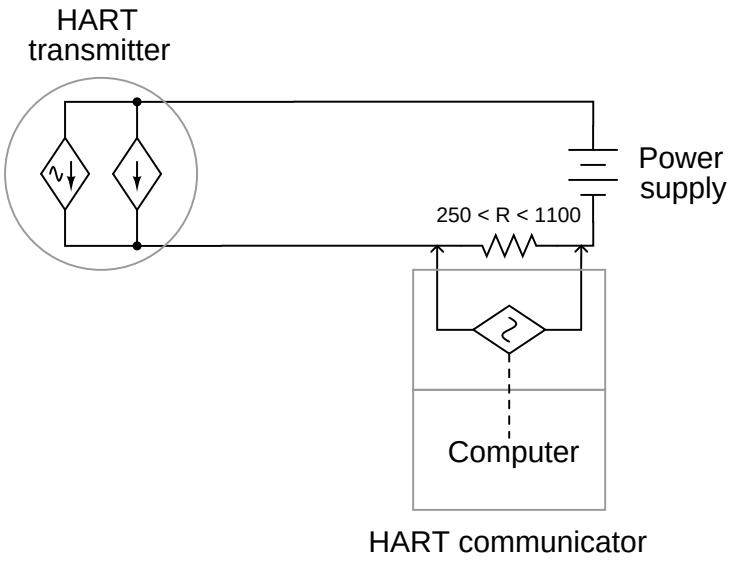
Loop resistance must be at least 250 ohms to allow the 1 mA P-P AC signal to develop enough voltage to be reliably detected by the HART modem in the listening device. The upper limit (1100 ohms) is not a function of HART communication so much as it is a function of the DC voltage drop, and the need to maintain a minimum DC terminal voltage at the transmitter for its own operation. If there is too much loop resistance, the transmitter will become "starved" of voltage and act erratically. In fact, even 1100 ohms of loop resistance may be too much if the DC power supply voltage is insufficient.

Loop resistance is also necessary for the HART transmitter to receive data signals transmitted by the HART communicator. If we analyze the circuit when the HART communicator's voltage source is active, we get this result:



Without the loop resistance in place, the DC power supply would "short out" the communicator's AC voltage signal just as effectively as it shorted out the transmitter's AC current signal. The presence of a loop resistor in the circuit prevents the DC power supply from "loading" the AC voltage signal by the communicator. This AC voltage is seen in the diagram as being directly in parallel with the transmitter, where its internal HART modem receives the audio tones and processes the data packets.

Manufacturers' instructions generally recommend HART communicator devices be connected directly in parallel with the HART field instrument, as shown in the previous schematic diagrams. However, it is also perfectly valid to connect the communicator device directly in parallel with the loop resistor like this:



Connected directly in parallel with the loop resistor, the communicator is able to receive transmissions from the HART transmitter just fine, as the DC power source acts as a dead short to the AC current HART signal and passes it through to the transmitter.

This is nice to know, as it is often easier to achieve an alligator-clip connection across the leads of a resistor than it is to clip in parallel with the loop wires when at a terminal strip or at the controller end of the loop circuit.

HART technology has given a new lease on the venerable 4-20 mA analog instrumentation signal standard. It has allowed new features and capabilities to be added on to existing analog signal loops without having to upgrade wiring or change all instruments in the loop. Some of the features of HART are listed here:

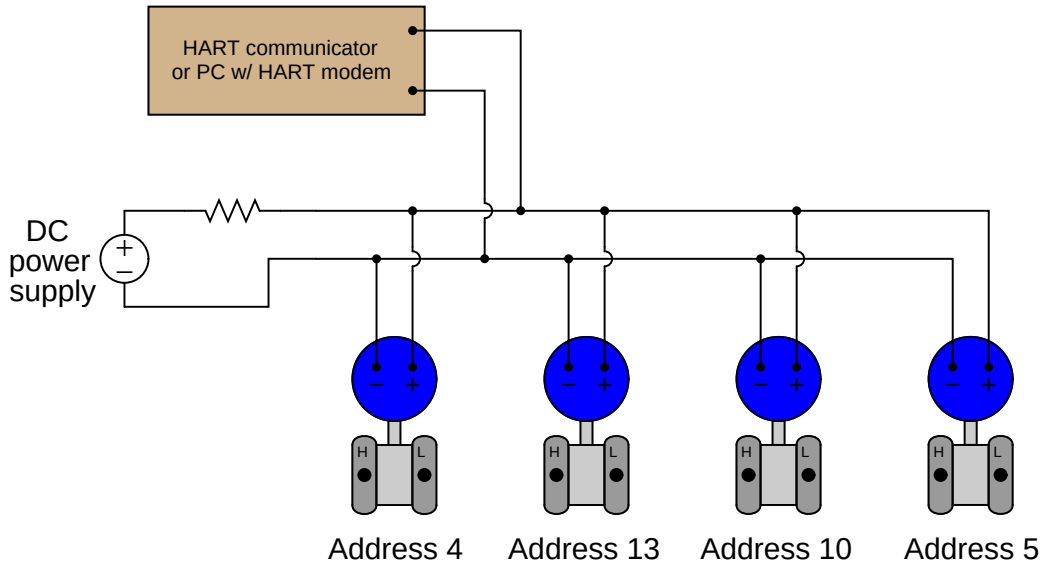
- Diagnostic data may be transmitted by the field device (self-test results, out-of-limit alarms, preventative maintenance alerts, etc.)
- Field instruments may be re-ranged remotely through the use of HART communicators
- Technicians may use HART communicators to force field instruments into different "manual" modes for diagnostic purposes (e.g. forcing a transmitter to output a fixed current so as to check calibration of other loop components, manually stroking a valve equipped with a HART-capable positioner)
- Field instruments may be programmed with identification data (e.g. tag numbers corresponding to plant-wide instrument loop documentation)

15.7.1 HART multidrop mode

The HART standard also supports a mode of operation that is totally digital, and capable of supporting multiple HART instruments on the same pair of wires. This is known as *multidrop mode*.

Every HART instrument has an *address* number, which is typically set to a value of zero (0). A network address is a number used to distinguish one device from another on a broadcast network, so messages broadcast across the network may be directed to specific destinations. When a HART instrument operates in digital/analog hybrid mode, where it must have its own dedicated wire pair for communicating the 4-20 mA DC signal between it and an indicator or controller, there is no need for a digital address. An address becomes necessary only when multiple devices are connected to the same network wiring, and there arises a need to digitally distinguish one device from another on the same network.

This is a functionality the designers of HART intended from the beginning, although it is frequently unused in industry. Multiple HART instruments may be connected directly in parallel with one another along the same wire pair, and information exchanged between those instruments and a host system, if the HART address numbers are set to non-zero values (between 1 and 15):



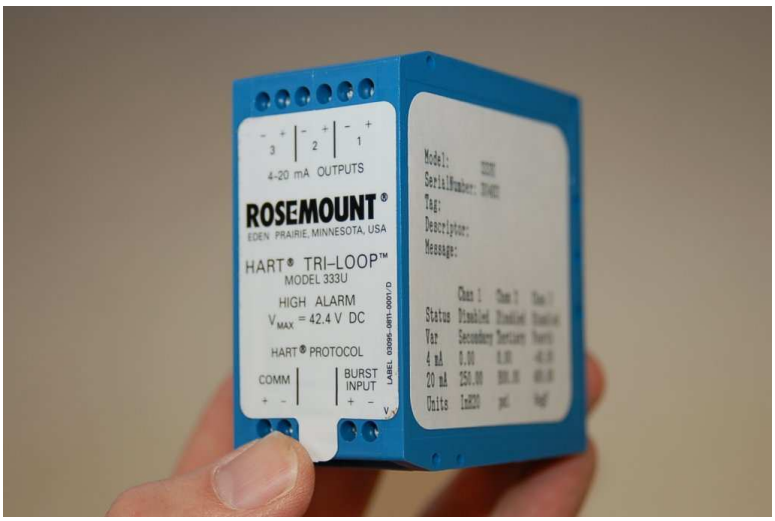
Setting an instrument's HART address to a non-zero value is all that is necessary to engage multidrop mode. The address numbers themselves are irrelevant, as long as they fall within the range of 1 to 15 and are unique to that network.

The major disadvantage of using HART instruments in multidrop mode is its slow speed. Due to HART's slow data rate (1200 bits per second), it may take several seconds to access a particular instrument's data on a multidropped network. For some applications such as temperature measurement, this slow response time may be acceptable. For inherently faster processes such as liquid flow control, it would not be nearly fast enough to provide up-to-date information for the control system to act upon.

15.7.2 HART multi-variable transmitters

Some "smart" instruments have the ability to report multiple process variables. A good example of this is Coriolis-effect flowmeters, which by their very nature simultaneously measure the density, flow rate, and temperature of the fluid passing through them. A single pair of wires can only convey one 4-20 mA analog signal, but that same pair of wires may convey multiple digital signals encoded in the HART protocol. Digital signal transmission is required to realize the full capability of such "multi-variable" transmitters.

If the host system receiving the transmitter's signal(s) is HART-ready, it may digitally poll the transmitters for all variables. If, however, the host system does not "talk" using the HART protocol, some other means must be found to "decode" the wealth of digital data coming from the multi-variable transmitter. One such device is Rosemount's model 333 HART "Tri-Loop" demultiplexer shown in the following photograph:



This device polls the multi-variable transmitter and converts up to three HART variables into independent 4-20 mA analog output signals, which any suitable analog indicator or controller device may receive.

It should be noted that the same caveat applicable to multidrop HART systems (i.e. slow speed) applies to HART polling of multi-variable transmitters. HART is a relatively slow digital bus standard, and as such it should never be considered for applications demanding quick response. In applications where speed is not a concern, however, it is a very practical solution for acquiring multiple channels of data over a single pair of wires.

15.8 Modbus

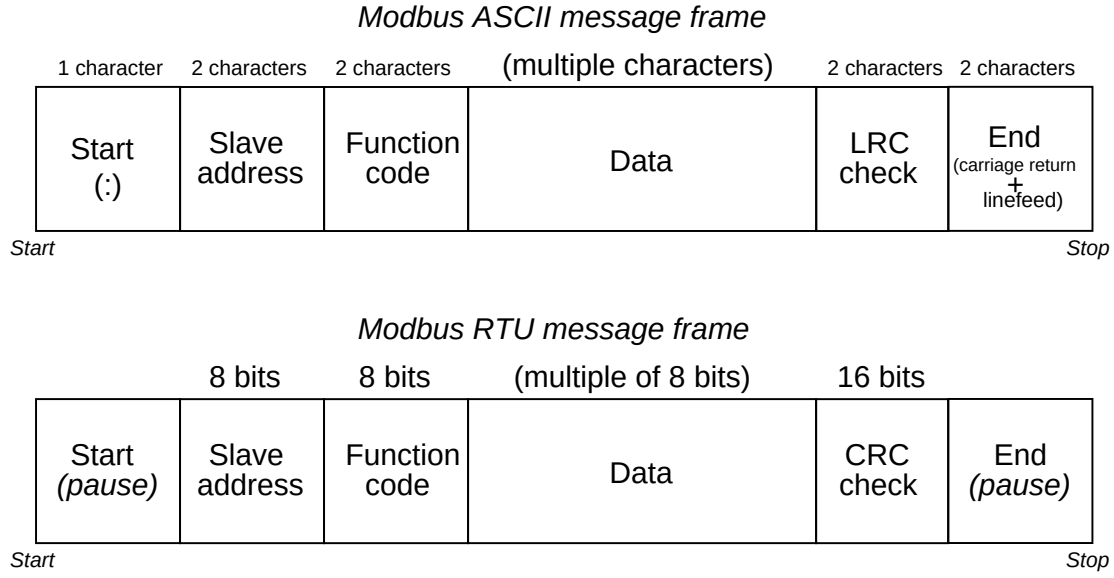
Developed by the Modicon company (the original manufacturer of the *Programmable Logic Controller*, or *PLC*) in 1979 for use in its industrial control products, *Modbus* is a protocol designed specifically for exchanging process data between industrial control devices. The Modbus standard does not specify any details of physical networking, and thus may be deployed on any number of physical networks such as EIA/TIA-232, EIA/TIA-485, Ethernet (the latter via TCP/IP), and a special token-passing network also developed by Modicon called *Modbus Plus*. Modbus itself is primarily concerned with details at layer 7 of the OSI Reference Model, the so-called *Application Layer*.

Modbus, especially when implemented over simple serial networks such as EIA/TIA-232 and EIA/TIA-485, is a rather primitive protocol. The seemingly arbitrary numerical codes used to issue commands and specify addresses is antiquated by modern standards. For better or for worse, though, a great many digital industrial devices "speak" Modbus, even if they are also capable of communicating via some other network protocol.

15.8.1 Modbus data frames

Modbus defines a set of commands for reading (receiving) and writing (transmitting) data between a master device and one or more slave devices connected to the network. Each of these commands is referenced by a numerical code, with addresses of the master and slave devices' internal registers (data sources and data destinations) specified along with the function code in the Modbus message frame.

Two different formats are specified in the Modbus standard: *ASCII* and *RTU*. The difference between these two modes is how addresses, function codes, data, and error-checking bits are represented. In Modbus ASCII mode, all slave device addresses, function codes, and data are represented in the form of ASCII characters (7 bits each), which may be read directly by any terminal program (e.g. `minicom`, `Hyperterminal`, `kermit`, etc.) intercepting the serial data stream. This makes troubleshooting easier: to be able to directly view the Modbus data frames in human-readable form. In Modbus RTU mode, all slave device addresses, function codes, and data are represented directly in hexadecimal form, with four bits per hexadecimal character. Different error-checking techniques are used for ASCII and RTU modes as well. The following diagram compares data frames for the two Modbus communication modes:



As you can see from a comparison of the two frames, ASCII frames require nearly twice the number of bits as RTU frames, making Modbus ASCII slower than Modbus RTU for any given data rate (bits per second).

The contents of the "Data" field vary greatly depending on which function is invoked, and whether or not the frame is issued by the master device or from a slave device. More details on Modbus "Data" field contents will appear in a later subsection.

Since Modbus is strictly a "layer 7" protocol, these message frames are usually embedded within other data frames specified by lower-level protocols. For example, Modbus over TCP/IP encapsulates

individual Modbus data frames as TCP/IP packets, which are then (usually) encapsulated again as Ethernet packets to arrive at the destination device. This "multi-layered" approach inherent to Modbus being such a high-level protocol may seem cumbersome, but it offers great flexibility in that Modbus frames may be communicated over nearly any kind of virtual and physical network type.

15.8.2 Modbus function codes and addresses

A listing of commonly-used Modbus function codes appears in the following table:

Modbus code (decimal)	Function
01	Read one or more PLC output "coils" (1 bit each)
02	Read one or more PLC input "contacts" (1 bit each)
03	Read one or more PLC "holding" registers (16 bits each)
04	Read one or more PLC analog input registers (16 bits each)
05	Write (force) a single PLC output "coil" (1 bit)
06	Write (preset) a single PLC "holding" register (16 bits)
15	Write (force) multiple PLC output "coils" (1 bit each)
16	Write (preset) multiple PLC "holding" registers (16 bits each)

Modbus "984" addressing defines sets of fixed numerical ranges where various types of data may be found in a PLC or other control device. The absolute address ranges (according to the Modbus 984 scheme) are shown in this table:

Modbus codes (decimal)	Address range (decimal)	Purpose
01, 05, 15	00001 to 09999	Discrete outputs ("coils")
02	10001 to 19999	Discrete inputs ("contacts")
04	30001 to 39999	Analog input registers
03, 06, 16	40001 to 49999	"Holding" registers

This fixed addressing scheme usually does not match conveniently to the addressing within the master or slave devices. Manufacturer's documentation for Modbus-compatible devices normally provide Modbus "mapping" references so technicians and engineers alike may determine which Modbus addresses refer to specific bit or word registers in the device.

Note how all the Modbus address ranges begin at the number one, not zero as is customary for so many digital systems. For example, a PLC with sixteen discrete input channels numbered 0 through 15 by the manufacturer may "map" those inputs ("contacts") to Modbus addresses 10001 through 10016, respectively.

Coil, Contact, and Register addresses are specified within Modbus data frames relative to the starting points of their respective commands. For example, the function to read discrete inputs ("contacts") on a slave device only applies to Modbus absolute addresses 10001 through 19999, and so the actual Modbus command issued to read contact address 10005 will specify the address as 0004 (since 10005 is the *fourth* address up from the start of the contact address block 10001 through 19999), and this *relative* address value is always communicated as a hexadecimal (not decimal) value[48].

[48] If the process of converting a device's I/O addresses to Modbus commands sounds to you like it would be about as much fun as being beat with a rubber hose, you are correct. In order to appease those of us lacking masochistic tendencies, some digital device manufacturers include Modbus address "translation" features into their products, so the programmer (you) does not have to burden yourself with offset calculations and decimal-to-hexadecimal conversions just to move a few blocks of data between devices on a Modbus network.

15.8.3 Modbus function command formats

Every Modbus data frame, whether ASCII or RTU mode, has a field designated for "data." For each Modbus function, the content of this "data" field follows a specific format. It is the purpose of this subsection to document the data formats required for common Modbus functions, both the "Query" message transmitted by the Modbus master device to a slave device, and the corresponding "Response" message transmitted back to the master device by the queried slave device.

Since each Modbus data frame is packaged in multiples of 8 bits (RTU), they are usually represented in text as individual bytes (two hexadecimal characters). For example, a 16-bit "word" of Modbus data such as `1100100101011011` would typically be documented as `C9 5B` with a deliberate space separating the "high" (`C9`) and "low" (`5B`) bytes.

Function code 01 – Read Coil(s)

This Modbus function reads the statuses of slave device discrete outputs ("coils") within the slave device, returning those statuses in blocks of eight (even if the "number of coils" specified in the query is not a multiple of eight!). Relevant Modbus addresses for this function range from 00001 to 09999 (decimal) but the starting address is a hexadecimal number representing the $(n-1)^{th}$ register from the beginning of this range (e.g. decimal address 00100 would be specified as hexadecimal `00 63`).

Query message (Function code 01)

Start	Slave address	Function code	Data				Error check	End
			Starting address		Number of coils			
	XX	01	Hi	Lo	Hi	Lo	XX	

Start Stop

Response message (Function code 01)

Start	Slave address	Function code	Data				Error check	End
			Number of bytes	First byte (8 coils)	Second byte (8 coils)	Third byte (8 coils)		
	XX	01					XX	

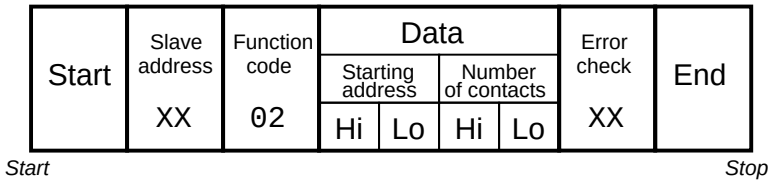
Start Stop

Note that the second and third bytes representing coil status are shown in grey, because their existence assumes more than one byte worth of coils has been requested in the query.

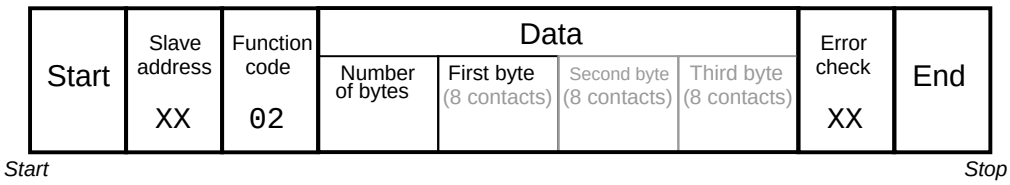
Function code 02 – Read Contact(s)

This Modbus function reads the statuses of slave device discrete inputs ("contacts") within the slave device, returning those statuses in blocks of eight (even if the "number of contacts" specified in the query is not a multiple of eight!). Relevant Modbus addresses for this function range from 10001 to 19999 (decimal), but the starting address is a hexadecimal number representing the $(n-1)^{th}$ register from the beginning of this range (e.g. decimal address 10256 would be specified as hexadecimal `00 FF`).

Query message (Function code 02)



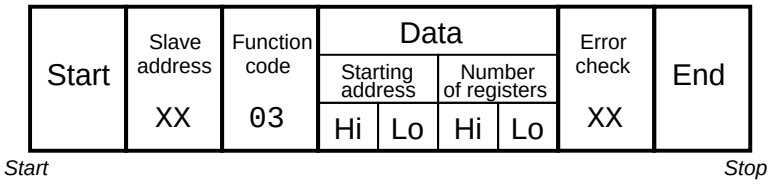
Response message (Function code 02)



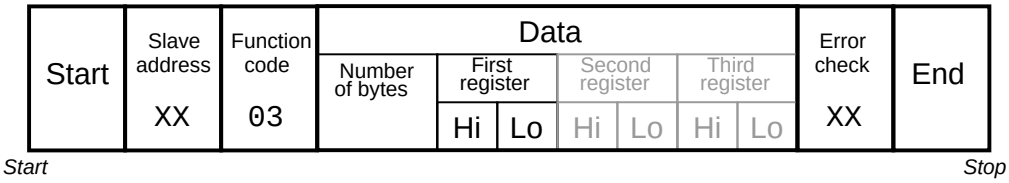
Function code 03 – Read Holding Register(s)

This Modbus function reads the statuses of "holding" registers within the slave device, with the size of each register assumed to be two bytes (16 bits). Relevant Modbus addresses for this function range from 40001 to 49999 (decimal), but the starting address is a hexadecimal number representing the $(n-1)^{th}$ register from the beginning of this range (e.g. decimal address 40980 would be specified as hexadecimal `03 D3`).

Query message (Function code 03)



Response message (Function code 03)

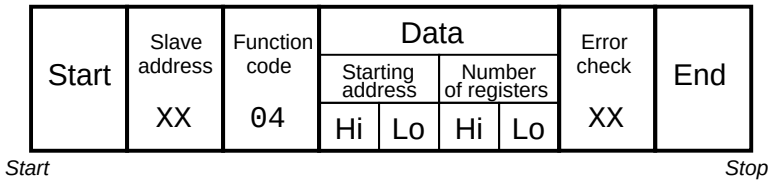


Note that since the query message specifies the number of registers (each register being two bytes in size), and the response message replies with the number of *bytes*, the response message's "number of bytes" field will have a value twice that of the query message's "number of registers" field. Note also that the maximum number of registers which may be requested in the query message (65536) with "high" and "low" byte values grossly exceeds the number of bytes the response message can report (255) with its single byte value.

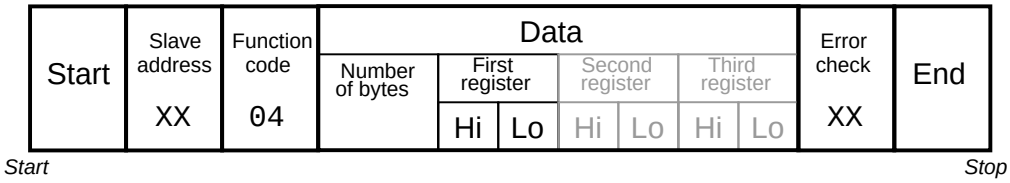
Function code 04 – Read Analog Input Register(s)

This Modbus function is virtually identical to 03 (Read Holding Registers) except that it reads "input" registers instead: addresses 30001 through 39999 (decimal). As with all the Modbus relative addresses, the starting address specified in both messages is a hexadecimal number representing the $(n-1)^{th}$ register from the beginning of this range (e.g. decimal address 32893 would be specified as hexadecimal `0B 4C`).

Query message (Function code 04)



Response message (Function code 04)

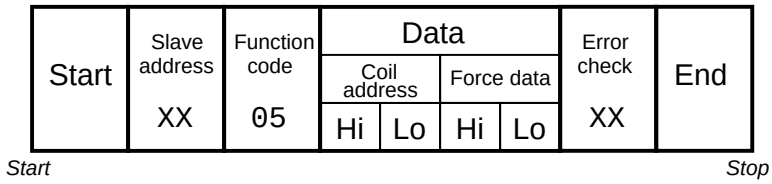


Note that since the query message specifies the number of registers (each register being two bytes in size), and the response message replies with the number of *bytes*, the response message's "number of bytes" field will have a value twice that of the query message's "number of registers" field. Note also that the maximum number of registers which may be requested in the query message (65536) with "high" and "low" byte values grossly exceeds the number of bytes the response message can report (255) with its single byte value.

Function code 05 – Write (Force) Single Coil

This Modbus function writes a single bit of data to a discrete output ("coil") within the slave device. Relevant Modbus addresses for this function range from 00001 to 09999 (decimal) but the starting address is a hexadecimal number representing the $(n-1)^{th}$ register from the beginning of this range (e.g. decimal address 07200 would be specified as hexadecimal `1C 1F`).

Query/Response message (Function code 05)



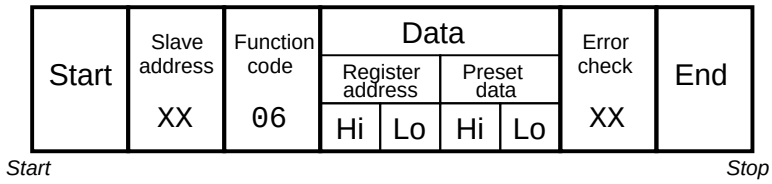
The "force data" for a single coil consists of either `00 00` (force coil off) or `FF 00` (force coil on). No other data values will suffice – anything other than `00 00` or `FF 00` will be ignored by the slave device.

A normal response message will be a simple echo (verbatim repeat) of the query message.

Function code 06 – Write (Preset) Single Holding Register

This Modbus function writes data to a single "holding" register within the slave device. Relevant Modbus addresses for this function range from 40001 to 49999 (decimal) but the starting address is a hexadecimal number representing the $(n-1)^{th}$ register from the beginning of this range (e.g. decimal address 40034 would be specified as hexadecimal `00 21`).

Query/Response message (Function code 06)

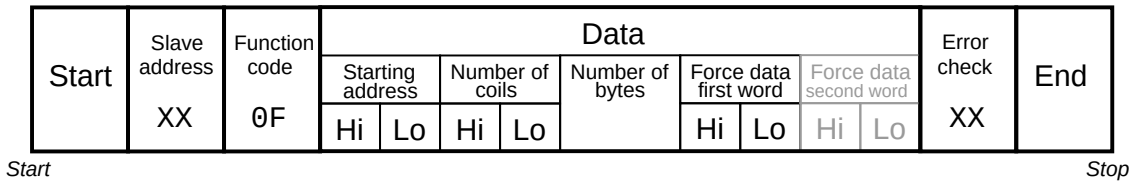


A normal response message will be a simple echo (verbatim repeat) of the query message.

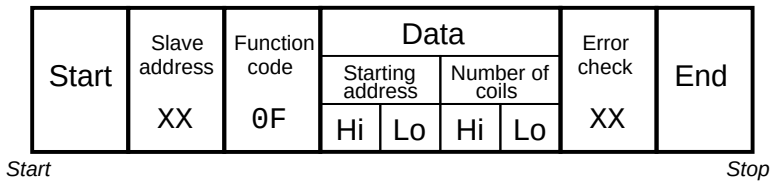
Function code 15 – Write (Force) Multiple Coils

This Modbus function writes multiple bits of data to a set of discrete outputs ("coils") within the slave device. Relevant Modbus addresses for this function range from 00001 to 09999 (decimal) but the starting address is a hexadecimal number representing the $(n-1)^{th}$ register from the beginning of this range (e.g. decimal address 03207 would be specified as hexadecimal `0C 86`).

Query message (Function code 15)



Response message (Function code 15)



Note that the query message specifies both the number of coils (bits) and the number of bytes.

Function code 16 – Write (Preset) Multiple Holding Register

This Modbus function writes multiple words of data to a set of "holding" registers within the slave device. Relevant Modbus addresses for this function range from 40001 to 49999 (decimal) but the starting address is a hexadecimal number representing the $(n-1)^{th}$ register from the beginning of this range (e.g. decimal address 47441 would be specified as hexadecimal `1D 10`).

Query message (Function code 16)

Start	Slave address	Function code	Data								Error check	End
			Starting address		Number of registers		Number of bytes	Preset data first register		Preset data second register		
	XX	10	Hi	Lo	Hi	Lo		Hi	Lo	Hi	Lo	XX
Start												*Stop*

Response message (Function code 16)

Start	Slave address	Function code	Data				Error check	End
			Starting address		Number of registers			
	XX	10	Hi	Lo	Hi	Lo	XX	
Start								*Stop*

Note that the query message specifies both the number of registers (16-bit words) and the number of bytes, which is redundant (the number of bytes must *always* be twice the number of registers, given that each register is two bytes[49] in size). Note also that the maximum number of registers which may be requested in the query message (65536) with "high" and "low" byte values grossly exceeds the number of bytes the response message can report (255) with its single byte value.

[49] Even for devices where the register size is less than two bytes (e.g. Modicon M84 and 484 model controllers have 10 bits within each register), data is still addressed as two bytes' worth per register, with the leading bits simply set to zero to act as placeholders.

References

"422 and 485 Standards Overview and System Configurations" Application Report SLLA070C, Texas Instruments Incorporated, Dallas, TX, 2002.

"B&B Converters for the Industrial Bus World" Technical Article 13, B&B Electronics Manufacturing Company, Ottawa, IL, 2000.

Floyd, Thomas L., *Digital Fundamentals*, 6th edition, Prentice-Hall, Inc., Upper Saddle River, NJ, 1997.

"FOUNDATION Fieldbus System Engineering Guidelines" (AG 181) Revision 2.0, The Fieldbus Foundation, 2004.

"FOUNDATION Specification System Architecture" (FF 581) Revision FS 1.1, The Fieldbus Foundation, 2000.

"Fundamentals of RS-232 Serial Communications" Application Note 83 (AN83), Maxim Integrated Products, 2001.

Giancoli, Douglas C., *Physics for Scientists & Engineers*, Third Edition, Prentice Hall, Upper Saddle River, NJ, 2000.

Graham, Frank D., *Audels New Electric Library, Volume IX*, Theo. Audel & Co., New York, NY, 1942.

HART Communications, Technical Information L452 EN; SAMSON AG, 1999.

Horak, Ray, *Telecommunications and Data Communications Handbook*, John Wiley & Sons, Inc., New York, NY, 2007.

Horak, Ray, *Webster's New World Telecom Dictionary*, Wiley Publishing, Inc., Indianapolis, IN, 2008.

Hutchinson, Chuck, *The ARRL Handbook For Radio Amateurs*, 2001 edition, The American Radio Relay League, CT, 2000.

"Industrial Electronics Reference Book", Westinghouse Electric Corporation, John Wiley & Sons Inc., New York, NY, 1948.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Software and Digital Networks*, Third Edition, CRC Press, New York, NY, 2002.

"Modbus Application Protocol Specification", version 1.1b, Modbus-IDA, Modbus Organization, Inc., 2006.

"Modbus Messaging on TCP/IP Implementation Guide", version 1.0b, Modbus-IDA, Modbus Organization, Inc., 2006.

"Modicon Modbus Protocol Reference Guide", (PI-MBUS-300) revision J, Modicon, Inc. Industrial Automation Systems, North Andover, MA, 1996.

Newton, Harry, *Newton's Telecom Dictionary*, CMP Books, San Francisco, CA, 2005.

Park, John; Mackay, Steve; Wright, Edwin; *Practical Data Communications for Instrumentation and Control*, IDC Technologies, published by Newnes (an imprint of Elsevier), Oxford, England, 2003.

"Selecting and Using RS-232, RS-422, and RS-485 Serial Data Standards" Application Note 723 (AN723), Maxim Integrated Products, 2000.

Smith, Steven W., *The Scientist and Engineer's Guide to Digital Signal Processing*, California Technical Publishing, San Diego, CA, 1997.

Smith, W. W., *The "Radio" Handbook*, Sixth Edition, Radio Ltd., Santa Barbara, CA, 1939.

Spurgeon, Charles E., *Ethernet: The Definitive Guide*, O'Reilly Media, Inc., Sebastopol, CA, 2000.

Svacina, Bob, *Understanding Device Level Buses: A Tutorial*, *Interlink*BT, LLC, Minneapolis, MN, 1998.

Welsh, Matt and Kaufman, Lar, *Running Linux*, Second Edition, O'Reilly & Associates, Sebastopol, CA, 1996.

Chapter 16

FOUNDATION Fieldbus instrumentation

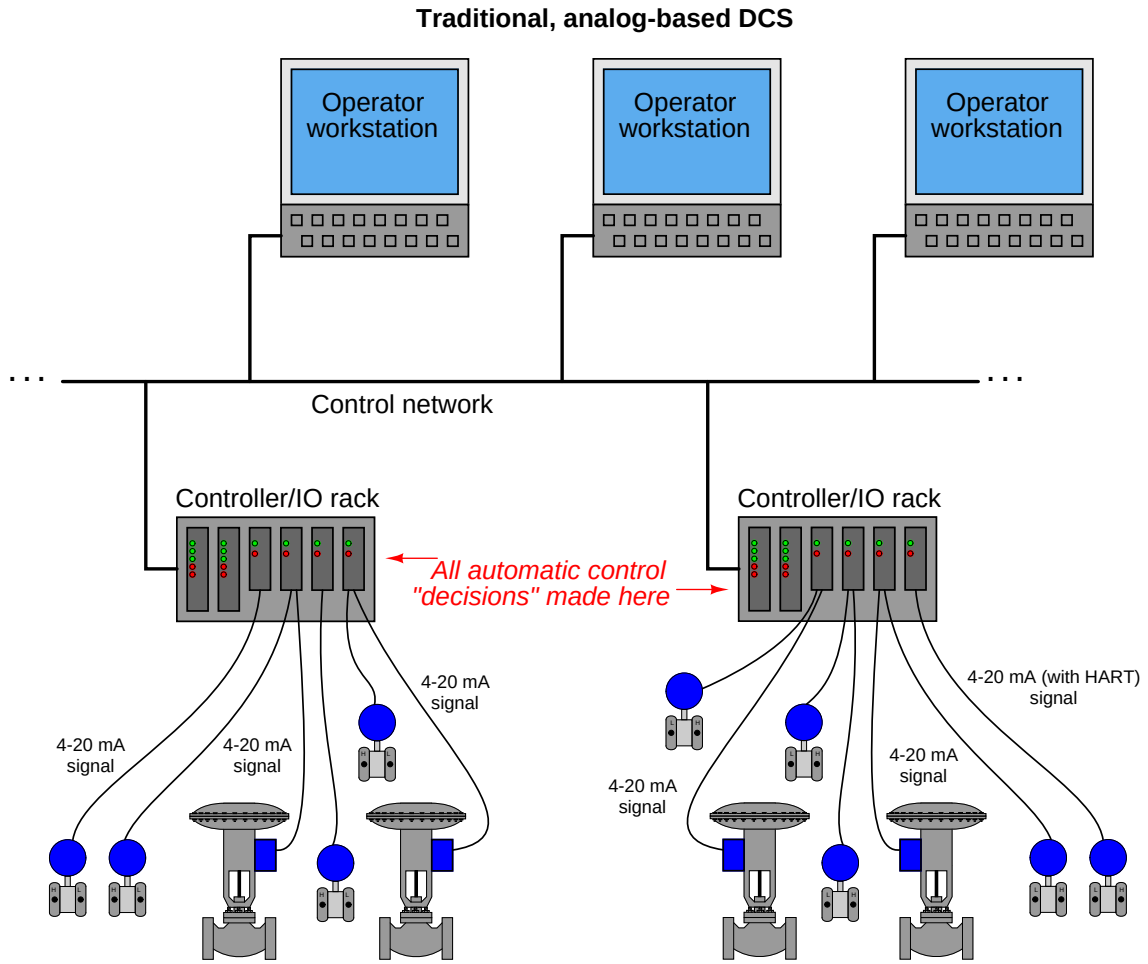
FOUNDATION Fieldbus is a standard for digital field instrumentation enabling field instruments to not only communicate with each other digitally, but also to execute all continuous control algorithms (such as PID, ratio control, cascade control, feedforward control, etc.) traditionally implemented in dedicated control devices. In essence, FOUNDATION Fieldbus extends the general concept of a distributed control system (DCS) all the way to the field devices themselves. In this way, FOUNDATION Fieldbus sets itself apart as more than just another digital communication "bus" for industry – it truly represents a new way to implement measurement and control systems. This chapter is devoted to a discussion of FOUNDATION Fieldbus instrumentation, building on general concepts of digital data acquisition and communication previously explored in this book.

For brevity, "FOUNDATION Fieldbus" will be abbreviated as *FF* throughout the rest of this chapter.

This particular industrial network standard was first proposed as a concept in 1984, and officially standardized by the Fieldbus Foundation (the organization overseeing all FF standards and validation) in 1996. To date, adoption of FF has been somewhat slow, mostly limited to new construction projects. One of the "selling points" of FF is decreased installation time, which makes it a more attractive technology for brand-new installations than for retrofit projects.

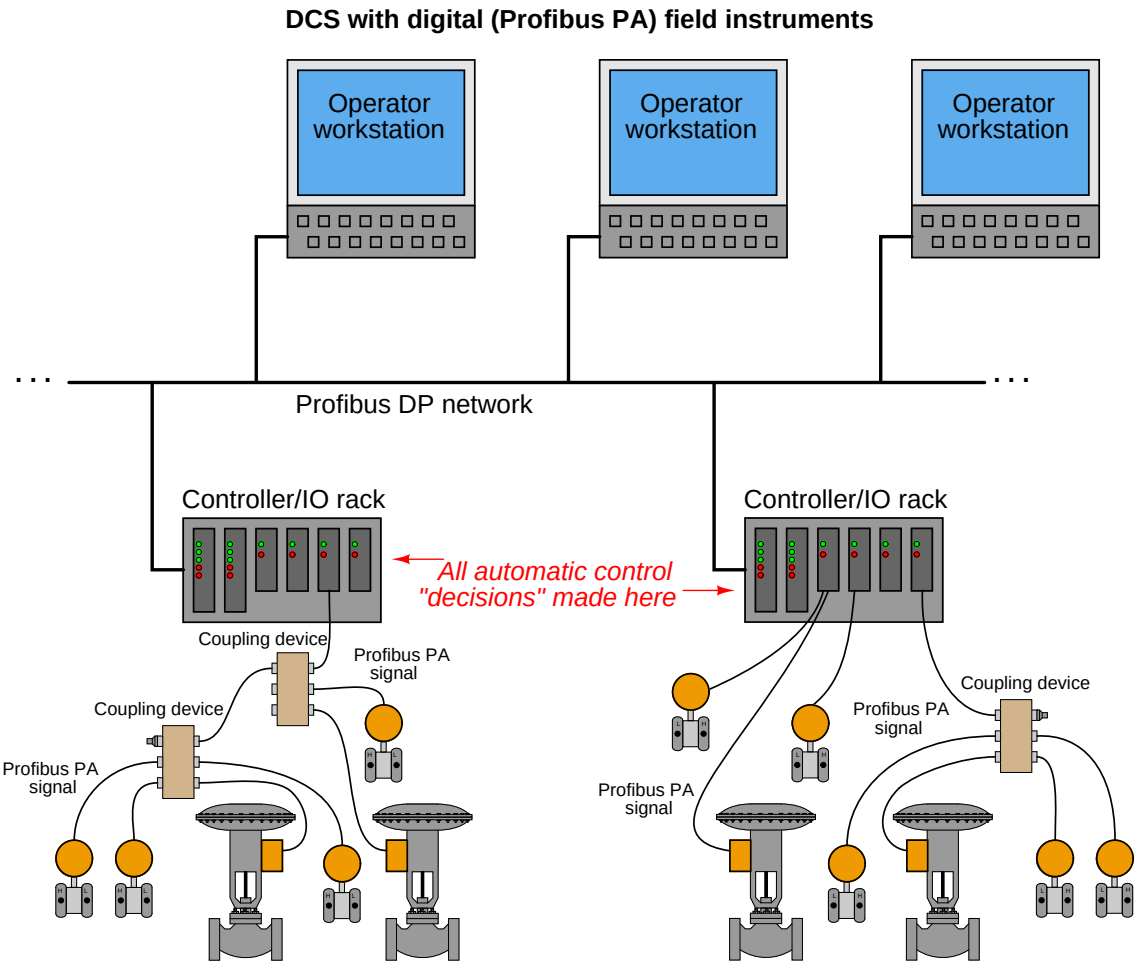
16.1 FF design philosophy

To understand just how different FF is from other digital instrument systems, consider a typical layout for a distributed control system (DCS), where all the calculations and logical "decisions" are made in dedicated *controllers*, usually taking the form of a multi-card "rack" with processor(s), analog input cards, analog output cards, and other types of I/O (input/output) cards:



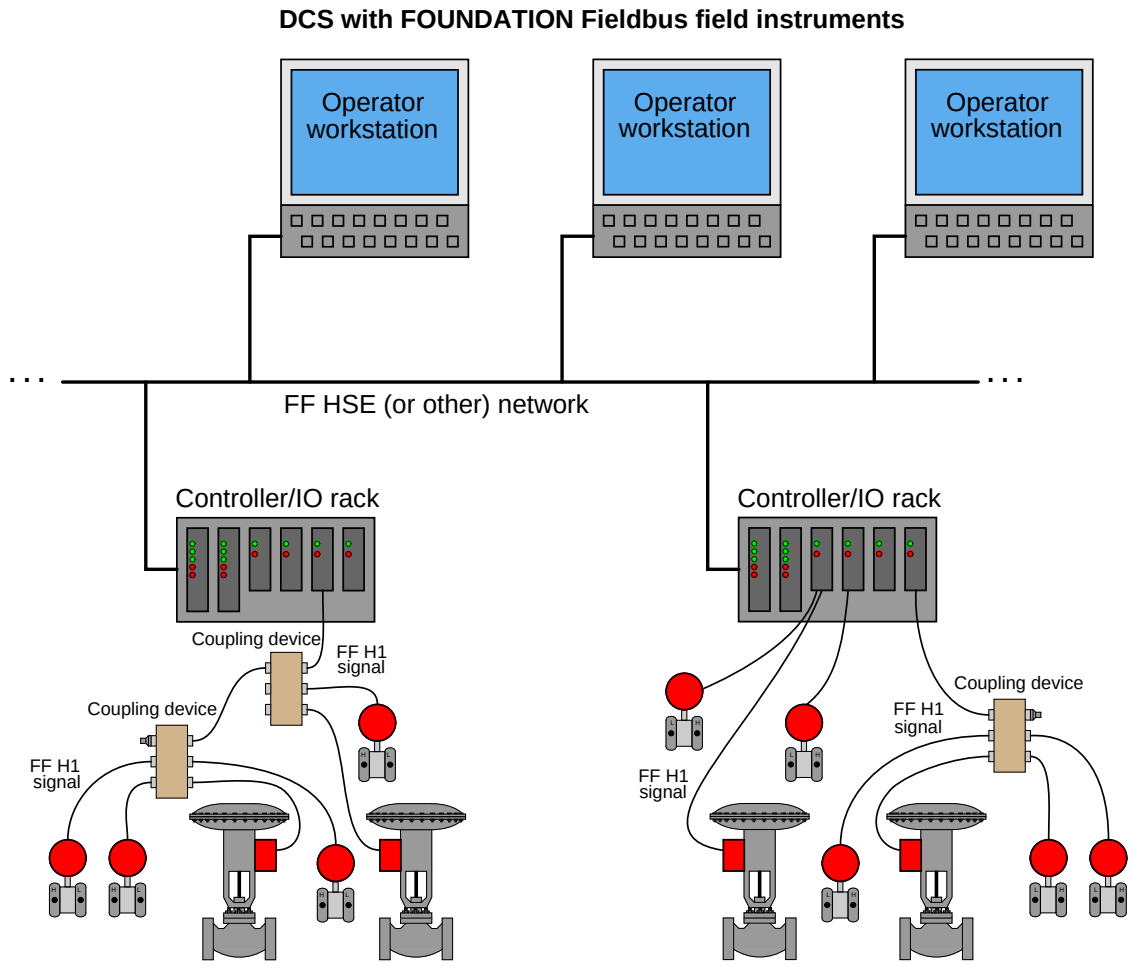
Information is communicated in analog form between the DCS controllers and the field instruments. If equipped with the proper types of I/O cards, the DCS may even communicate digitally with some of the field instruments using *HART* protocol. This allows multivariable instruments to communicate multiple variables to and from the DCS controllers (albeit slowly) over a single wire pair.

It is even possible to build a control system around a DCS using all digital field instruments, using a protocol such as *Profibus PA* to exchange process variable (PV) and manipulated variable (MV) signals to and from the DCS controllers:



Now, multivariable field instruments have the ability to quickly exchange their data with the DCS, along with maintenance-related information (calibration ranges, error messages, and alarms). Each "fieldbus" cable is a (potential) two-way path for digital information flow. Field wiring is reduced in cable length and connection count due to the use of *coupling devices* to connect multiple instruments to single "home run" network cables leading to the DCS. Still, however, all the automatic control algorithms are implemented in the DCS.

An FF system, by contrast, allows the embedding of all control algorithms within the field instruments rather than relying on the DCS controllers to execute automatic "decisions." In fact, the DCS would not even be necessary if not for the need of operations personnel to monitor and alter control system status:



That being said, it is possible (and in fact common) for control algorithms to be placed in the DCS controllers in addition to algorithms executed by FF field devices.

When the FF standard was being designed, two different network levels were planned: a "low speed" network for the connection of field instruments to each other to form network *segments*, and a "high speed" network for use as a plant-wide "backbone" for conveying large amounts of process data over longer distances. The low-speed (field) network was designated *H1*, while the high-speed (plant) network was designated *H2*. Later in the FF standard development process, it was realized that

existing Ethernet technology would address all the basic requirements of a high-speed "backbone," and so it was decided to abandon work on the H2 standard, settling on an extension of 100 Mbps Ethernet called *HSE* ("High Speed Ethernet") as the backbone FF network instead.

The bulk of this chapter will focus on H1 rather than HSE.

16.2 H1 FF Physical layer

Layer 1 of the OSI Reference Model is where we define the "physical" elements of a digital data network. The H1 FF network exhibits the following properties:

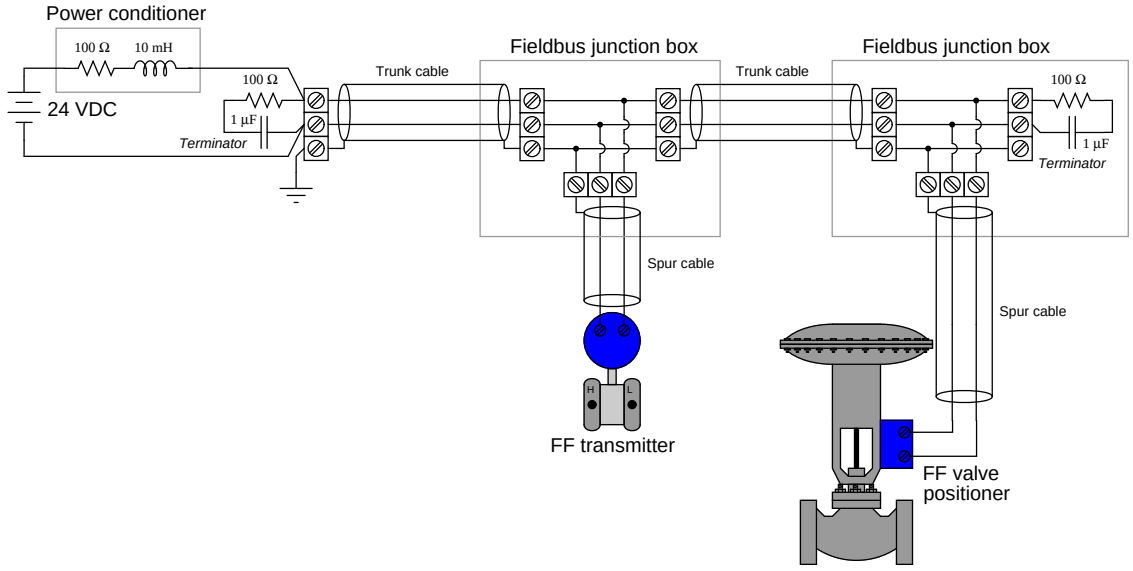
- Two-wire (ungrounded) network cable
- 100 ohm (nominal) characteristic impedance
- DC power is conveyed over the same two wires as digital data
- 31.25 kbps data rate
- Differential voltage signaling (0.75 volts peak-to-peak transmit minimum ; 0.15 volts peak-to-peak receive threshold minimum)
- Manchester encoding

Since DC power is conveyed over the same two wires as the digital data, it means each device only needs to connect to two wires in order to function on an H1 network segment. The choice of a (relatively) slow 31.25 kbps data rate allows for imperfect cables and terminations which would otherwise plague a faster network. Manchester encoding embeds the network clock pulse along with the digital data, simplifying synchronization between devices.

As you can see, the layer 1 design parameters were chosen to make FF H1 networks easy to build in unforgiving industrial environments. The physical layer of FOUNDATION Fieldbus happens to be identical to that of Profibus-PA, further simplifying installation by allowing the use of certain network validation tools and connection hardware developed for this other network.

16.2.1 Segment topology

A minimal FF H1 segment consists of a DC power supply, a "power conditioner," exactly two terminator resistors[1] (one at each extreme end of the cable), a shielded and twisted-pair cable, and of course at least two FF instruments to communicate with each other. The cable connecting each instrument to the nearest junction is called a *spur* (or sometimes a *stub* or a *drop*), while the cable connecting all junctions to the main power source (where a host DCS would typically be located) is called a *trunk* (or sometimes a *home run* for the section leading directly to a host system):

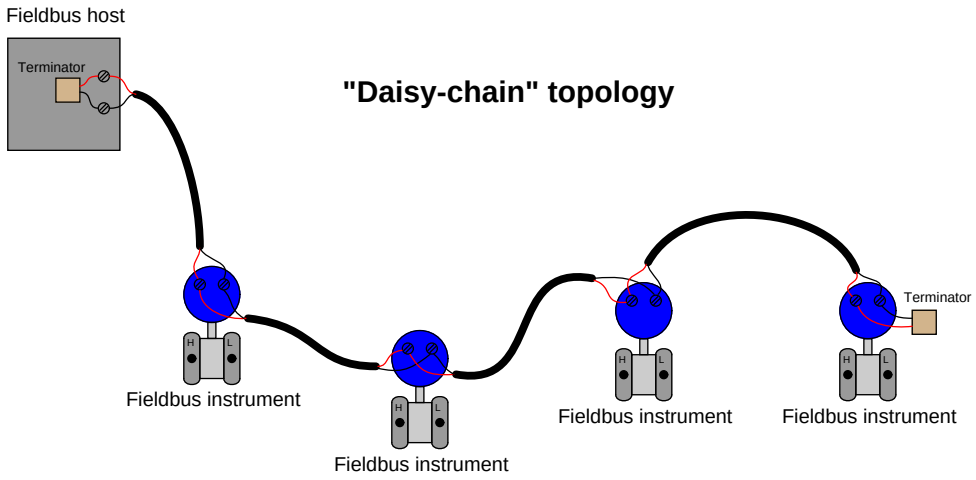


Normally, we would find more than two FF devices connected to a trunk cable, as well as a "host" system such as a DCS FF card for presenting data from the FF instruments, performing maintenance tasks, and integrating with other control loops. Regardless of how many (or how few) FF devices connect to an H1 segment, though, there should always be *exactly two* terminating resistors in each segment – one at each end[2] of the trunk cable. These resistor/capacitor networks serve the sole purpose of eliminating signal reflections off the ends of the trunk cable, making the cable look infinitely long from the perspective of the propagating pulse signals. Missing terminators will result in signal reflections off the unterminated line end(s), while extra terminators have the equally deleterious effect of attenuating signal strength (as well as potentially causing signal reflections of opposite phase).

[1] Each FF terminator resistor is actually a series resistor/capacitor network. The resistor blocks direct current, so that the 100 Ω resistor does not present a DC load to the system.

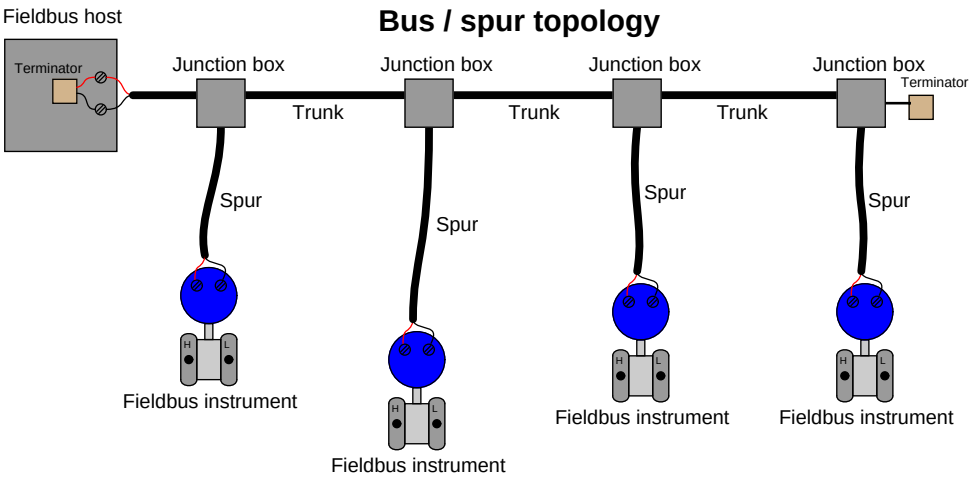
[2] Be sure to check the specifications of the host system H1 interface card, because many are equipped with internal terminating resistors given the expectation that the host system will connect to one far end of the trunk!

All H1 networks are essentially parallel electrical circuits, where the two connection terminals of each field instrument are paralleled to each other. The physical arrangement of these transmitters, though, may vary substantially. The simplest way to connect FF H1 devices together is the so-called "daisy-chain" method, where each instrument connects to two cable lengths, forming an uninterrupted "chain" network from one end of the segment to the other:



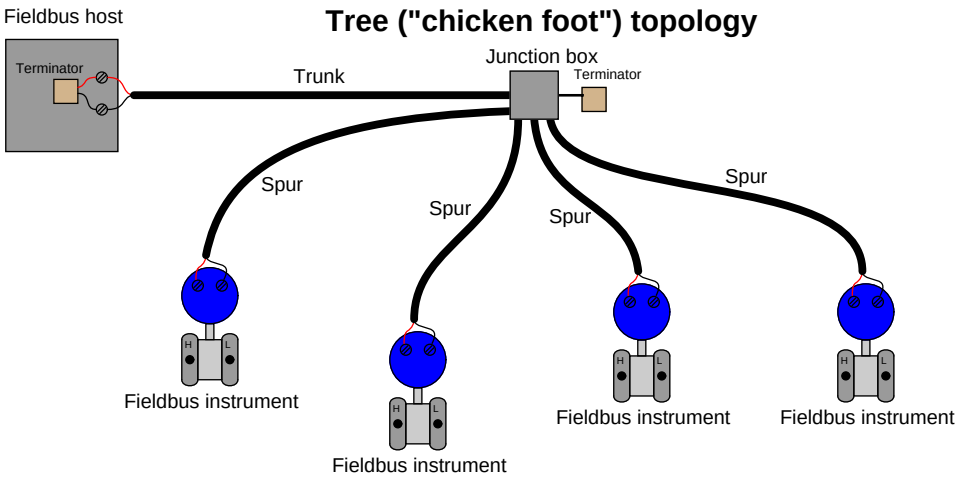
As simple as this topology is, it suffers from a major disadvantage: it is impossible to disconnect any device in the segment without interrupting the network's continuity. Disconnecting (and reconnecting for that matter) any device necessarily results in all "downstream" devices losing signal, if only for a brief time. This is an unacceptable state of affairs for most applications.

An alternative topology is the *bus* layout, where short "spur" cables connect instruments to a longer "trunk" cable. Terminal blocks – or even quick-disconnect couplings – within each junction box provide a convenient means of disconnecting individual devices from the segment without interrupting data communication with the other devices:



The ideal arrangement for a "bus" network is to minimize the length of each spur cable, so as to minimize the delay of reflected signals off the unterminated ends of the drops. Remember that only *two* termination resistors are allowed in any electrically continuous network segment, and so this rule forbids the addition of terminators to the end of each spur cable.

Yet another alternative topology for H1 networks is the so-called *chicken-foot* arrangement, where a long trunk cable terminates at a multi-point junction along with several field devices and their spur cables:



Most FF systems resemble a combination of "bus" and "chicken-foot" topologies, where multiple junction devices serve as connection points for two or more field instruments per junction.

16.2.2 Coupling devices

In order to simplify the task of connecting Fieldbus devices to such a network segment, multiple manufacturers sell *coupling devices* (often informally referred to as *bricks*) with quick-disconnect electrical fittings so the end-user does not have to build and commission junction boxes using standard terminal blocks. A photograph of a Turck brand Fieldbus coupling device appears here, showing multiple spur cables plugged into it:

Coupling devices are highly recommended for all industrial fieldbus systems, FF or otherwise. Not only do these devices provide a convenient means of forming highly reliable connections between field instruments and the trunk cable, but many of them are equipped with features such as short-circuit protection (so that a shorted spur cable or field instrument does not cause the entire segment to stop communicating) and LED indication of spur status.

Cables connecting to a coupling device must be equipped with special plugs matching the sockets on the coupler. This presents a bit of a problem when attempting to pull such a cable through electrical conduit: the bulky plug requires either over-sized conduit to accommodate the plug's width, or requires the plug be installed on the cable after pulling through the conduit. Both approaches are expensive, the first in terms of capital cost and the second in terms of installation labor. For this reason, many installers abandon electrical conduit altogether in favor of *ITC* ("Instrument Tray Cable").

A wider-angle photograph of the coupling device previously shown reveals many ITC cables and their routing through wire "basket" style trays among process instruments and vessels:

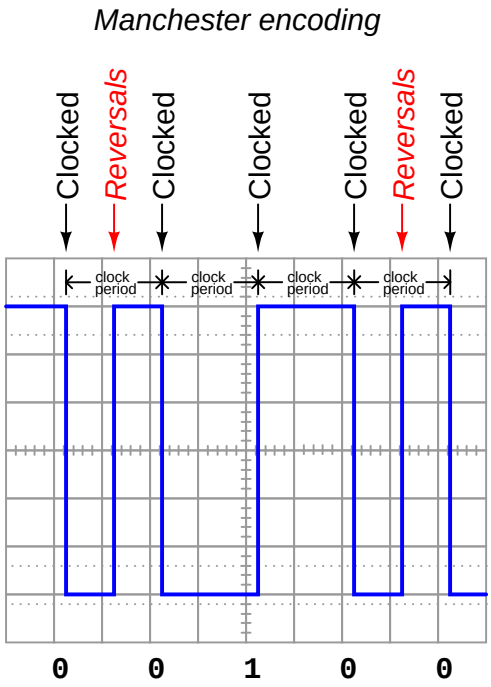
As evident in this photograph, ITC is obviously rated for continuous exposure to direct sunlight and moisture, as well as a certain amount of physical distress (abrasion, high and low temperatures, etc.). Article 727 of the National Electrical Code (NEC) defines the acceptable uses and installations of ITC[3].

It should be noted that while a properly shielded and grounded FF cable is quite resistant to radio-frequency interference, coupling devices may present "weak spots" where radio interference may find its way onto the segment. Different styles of coupling devices offer differing levels of immunity to RF (Radio Frequency) noise. Those made of metal and properly bonded to ground will be well-shielded, while those made of plastic having exposed connection terminals offer little or no protection. In any case, it is a good practice to avoid "keying" any portable radio transmitter in the near vicinity of a Fieldbus coupling device.

[3]You should consult an NEC code book regarding specific limitations of ITC wiring. Some of the main points include limiting individual ITC cable lengths to a maximum of 50 feet, and mechanically securing the cable at intervals not to exceed 6 feet.

16.2.3 Electrical parameters

FOUNDATION Fieldbus H1 networks use Manchester encoding to represent bit states: a "high-to-low" transition represents a logical zero (0), while a "low-to-high" transition represents a logical one (1). The following illustration shows how the data stream `00100` would be represented in Manchester encoding:



FF devices must be able to correctly distinguish between rising- and fall-edge signals in order to properly interpret the bit states of a Manchester-encoded signal. Any device interpreting these pulse edges "backwards" will invert every single bit! Thankfully, this problem is easy to avoid because the DC power supplied by the H1 segment wiring provides a "key" to identifying which wire is which, and therefore which pulses are rising-edge versus which pulses are falling-edge. For this reason, many (but not all!) FF devices are polarity-insensitive, automatically detecting the polarity of the network segment and compensating accordingly.

Every FF device draws at least 10 mA of current from the segment, and this current does not vary in the same manner that an analog (4-20 mA) device draws differing amounts of current under different operating conditions. Always remember that a Fieldbus device signals its variable(s) digitally, not by varying current. Old habits (and thought patterns) die hard, and so Fieldbus systems present challenges to technicians familiar with the behavior of analog current loop instrumentation. The amount of current drawn by any particular FF device depends on that device's functionality – obviously, some will require more current[4] for their operation than others. 10 mA to 30 mA should

[4] Perusing documentation on an assortment of Emerson/Rosemount FF products, I found the following data: model 752 indicator = 17.5 mA, model 848L logic = 22 mA, model 848T temperature = 22 mA maximum, model 3244MV temperature = 17.5 mA nominal, model DVC6000f valve positioner = 18 mA maximum, model 848L logic = 22

be considered a general range of current drawn by each FF device.

The standard operating voltage range for FF devices is between 9 and 32 volts DC. It is important to note, however, that not all manufacturers' devices are in full compliance with the Fieldbus Foundation standard, and as such some may not operate properly at low voltages (near 9 volts DC)!

The minimum transmission voltage of a FF device is 750 millivolts peak-to-peak, while the minimum signal level for reception by a FF device is 150 millivolts peak-to-peak. This represents an acceptable attenuation of 5:1, or -14 dB between any two devices.

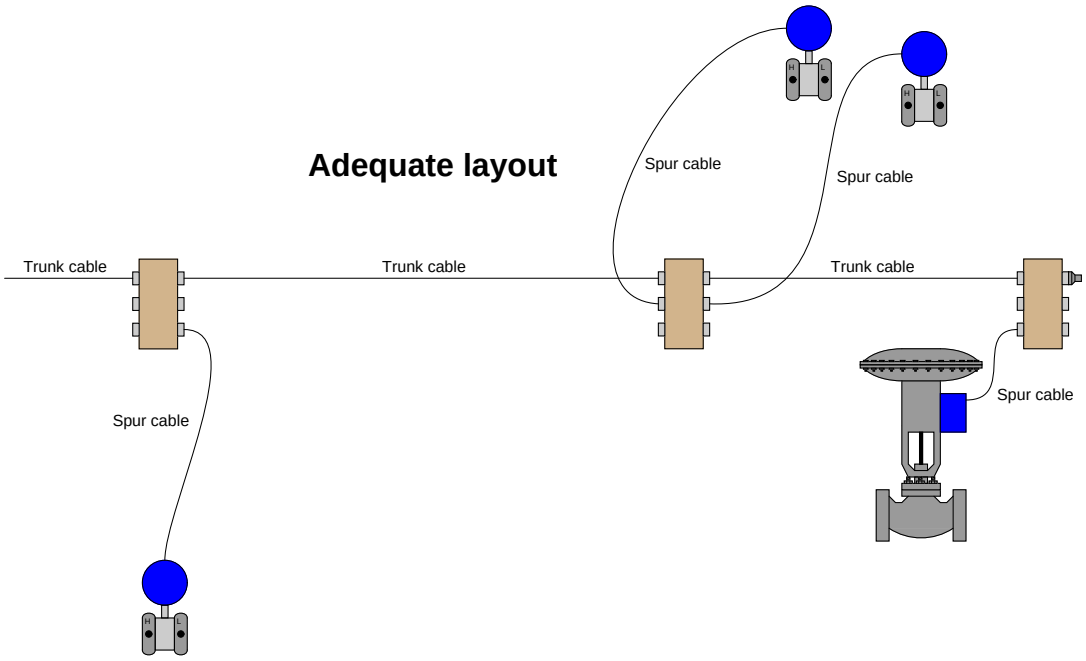
mA, model 848T temperature = 22 mA maximum, model 3244MV temperature = 17.5 mA nominal, model 5500 guided-wave radar level = 21 mA, model 3095MV flow (differential pressure) = 17 mA approximate, model DVC6000f valve positioner = 18 mA maximum.

16.2.4 Cable types

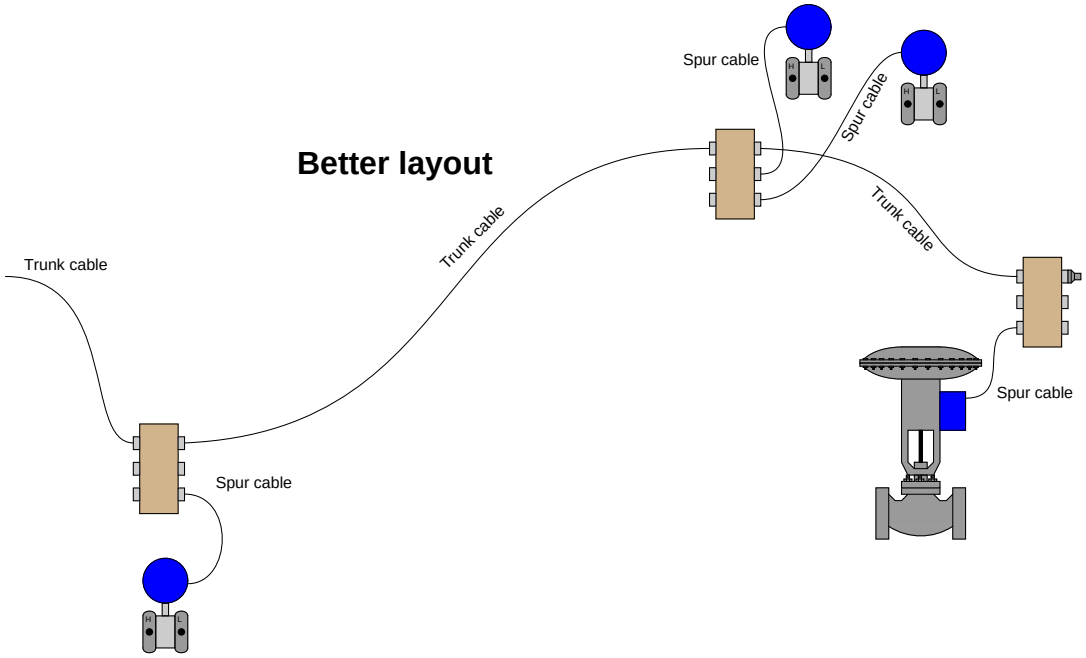
Fieldbus cable is rated according to a four-level code (A, B, C, or D), each successive letter representing a cable of lower quality[5]. The following table gives *minimum specifications* for each FF cable type:

Cable Type	Type A	Type B	Type C	Type D
Wire size	AWG 18	AWG 22	AWG 26	AWG 16
Char. Impedance	100 Ω ± 20%	100 Ω ± 30%	–	–
Shielding	1 for each pair	1 for entire cable	none	none
Twisted pairs	Yes	Yes	Yes	No
Max. length	1900 m	1200 m	400 m	200 m

Bear in mind that the maximum length given for each cable type is the *total length* of all cables in a segment, trunk length plus all spur lengths. As a general rule, spur lengths should be kept as short as possible. It is better to route the trunk cable in a serpentine fashion to locate coupling devices close to their respective instruments than it is to streamline the trunk cable routing. The following illustrations contrast the two approaches:



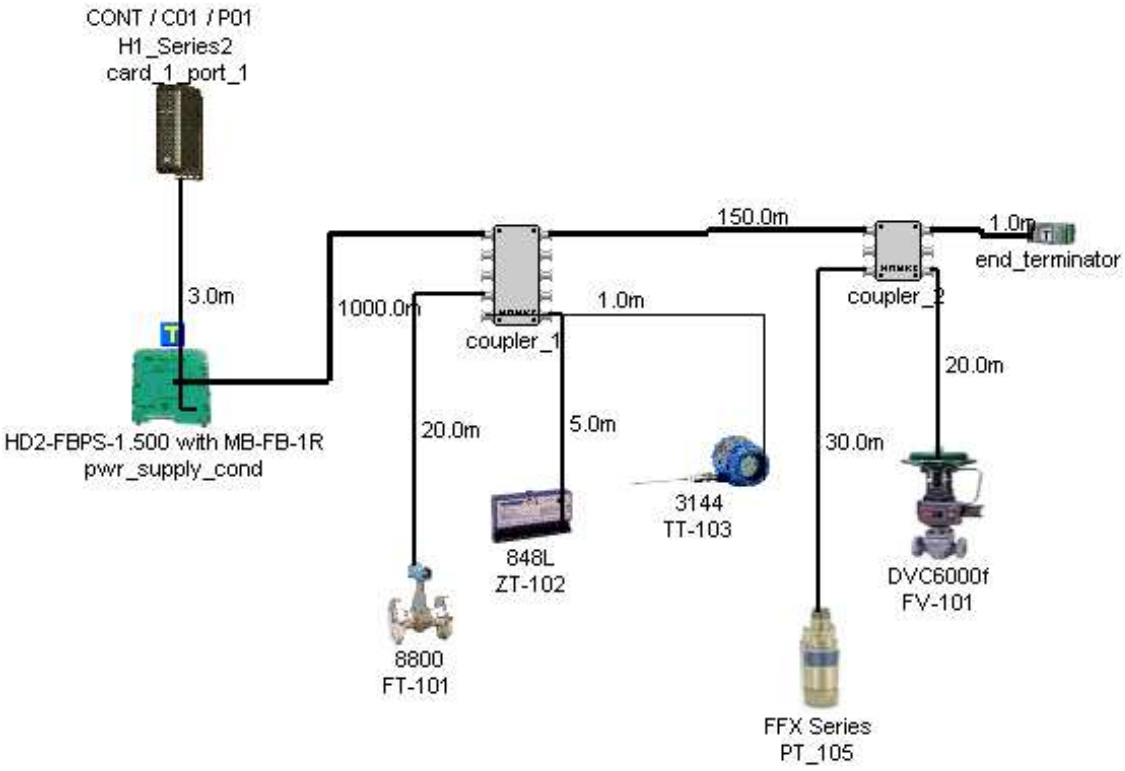
[5] I have successfully built several "demonstration" FF systems using cables of questionable quality, including lamp ("zip") cord, with no termination resistors whatsoever! If the distances involved are short, just about any cable type or condition will suffice. When planning the installation of any real Fieldbus installation, however, you should never attempt to save money by purchasing lesser-grade cable. The problems you will likely encounter as a consequence of using sub-standard cable will more than offset the initial cost saved by its purchase.



If greater lengths are required for a network segment, devices known as *repeaters* may be added which sense and re-broadcast the Manchester-encoded FF signal between trunk cables. A maximum of four repeaters may be used to extend any H1 segment.

16.2.5 Segment design

In addition to maximum (total) cable length and repeater count, a host of other details[6] conspire to limit how any particular H1 segment is wired. To help engineers and technicians alike deal with these details, manufacturers often provide free *segment design tool* software to pre-validate a segment design on computer before purchasing components and installing them in the field. A screenshot taken from Emerson's offering shows what a typical FF segment layout might look like:



A very nice feature of these segment design packages is their built-in database of FF components. Every time you "pick" a particular component to place in your simulated segment, the program references data for that device's current draw and other electrical parameters relevant to the performance of the segment. Of course, each manufacturer will tend to feature their own devices more prominently, and so these software tools sometimes have the flavor of a promotional advertisement. Despite the commercial aspect of their design, however, they are extremely useful in the planning stages of a FF network, and should be used whenever possible.

Another reason to use segment design tool software is to document the wiring of each FF segment. One of the casualties of the new Fieldbus paradigm is the traditional *loop diagram* (or "loop sheet"), the purpose of which is to document the signal wiring dedicated for each measurement and control

[6] Total device current draw, spur length versus number, intrinsic safety voltage and current limitations, etc.

loop. In FOUNDATION Fieldbus, the control "loop" is virtual rather than physical, being comprised of digital data sent between field instruments, the path of which being defined by the instruments' programming. The only physical wiring entity to document in a FF system is the segment, and each segment most likely hosts more than one measurement and/or control loop. Unless and until a standardized documentation format[7] is invented for Fieldbus network segments, the graphic image provided by segment design tool software is as good as anything.

[7] At the time of this writing (2009), the ISA has yet to standardize new methods of FF documentation in the style of loop sheets and P&IDs. This is one of those circumstances where technology has outpaced convention.

16.3 H1 FF Data Link Layer

Layer 2 of the OSI Reference Model is where we define the "data link" elements of a digital data network. The H1 FF network exhibits the following properties:

- Master/slave network behavior for cyclic communications (i.e. one device polls the others, and the others merely respond)
- Delegated token network behavior for acyclic communications (i.e. devices serially granted time to broadcast at will)
- Dedicated "scheduler" device for coordinating all segment communications
- 8-bit address field (0 through 255 possible)
- Maximum of 32 "live" devices on a segment

On an operating H1 segment, one device called the *Link Active Scheduler* (abbreviated LAS) functions as the "master" device for coordinating all network communications. For time-critical transmissions, the LAS polls the various field instruments to transmit their process control data (process variables, PID control output values, and other variables essential for loop monitoring and control), while these other devices respond in answer to the LAS's queries. These critical communications occur on a regular schedule, and therefore are referred to as *scheduled* or *cyclic* communications. Cyclic communication operates in a "master-slave" fashion, with the LAS acting as the master (commanding slave devices to broadcast their critical data), and all other devices responding only when called upon by the LAS.

Periods of time in between these critical transmissions are used for device's internal processing (e.g. PID algorithm execution, diagnostic checking) and also for less-critical data transmission. It is during these *unscheduled* or *acyclic* times that devices are sequentially given permission by the LAS to broadcast data of less importance such as operator setpoints, PID tuning constant updates, alarm acknowledgments, and diagnostic messages. Acyclic communication operates in a manner similar to "token-passing," with the LAS issuing time-limited tokens to the other devices in sequence permitting them to freely broadcast whatever other data they have to share.

The scheduled nature of cyclic communication guarantees a certain maximum response time to critical control functions, an important property of control networks called *determinism*. Without determinism, a control system cannot be relied upon to perform critical regulatory functions in a timely[8] manner, and sequencing[9] of control functions such as PID, summers, subtractors, ratio multipliers, and the like may be compromised.

[8] While many industrial control systems have been built using networks that are not strictly deterministic (e.g. Ethernet), generally good control behavior will result if the network latency time is arbitrarily short. Lack of "hard" determinism is more of a problem in safety shutdown systems where the system *must* respond within a certain amount of time in order to be effective in its safety function.

[9] By "sequencing," I mean the execution of all antecedent control functions prior to "downstream" functions requiring the processed data. If in a chain of function blocks we have some blocks lagging in their execution, other blocks relying on the output signals of those lagging blocks will be functioning on "old" data. This effectively adds dead time to the control system as a whole. The more antecedent blocks in the chain that lag in time behind the needs of their consequent blocks, the more dead time will be present in the entire system. To illustrate, if block A feeds data into block B which feeds data into block C, but the blocks are executed in reverse order (C, then B, then A) on the same period, a lag time of *three whole execution periods* will be manifest by the A-B-C algorithm.

16.3.1 Device addressing

FOUNDATION Fieldbus devices (also called *nodes*) are addressed by an eight-bit binary number when functioning on an H1 segment. This binary number field naturally supports a maximum addressing range of 0 to 255 (decimal), or 00 to FF hexadecimal. This address range is divided into the following sub-ranges by the Fieldbus Foundation:

Address range (decimal)	**Address range** (hexadecimal)	**Allocation**
0 through 15	00 through 0F	Reserved
16 through 247	10 through F7	Permanent devices
248 through 251	F8 through FB	New or decommissioned devices
252 through 255	FC through FF	Temporary ("visitor") devices

Devices are usually assigned addresses to function on the segment by the host system (typically a DCS with FF capability), although it is possible to order FF instruments pre-configured at the factory with addresses specified by the customer upon order. Host systems are generally configured to automatically determine device addresses rather than require the technician or engineer to manually assign each address. This make the commissioning process more convenient.

The maximum number of "permanent" devices (installed field instruments) allowed on an H1 segment for operational reasons is 32, and as you can see the addressing scheme offers far more valid addresses than that. One of the many tasks given to a segment's Link Active Scheduler (LAS) device is to probe for new devices connected to the segment. This is done on a one-at-a-time basis, with the LAS sequentially polling for uncommissioned addresses within the valid address range. Obviously, this can be a waste of time with only 32 addresses capable of active service at any given time and over 200 valid address numbers. A practical solution to this problem is to specify an "unused" address range for the LAS to skip, so it does not waste time probing for devices (nodes) within a certain range. This address range is specified as a set of two numbers: one for the First Unused Node (abbreviated *FUN*), and another specifying the Number of Unused Nodes (abbreviated *NUN*). For example, if one wished to have the LAS on a particular H1 segment skip device addresses 40 through 211, one would configure the FUN to equal 40 and the NUN to equal 172, since the address range 40 through 211 is one hundred seventy two addresses (inclusive of both 40 and 211).

Even with a maximum operational limit of 32 devices to an H1 segment, it is rare to find segments operating with more than 16 devices. One reason for this is speed: with additional devices requiring time to broadcast and process data, the total *macrocycle* time (the time period between guaranteed delivery of the same process data from any one device – the determinism time) must necessarily increase. According to the Fieldbus Foundation's engineering recommendations guide, there must be no more than twelve devices on a segment (including no more than two final control elements) in order to achieve a 1-second or less macrocycle time. For half-second update times, the recommended maximum is six devices (with no more than two final control elements). For quarter-second update times, the limit drops to a total of three devices, with no more than one final control element. Macrocycle time is essentially dead time, which is worse than lag time for any form of feedback control. When controlling certain fast processes (such as liquid pressure or flow rate), dead times on the order of one second are a recipe for instability.

Another limitation to the number of operational addresses on an H1 segment is current draw. FF devices draw 10 mA of current *minimum*. A FF segment with sixteen parallel-connected devices

would see a total current of 160 mA minimum, with a more realistic value being in excess of 300 mA.

In addition to network addresses, each FF device bears an absolutely unique identifier (a 32-byte binary number) to distinguish it from any other FF device in existence. This identifier serves much the same purpose as a *MAC address* on an Ethernet device. However, the identifier field for FF devices allows a far greater instrument count than Ethernet: 32 *bytes* for FF instruments versus 48 bits for Ethernet devices. While the Ethernet MAC address field only allows for a paltry 2.815×10^{14} unique devices, the FF identifier allows 1.158×10^{77} devices! The distinction between a FF device's network address and the device's identifier is virtually identical to the distinction between an Ethernet device's IP address assigned by the end-user and its MAC address number assigned by the manufacturer.

This identifier value is usually expressed as 32 ASCII-encoded characters for brevity (one alphanumeric character per byte), and is subdivided into byte groups as follows:

First 6 bytes	Middle 4 bytes	Last 22 bytes
Manufacturer code	Device type code	Serial number

For example, the identifiers for all *Fisher* brand devices begin with the first six characters `005100`. The identifiers for all *Smar* devices begin with the characters `000302`. The identifiers for all *Rosemount*[10] brand devices begin with `001151`. A typical identifier (this particular one for a Fisher model DVC5000f valve positioner) appears here:

```
005100 0100 FISHERDVC0440761498160
```

Normally, these identifiers appear as 32-character strings, without spaces at all. I have inserted spaces within this string to make the character groupings easier to see.

16.3.2 Communication management

In a FF network segment, the Link Active Scheduler (LAS) device coordinates all communications between segment devices. Among the many responsibilities the LAS is tasked with are the following:

- Commands non-LAS devices to broadcast data to the segment with "Compel Data" (CD) messages, issued at regular time intervals to specific devices (one at a time)
- Grants permission for non-LAS devices to communicate with "Pass Token" (PT) messages, issued during unscheduled time slots to specific devices (one at a time, in ascending order of address number)
- Keeps all segment devices synchronized with a regular "Time Distribution" (TD) message
- Probes for new devices on the segment with a "Probe Node" (PN) message
- Maintains and publishes a list of all active devices on the network (the *Live List*)

[10] The engineers there are not without a sense of humor, choosing for their manufacturer code the same model number as the venerable 1151 differential pressure transmitter, perhaps the most popular Rosemount industrial instrument in the company's history!

Scheduled versus unscheduled communication

As previously mentioned, Fieldbus H1 network communication may be divided into two broad categories: *scheduled* (cyclic) and *unscheduled* (acyclic). Scheduled communication events are reserved for exchanging critical control data such as process variable measurements, cascaded setpoints, and valve position commands. These scheduled communications happen on a regular, timed schedule so that loop determinism is guaranteed. Unscheduled communications, by contrast, are the way in which all other data is communicated along an H1 segment. Manual setpoint changes, configuration updates, alarms, and other data transfers of lesser importance are exchanged between devices in the times between scheduled communication events.

Both forms of communication are orchestrated by the Link Active Scheduler (LAS) device, of which there is but one active at any given time[11] on an H1 segment. The LAS issues "token" messages to non-LAS devices commanding (or merely authorizing) them to broadcast to the segment one at a time. Each token message issued by the LAS grants transmission rights to an FF device either for a limited purpose (i.e. the precise message to be transmitted) or for a limited time (i.e. giving that device the freedom to transmit whatever data it desires for a short duration), after which transmission rights return to the LAS. CD tokens are message-specific: each one issued by the LAS commands a single device to immediately respond with a broadcast of some specific data. This is how scheduled (cyclic) communication is managed. PT tokens are time-specific: each one issued by the LAS grants a single device free time to transmit data of lesser importance. This is how unscheduled (acyclic) communication between devices is managed.

The LAS also issues at third type of token message: the "Probe Node" (PN) token intended to elicit a response from any new devices connected to the network segment.

In addition to transmitting tokens – which by definition are messages granting another device permission to transmit to the network – the LAS also broadcasts other messages necessary for the function of an H1 segment. For example, the "Time Distribution" (TD) message regularly broadcast by the LAS keeps all devices' internal clocks synchronized, which is important for the coordinated transfer of data.

One of the "internal" tasks of the LAS (not requiring network broadcasts) is the maintenance of the *Live List*, which is a list of all known devices functioning on the network segment. New devices responding to "Probe Node" messages will be added to the Live List when detected. Devices failing to return or use PT tokens issued to them are removed from the Live List after a number of attempts. When "backup" LAS devices exist on the segment, the LAS also publishes updated copies of the Live List to them, so they will have the most up-to-date version should the need arise to take over for the original LAS (in the event of an LAS device failure).

In "busy" H1 segments where multiple devices are exchanging data with each other, a heavy traffic load of scheduled communications (CD tokens and their responses) makes it difficult for substantial unscheduled (acyclic) data exchanges to occur. For example, if a device happens to be maintaining a lengthy list of client/server requests in its queue, which it may address only during its allotted acyclic time slots (i.e. when it has been given the PT token from the LAS), it is quite possible the PT token will expire before all the device's transactions have been completed. This means the device will have to wait for the next acyclic period before it can complete all the unscheduled communication tasks

[11] In addition to the main LAS, there may be "backup" LAS devices waiting ready to take over in the event the main LAS fails for any reason. These are Link Master devices configured to act as redundant Link Active Schedulers should the need arise. However, at any given time there will be only *one* LAS.

in its queue. The Fieldbus Foundation recommends new H1 segments be configured for no more than 30% scheduled communications during each macrocycle (70% unscheduled time). This should leave plenty of "free time" for all necessary acyclic communications to take place without having to routinely wait multiple macrocycles.

Virtual Communication Relationships

A term you will frequently encounter in FF literature is *VCR*, or "Virtual Communication Relationships." There are three different types of VCRs in FF, describing three different ways in which data is communicated between FF devices:

- Publisher / Subscriber (scheduled), otherwise known as Buffered Network-Scheduled Unidirectional (BNU)
- Client / Server (unscheduled), otherwise known as Queued User-Triggered Bidirectional (QUB)
- Source / Sink (unscheduled), otherwise known as Queued User-Triggered Unidirectional (QUU)

Publisher / Subscriber: this VCR describes the action of a Compel Data token. The Link Active Scheduler (LAS) calls upon a specific device on the network to transmit specific data for a time-critical control purpose. When the addressed device responds with its data, multiple devices on the network "subscribing" to this published data receive it simultaneously. The publisher/subscriber VCR model is highly deterministic.

Client / Server: this VCR describes one class of unscheduled communications, permitted when a device receives a Pass Token (PT) message from the LAS. Each device maintains a queue (list) of data requests issued by other devices (clients), and responds to them in order as soon as it receives the Pass Token. By responding to client requests, the device acts as a server. Likewise, each device can use this time to act as a client, posting their own requests to other devices, which will act as servers when they receive the PT token from the LAS. This is how non-critical messages such as maintenance and device configuration data, operator setpoint changes, diagnostic messages, alarm acknowledgments and PID tuning values, etc. are exchanged between devices on an H1 segment Client/server communications are checked for data corruption by their receivers, to ensure reliable data flow.

Source / Sink (also called Report Distribution): this VCR describes another class of unscheduled communications, permitted when a device receives a Pass Token (PT) message from the LAS. This is where a device broadcasts data out to a "group address" representing many devices. Source/sink communications are not checked for data corruption, as are client/server communications.

An analogy for making sense of VCRs is to imagine lines drawn between FF devices on a segment to connect their various messages to other devices. Each line represents an individual transmission which must take place some time during the macrocycle. Each line is a VCR, some handled differently than others, some more critical than others, but all are nothing more than communication events in time. Later in this chapter, when you see function blocks connected together to form working control systems, think of the lines connecting blocks in different devices as VCRs[12].

[12] In the specific case of function block connections, each of those lines is a Publisher/Subscriber VCR, because each of those lines represents critical control data that must be passed from device to device in order for the function-block "program" to perform its control task.

16.3.3 Device capability

Not all FF devices are equally capable in terms of Data Link (layer 2) functions. The FF standard divides data link device functionality into three distinct groups, shown here in order of increasing capability:

- Basic devices
- Link Master devices
- Bridge devices

A *Basic* device is one capable of receiving and responding to tokens issued by the Link Active Scheduler (LAS) device. As discussed previously, these tokens may take the form of Compel Data (CD) messages which command immediate response from the Basic device, or Pass Token (PT) messages which grant the Basic device time-limited access to the segment for use in broadcasting data of lesser importance.

A *Link Master* device is one with the ability to be configured as the LAS for a segment. Not all FF devices have this ability, due to limited processing capability, memory, or both[13].

A *Bridge* device links multiple H1 segments together to form a larger network. Field instruments are never Bridge devices – a Bridge is a special-purpose device built for the express purpose of joining two or more H1 network segments.

16.4 FF function blocks

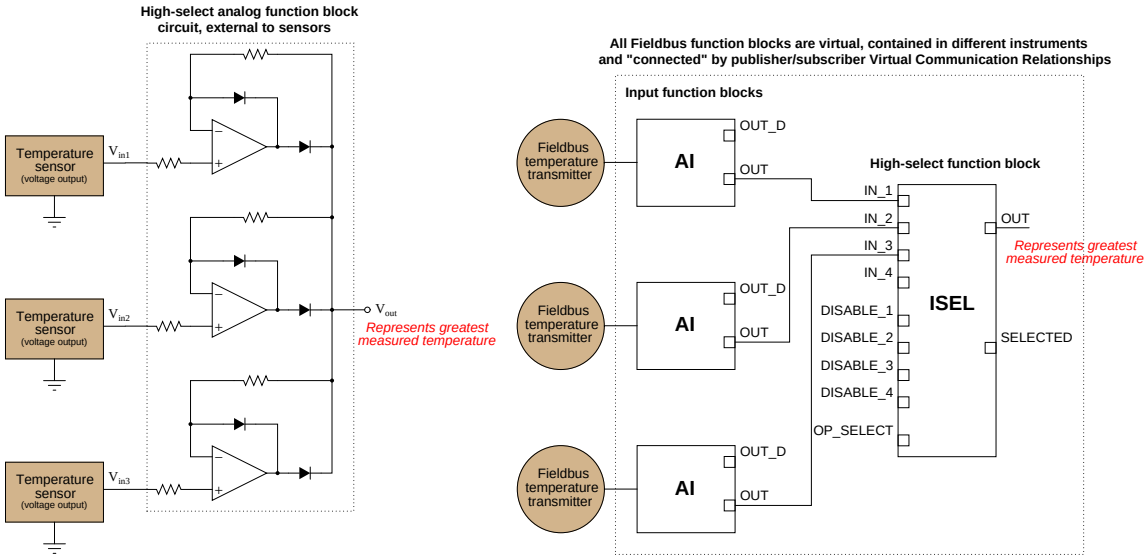
Data handled within FF systems are organized into modules known as *function blocks*. Sometimes these blocks serve merely to catalogue data, while in other instances the blocks execute specific algorithms useful for process measurement and control. These "blocks" are not physical entities, but rather abstract software objects – they exist only as bits of data and instructions in computer memory. However, the blocks are represented on a computer screen as rectangular objects with input ports on the left-hand side and output ports on the right-hand side. The construction of a working control system comprised of FF devices consists of linking the outputs of certain function blocks with the inputs of other function blocks via configuration software and computer-based tools. This usually takes the form of using a computer to draw connecting lines between the output and input ports of different function blocks.

[13] Some FF devices capable of performing advanced function block algorithms for certain process control schemes may have the raw computational power to be an LAS, but the manufacturer has decided not to make them Link Master capable simply to allow their computational power to be devoted to the function block processing rather than split between function block tasks and LAS tasks.

16.4.1 Analog function blocks versus digital function blocks

Function-block programming in general strongly resembles the design philosophy of legacy analog-based computer systems, where specific functions (addition, subtraction, multiplication, ratio, time-integration, limiting, and others) were encapsulated in discrete operational amplifier circuits, and whole systems were built by connecting function blocks together in whatever patterns were desired to achieve a design goal. Here with Fieldbus programming, the function blocks are virtual (bits and data structures in digital memory) rather than real analog circuits, and the connections between blocks are merely pointer assignments in digital memory rather than actual "patch cable" connections between circuit boards.

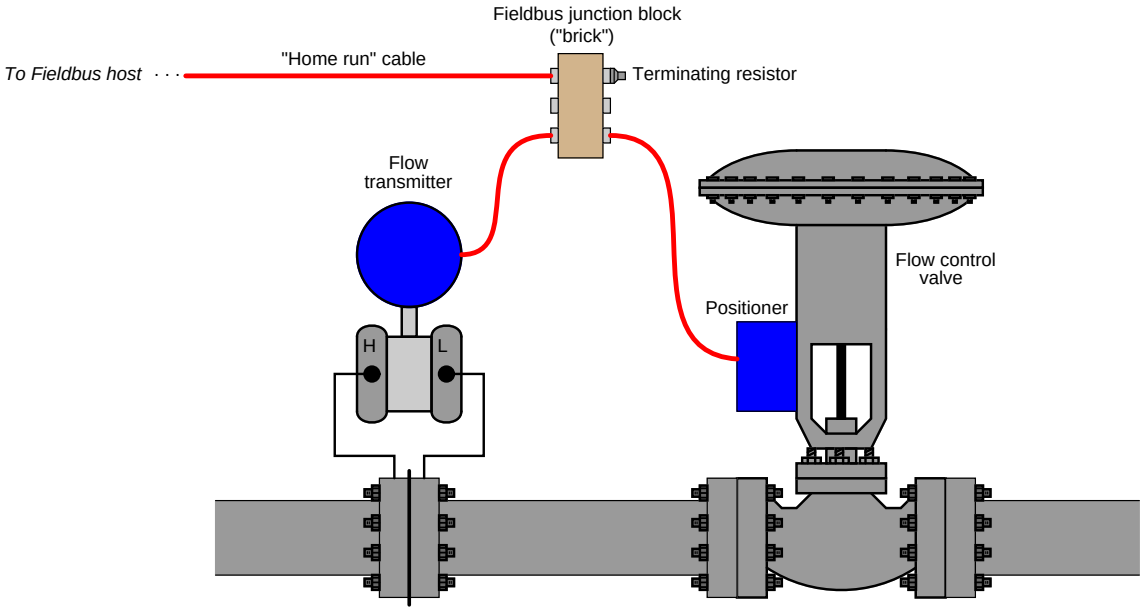
An example contrasting analog circuit design with Fieldbus function-block design appears here, both systems selecting the *greatest* temperature signal to be the output. The system on the left-hand side receives analog voltage signals from three temperature sensors, using a network of operational amplifiers, diodes, and resistors to select the greatest voltage signal to be the output. The system on the right-hand side uses three Fieldbus transmitters to sense temperature, selecting the greatest temperature by means of an algorithm executing in a Fieldbus device (it could be one of the FF transmitters, or even another device on the segment):



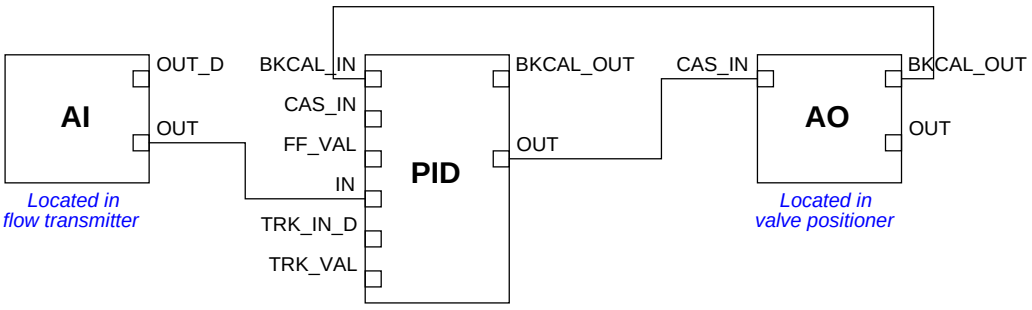
FOUNDATION Fieldbus abstracts the notion of discrete circuit modules performing compartmentalized tasks to software algorithms, "connecting" these algorithms together by a set of communication assignments whereby function blocks broadcast their output data to the network and the receiving blocks "listen" for the broadcasts at the right times.

16.4.2 Function block location

There is usually some freedom in where various function blocks may be located in a FF segment. Take for instance the example of a flow control loop, where a flow transmitter feeds measured flow data into a PID control function block, which then drives a control valve to whatever position necessary to regulate flow. The actual physical device layout might look something like this:



The function block connections necessary for this control scheme to work are shown in the next diagram, coupling the AI (analog input) block located in the transmitter to a PID control block to an AO (analog output) block located in the valve positioner:

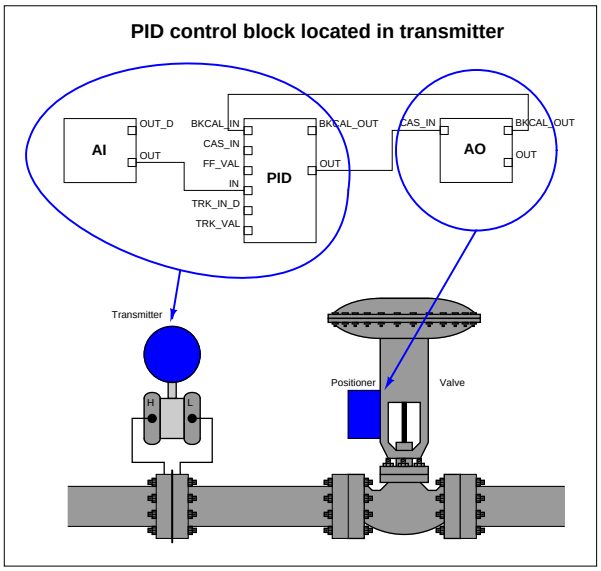


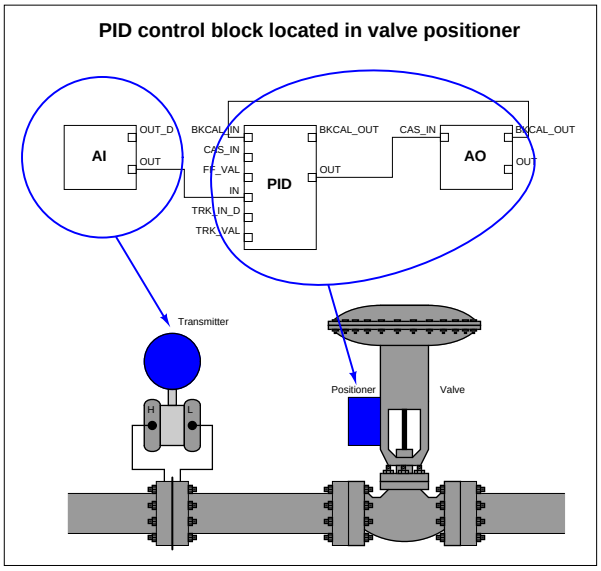
All function block inputs are on the left-hand sides of the blocks, and all outputs are on the right-hand sides. In this function block program, data from the analog input (AI) block flows into the PID block. After calculating the proper output value, the PID block sends data to the analog

output (AO) block where the final control element (e.g. valve, variable-speed motor) is adjusted. The AO block in turn sends a "back calculation" signal to the PID block to let it know the final control element has successfully reached the state commanded by the PID block's output. This is important for the elimination of *reset windup* in the event the final control element fails to respond to the PID block's output signal.

It should be obvious that the analog input (AI) block must reside in the transmitter, simply because only the transmitter is able to measure the process fluid flow rate. Likewise, it should be obvious that the analog output (AO) block must reside in the control valve positioner, simply because the valve is the only device capable of manipulating (exerting influence over) anything. However, given the lack of a separate controller device, the person configuring the Fieldbus loop may choose to locate the PID block in either the transmitter or the control valve positioner. So long as both FF devices possess PID function block capability, either location is possible for the PID function block.

The following illustrations show the two possible locations of the PID function block:

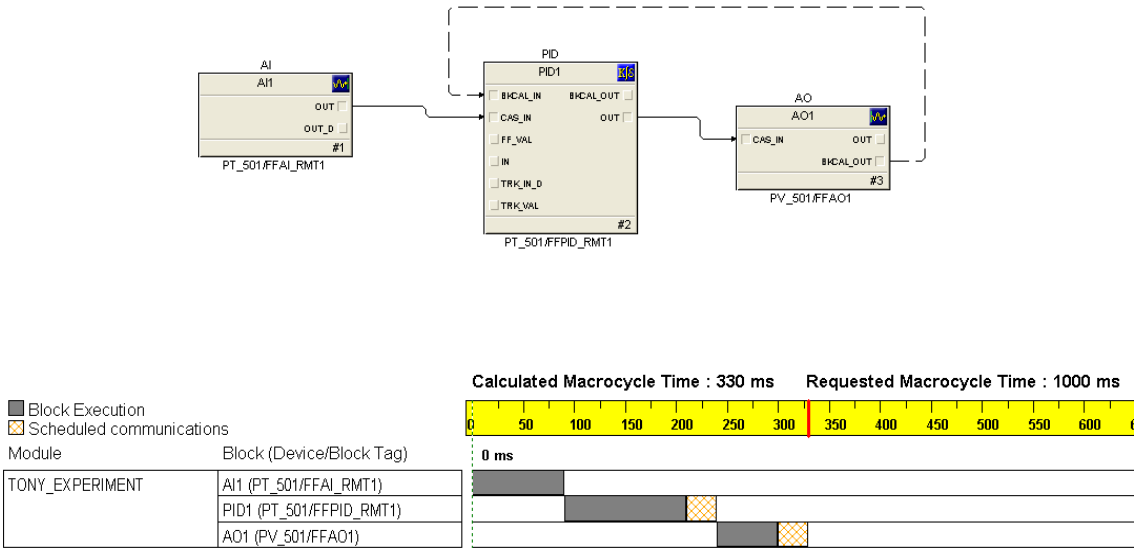




The only factor favoring one location over another for the PID function block is the number of communication broadcasts ("Compel Data" token distributions and replies) necessary per macrocycle. Note the lines connecting function blocks between the two instruments in the previous diagrams (lines crossing from one blue bubble to another). Each of these lines represents a VCR (Virtual Communication Relationship) – an instance during each macrocycle where data must be

transmitted over the network segment from one device to another. With the PID function block located in the flow transmitter, two lines connected the flow transmitter's PID block to the valve positioner's AO block. With the PID function block located in the valve positioner, only one line connected the flow transmitter's AI block to the valve positioner's PID block. Thus, locating the PID function block in the valve positioner means only one CD message/reply is necessary per macrocycle, making the network communication more efficient.

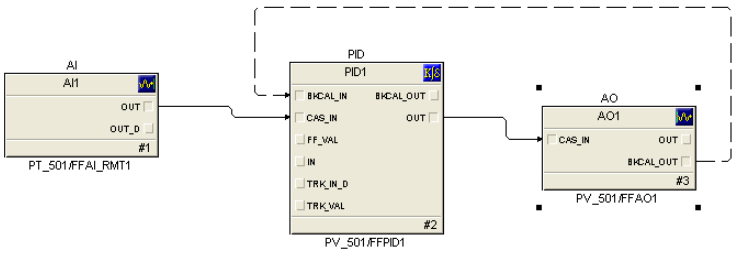
To illustrate the difference this re-location of the PID block makes, we will examine the function block diagram and macrocycle timing schedule on a simple pressure control FF loop, hosted on an Emerson DeltaV distributed control system. The first composite screenshot shows the function block diagram and schedule with the PID function block located in the transmitter (PT_501):

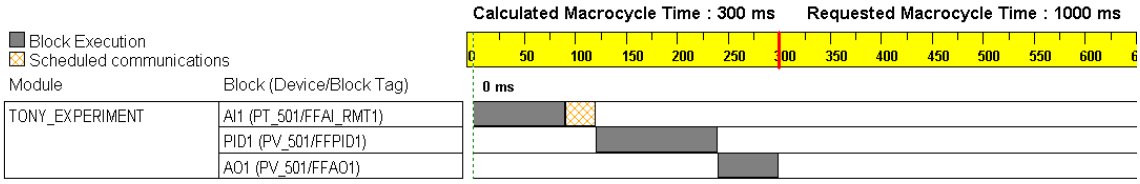


Note the two scheduled communication events (CD tokens and responses) necessary in the macrocycle schedule to enable communication between pressure transmitter PT_501's PID function block and valve positioner PV_501's analog output function block. The total (minimum) macrocycle time for this control loop is 330 milliseconds[14].

[14]This is not an unreasonable loop execution time for a pressure control system, especially if it is a gas process. Liquid pressure control is notoriously fast, and may experience trouble with a loop dead time of almost one-third of a second. For historical comparison, this execution time is on par with that of the original Honeywell TDC 2000 distributed control system, a hardware platform that has controlled many thousands of loops in oil refineries, pulp mills, chemical processing plants, and other industrial facilities worldwide. The practicality of a digital control loop with one-third second response is therefore proven by decades of practical application.

Now let's examine the same PID pressure control system with the PID function block moved to the valve. Here you see the function block diagram followed immediately by the updated macrocycle schedule:





Note that the macrocycle time is 30 milliseconds than before (300 milliseconds total as opposed to 330 milliseconds), since there is one less scheduled communications event happening. This represents a time reduction of almost 10% compared to the previous example, simply by assigning one function block to a different device on the segment.

16.4.3 Standard function blocks

The FF standard specifies many different function blocks for the construction of control algorithms. Ten of them are considered "basic" FF function blocks:

- `AI` – Analog Input
- `AO` – Analog Output
- `B` – Bias
- `CS` – Control Selector
- `DI` – Discrete Input
- `DO` – Discrete Output
- `ML` – Manual Loader
- `PD` – Proportional/Derivative control
- `PID` – Proportional/Integral/Derivative control
- `RA` – Ratio

Nineteen more "Advanced" function blocks are incorporated in the FF standard:

- Pulse Input
- Complex Analog Output
- Complex Discrete Output
- Step Output PID
- Device Control
- Setpoint Ramp
- Splitter
- Input Selector
- Signal Characterizer
- Dead Time
- Calculate
- Lead/Lag
- Arithmetic
- Integrator

- Timer
- Analog Alarm
- Discrete Alarm
- Analog Human Interface
- Discrete Human Interface

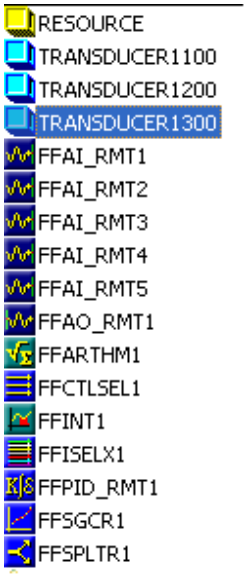
Five more function blocks are specified as well:

- Multiple Analog Input
- Multiple Analog Output
- Multiple Digital Input
- Multiple Digital Output
- Flexible Function Block

The primary benefit of standardization is that the end-user may choose FF instruments manufactured by any standard-compliant vendor, and those function blocks should behave the same as the equivalent function blocks within any other manufacturer's model of FF device. There are, of course, examples where manufacturers have equipped their FF devices with "extended" capability function blocks going beyond the Fieldbus Foundation standard, and the user must beware of this.

16.4.4 Device-specific function blocks

In addition to the function blocks necessary to construct control schemes, all FF instruments contain one *Resource* block and usually one or more *Transducer* blocks describing details specific to that instrument. The following screenshot shows all function blocks within a Rosemount model 3095MV Fieldbus transmitter:



The Resource block appears first in this list, followed by three transducer blocks, then followed by the palette of general function blocks for use in constructing control algorithms. Information contained in the Resource block of an FF instrument includes the following:

- Identifier (the 32-byte code unique to every FF device)
- Type of device
- Device revision level
- Memory total and available (free) capacity
- Computation time
- Available features listing
- Current device state (Initializing, Standby, On-line, Failed, etc.)

Transducer blocks provide a means of organizing data relevant to the actual sensing inputs, outputs, calculated variables, and graphic displays of a FF device. There need not be a one-to-one correspondence between the number of transducer blocks in an FF device and the number of physical I/O channels it has. For example, in the Rosemount 3095MV multivariable transmitter, transducer block 1100 handles all physical measurement inputs (pressure and temperature sensors)

while transducer block 1200 is reserved for inferred mass flow (based on calculations performed on the raw sensor measurements) and transducer block 1300 handles data for the liquid crystal display (LCD).

16.4.5 Status propagation

As mentioned earlier, function block programming bears a strong resemblance to analog function-block circuit design, where specific tasks are divided up into discrete elements, those elements connected together to form a larger system with more complex functionality. One of the important distinctions between the old-style analog function block circuit design and FF function block programming is the data content of the lines connecting blocks together. In the analog world, each connecting line (wire) carries exactly one piece of information: a single variable represented in analog form by a voltage signal. In the world of Fieldbus, each connecting line carries not only the variable's numerical value, but also a *status* and in some cases an *engineering unit* (a unit of measurement).

The inclusion of status along with data is a powerful concept, with roots in scientific practice. Scientists, as a rule, do their best to report the degree of *confidence* associated with the data they publish from experiments. Data is important, of course, but so is the degree of certainty with which that data was obtained. Obviously, data gathered with instruments of low quality (high uncertainty) will have different significance than data gathered with instruments of high precision and impeccable accuracy (low uncertainty). Any scientist wishing to base their theoretical work on a set of scientific data published by another scientist will have a measure of that data's significance – a very valuable detail.

By the same token, data "published" by a FF device is only as good as the health of that device. A FF transmitter exhibiting noisy or wildly fluctuating measurements might very well be nearing complete failure, and therefore its published data should be treated with skepticism. Since FF devices are "smart" (meaning, among other things, they have self-diagnostic capability), they have the ability to flag their own data as "Bad" if an internal fault is detected. The data still gets published and sent to other FF function blocks, but the status sent along with that data warns all downstream blocks of its uncertainty.

The three major status conditions associated with every FF signal passed between function blocks are:

- Good
- Bad
- Uncertain

Sub-status states also exist to further delineate the nature of the uncertainty. "Sensor Failure" is an example of a sub-status value, describing the reason for a "Bad" status value.

In computer science, there is a saying that "Garbage In equals Garbage Out," sometimes abbreviated as *GIGO*. No algorithm, no matter how advanced, can guarantee and output of good data from an input of bad data. This principle finds application in FF function block programming, as the blocks are programmed to switch mode when "Bad" or "Uncertain" input statuses are detected.

Furthermore, status values are *propagated* from the originating block all the way down through the last function block in the chain, reflecting the effect of an input signal's uncertainty on all consequent function outputs. For example, an analog input (AI) block sending a "Bad" status signal to the process variable input of a PID control block will have its "Bad" status propagated to the output of the PID block as well. Any function blocks receiving the PID block's output signal will likewise sense the "Bad" status and further propagate that status to their output signal(s).

16.4.6 Function block modes

All FF function blocks must support multiple *modes* of operation, describing how the block should execute its intended function. Several different function block modes are commonly found for FF function blocks, though not all FF function blocks support all of these modes:

- **OOS** (Out Of Service) – *All function blocks are required to support this mode, where the block freezes its output at the last calculated value and attaches a "Bad" status value*
- **Man** (Manual) – *the output of the block is determined by human control*
- **Auto** (Automatic) – *the function block processes information normally*
- **Cas** (Cascade) – *the function block processes information normally*
- **Iman** (Initialization Manual) – *the output of the block is fixed at its last calculated value, due to the output signal path being incomplete*
- **LO** (Local Override) – *the output of the block is fixed at its last calculated value, due to a detected fault condition within the device*
- **RCas** (Remote Cascade) – *the function block processes information normally based on a setpoint sent from a remote source to the block's RCas_In input*
- **ROut** (Remote Output) – *the function block passes data to its output sent from a remote source to the block's ROut_In input*

Instrumentation technicians and professionals are already familiar with the concept of a controller having "Automatic," "Manual," and even "Cascade" operating modes, but Fieldbus function block programming extends this general concept to each and every block. With FF, *each block* may be independently set into "Automatic" or "Manual" mode, which is a useful tool for testing FF algorithms and troubleshooting complex FF control schemes. The "Out of Service" mode, for instance, is commonly set by an instrument technician as he or she performs routine maintenance on an FF device (e.g. checking the calibration of an FF transmitter).

In addition to these operating modes for FF function blocks (not all of which are supported by all FF blocks), FF function blocks also have four mode categories describing valid modes for the block to be in under various conditions:

- Target
- Actual
- Permitted
- Normal

A block's "Target" mode is the mode it strives to be in if possible. The "Actual" mode is the mode the block is in at the present time. "Permitted" modes list all the different modes which may be used as "target" modes. "Normal" is a category describing to an operator interface what a block's normal operation mode should be, but the block itself does not heed this setting.

16.5 H1 FF device configuration and commissioning

Fieldbus devices require far more attention in their initial setup and commissioning than their analog counterparts. Unlike an analog transmitter, for example, where the only "configuration" settings are its zero and span calibration adjustments, a FF transmitter has a substantial number of parameters describing its behavior. Some of these parameters must be set by the end-user, while others are configured automatically by the host system during the start-up process, which we generally refer to as *commissioning.*

16.5.1 Configuration files

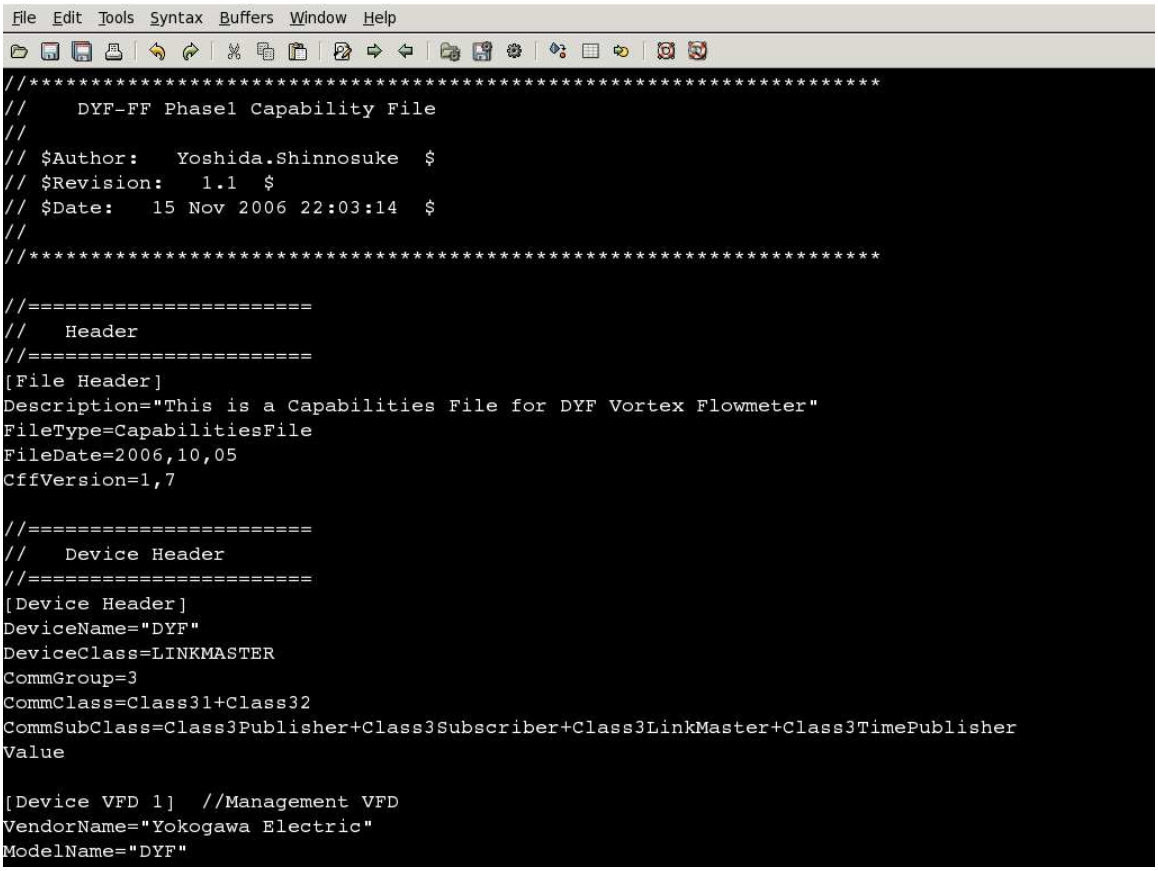
In order for a FF device to work together with a host system (which may be manufactured by a different company), the device must have its capabilities explicitly described so the host system "knows what to do with it." This is analogous to the need for *driver* files when interfacing a personal computer with a new peripheral device such as a printer, scanner, or modem.

A standardized language exists for digital instrumentation called the *Device Description Language*, or *DDL*. All FF instrument manufacturers are required to document their devices' capabilities in this standard-format language, which is then compiled by a computer into a set of files known as the *Device Description* (DD) files for that instrument. DDL itself is a text-based language, much like C or Java, written by a human programmer. The DD files are generated from the DDL source file by a computer, output in a form intended for another computer's read-only access. For FF instruments, the DD files end in the filename extensions `.sym` and `.ffo`, and may be obtained freely from the manufacturer or from the Fieldbus Foundation[15] The `.ffo` DD file is in a binary format readable only by a computer with the appropriate "DD services" software active. The `.sym` DD file is ASCII-encoded, making it viewable by a human by using a text editor program (although you should not attempt to edit the contents of a `.sym` file).

Other device-specific files maintained by the host system of a FF segment are the *Capability* and *Value* files, both referred to as *Common Format Files*, or `.cff` files. These are text-readable (ASCII encoded) digital files describing device capability and specific configuration values for the device, respectively. The Capability file for a FF device is typically downloaded from either the manufacturer's or the Fieldbus Foundation website along with the two DD files, as a three-file set (filename extensions being `.cff`, `.sym`, and `.ffo`, respectively). The Value file is generated by the host system during the device's configuration, storing the specific configuration values for that specific device and system tag number. The data stored in a Value file may be used to duplicate the exact configuration of a failed FF device, ensuring the new device replacing it will contain all the same parameters.

[15] One of the tasks of the Fieldbus Foundation is to maintain approved listings of FF devices in current manufacture. The concept is that whenever a manufacturer introduces a new FF device, it must be approved by the Fieldbus Foundation in order to receive the Fieldbus "badge" (a logo with a stylized letter "F"). Approved devices are cataloged by the Fieldbus Foundation, complete with their DD file sets. This process of approval is necessary for operational compatibility (called *interoperability*) between FF devices of different manufacture. Without some form of centralized standardization and approval, different manufacturers would invariably produce devices that were mutually incompatible with each other.

A screenshot of a `.cff` Capability file opened in a text editor program appears here, showing the first few lines of code describing the capabilities of a Yokogawa model DYF vortex flowmeter:

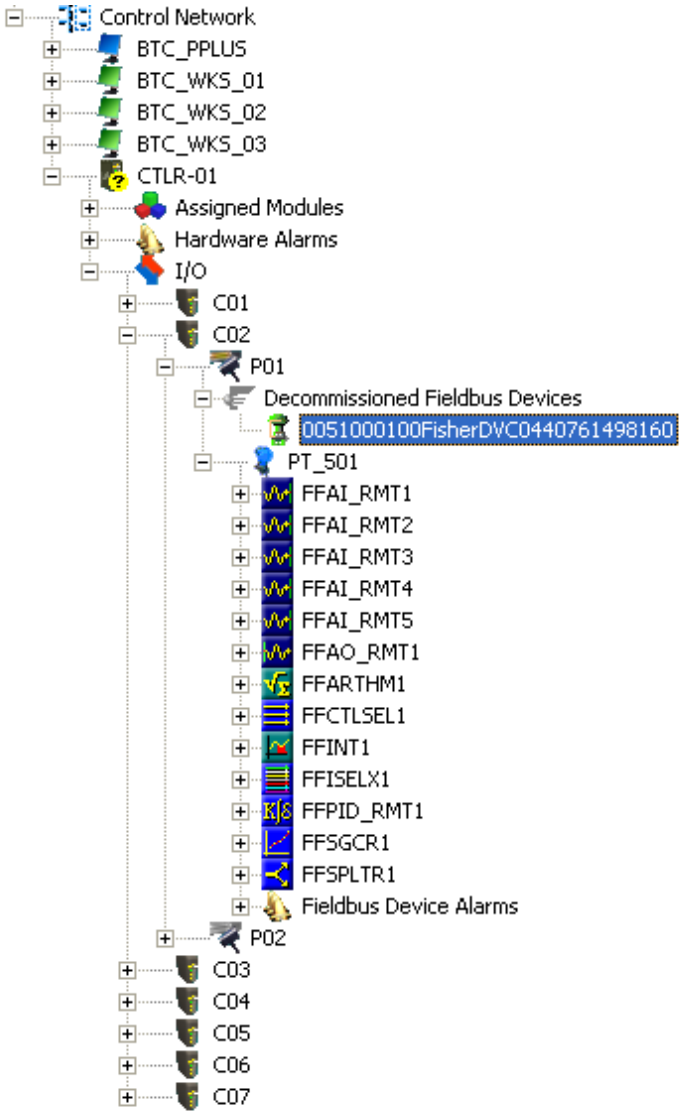


As with "driver" files needed to make a personal computer peripheral device function, it is important to have the correct versions of the Capability and DD files installed on the host system computer before attempting to commission the device. It is permissible to have Capability and DD files installed that are newer than the physical device, but not visa-versa (a newer physical device than the Capability and DD files). This requirement of proper configuration file management is a new task for the instrument technician and engineer to manage in their jobs. With every new FF device installed in a control system, the proper configuration files must be obtained, installed, and archived for safe keeping in the event of data loss (a "crash") in the host system.

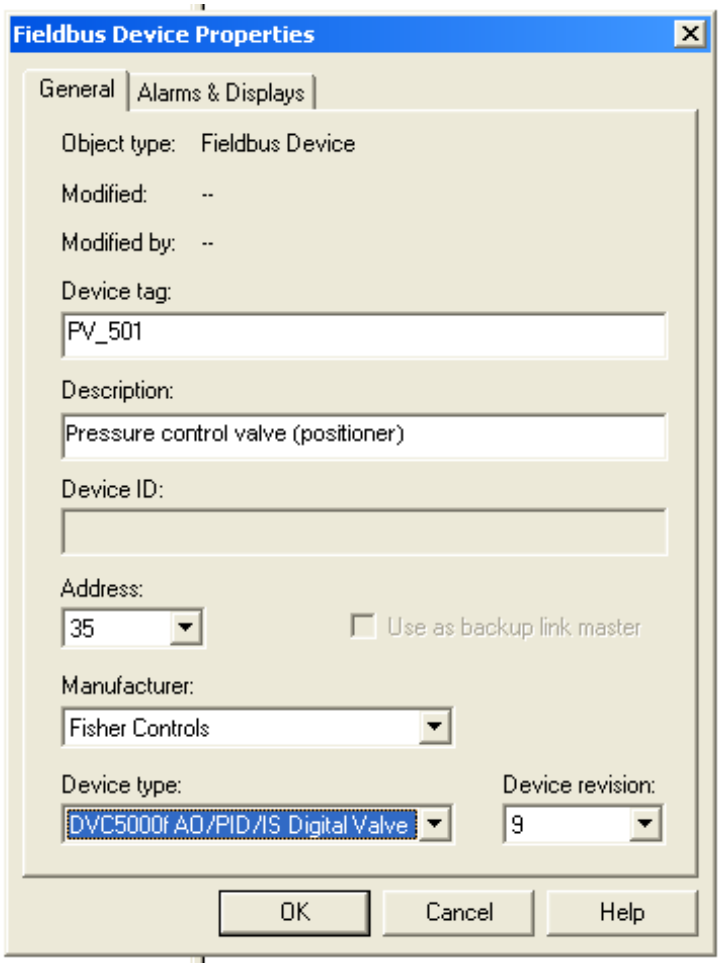
16.5.2 Device commissioning

This section illustrates the commissioning of a Fieldbus device on a real segment, showing screenshots of a host system's configuration menus. The particular device happens to be a Fisher DVC5000f valve positioner, and the host system is a *DeltaV* distributed control system manufactured by Emerson. All configuration files were updated in this system prior to the commissioning exercise. Keep in mind that the particular steps taken to commission any FF device will vary from one host system to another, and may not follow the sequence of steps shown here.

If an unconfigured FF device is connected to an H1 network, it appears as a "decommissioned" device. On the Emerson DeltaV host system, all decommissioned FF devices appear within a designated folder on the "container" hierarchy. Here, my Fisher DVC5000 device is shown highlighted in blue. A commissioned FF device appears just below it (PT_501), showing all available function blocks within that instrument:



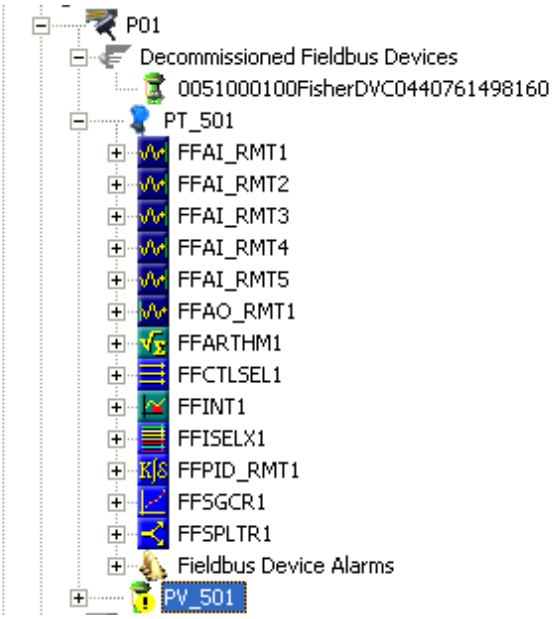
Before any FF device may be recognized by the DeltaV host system, a "placeholder" and tag name must be created for it within the segment hierarchy. To do this, a "New Fieldbus Device" must be added to the H1 port. Once this option is selected[16], a window opens up to allow naming of this new device:



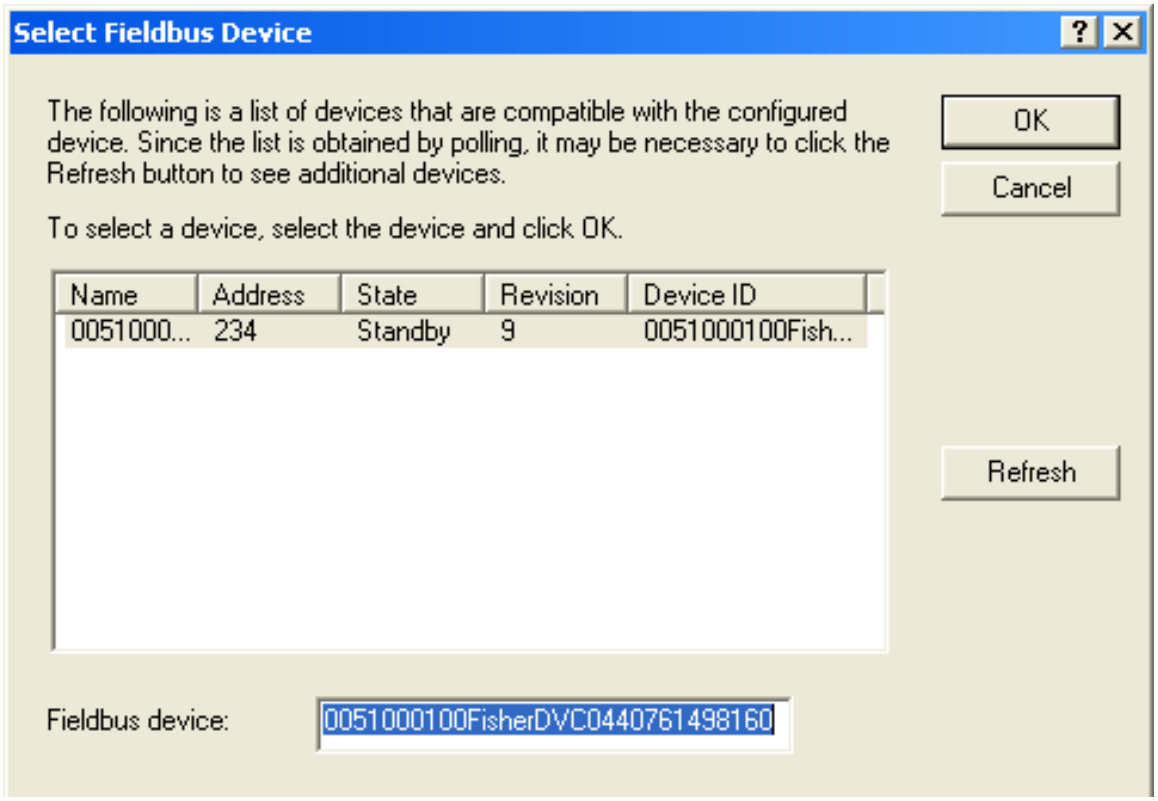
Here, the tag name "PV_501" has been chosen for the Fisher valve positioner, since it will work in conjunction with the pressure transmitter PT_501 to form a complete pressure control loop. In addition to a tag name (PV_501), I have also added a text description ("Pressure control valve (positioner)"), and specified the device type (Fisher DVC5000f with AO, PID, and IS function block capability). The DeltaV host system chose a free address for this device (35), although it is possible to manually select the desired device address at this point. Note the "Backup Link Master" check box in this configuration window, which is grey in color (indicating the option is not available with this device).

[16] On the Emerson DeltaV system, most options are available as drop-down menu selections following a right-mouse-button click on the appropriate icon.

After the device information has been entered for the new tag name, a "placeholder" icon appears within the hierarchy for the H1 segment (connected to Port 1). You can see the new tag name (PV_501) below the last function block for the commissioned FF instrument (PT_501). The actual device is still decommissioned, and appears as such:



By right-clicking on the new tag name and selecting the "Commission" option, a new window opens to allow you to select which decommissioned device should be given the new tag name. Since there is only one decommissioned device on the entire segment, only one option appears within the window:



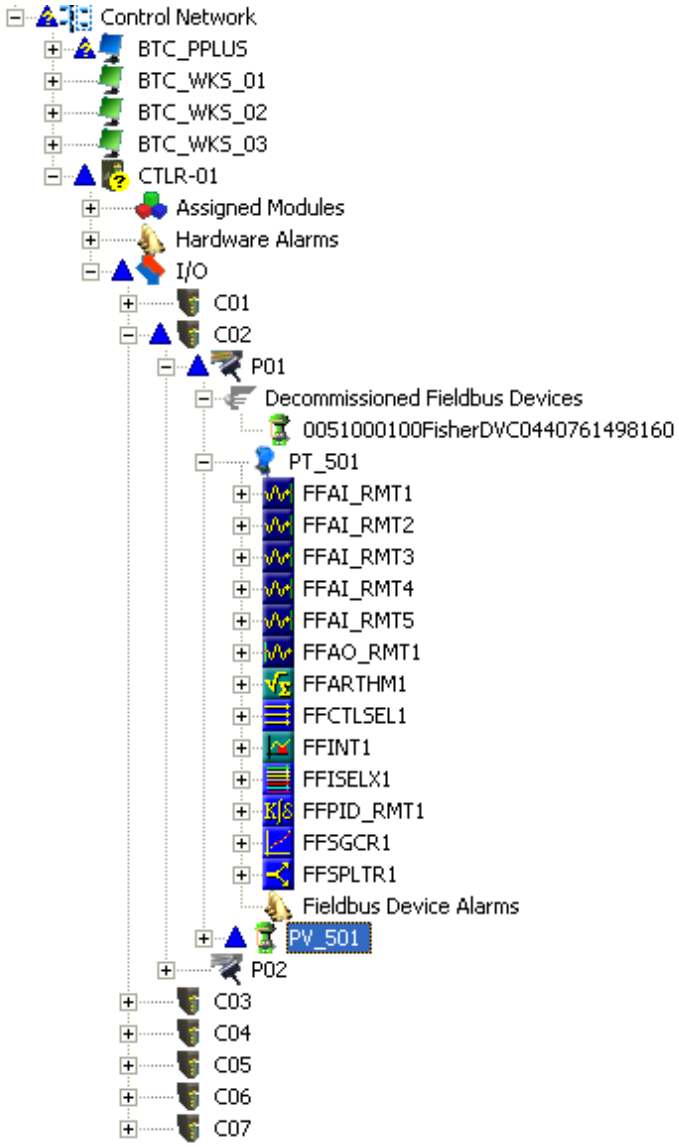
After selecting the decommissioned device you wish to commission, the DeltaV host system prompts you to reconcile any differences between the newly created tag name placeholder and the decommissioned device. It is possible the Resource and/or Transducer block parameters set within the placeholder do not match what is currently set in the decommissioned device, if that is what you desire. Otherwise, the existing block parameters within the decommissioned device will remain unchanged.

After selecting (or not selecting) the "reconcile" option, the DeltaV system prompts you to confirm commissioning of the device, after which it goes through a series of animated[17] display sequences as the device transitions from the "Standby" state to the "Commissioned" state:

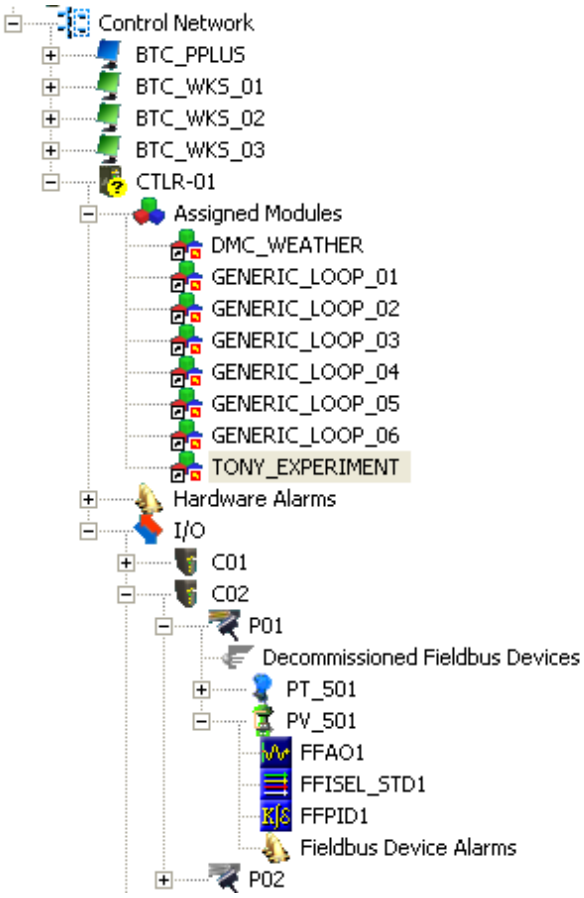
As you can see, the commissioning process is not very fast. After nearly one full minute of waiting, the device is still "Initializing" and not yet "Commissioned." The network speed of 31.25 kbps and the priority of scheduled communications are limiting factors when exchanging large quantities of configuration data over a FF H1 network segment. In order for device configuration to not interrupt or slow down process-critical data transfers, all configuration data exchanges must wait for unscheduled time periods, and then transmit at the relatively slow rate of 31.25 kbps when the alloted times arrive. Any technician accustomed to the fast data transfer rates of modern Ethernet devices will feel as though he or she has taken a step back in time when computers were *much* slower.

[17] Animated graphics on the Emerson DeltaV control system prominently feature an anthropomorphized globe valve named Duncan. There's nothing like a computer programmer with a sense of humor . . .

After commissioning this device on the DeltaV host system, several placeholders in the hierarchy appear with blue triangles next to them. In the DeltaV system, these blue triangle icons represent the need to download database changes to the distributed nodes of the system:



After "downloading" the data, the new FF valve positioner shows up directly below the existing pressure transmitter as a commissioned instrument, and is ready for service. The function blocks for pressure transmitter PT_501 have been "collapsed" back into the transmitter's icon, and the function blocks for the new valve positioner (PV_501) have been "expanded" for view:

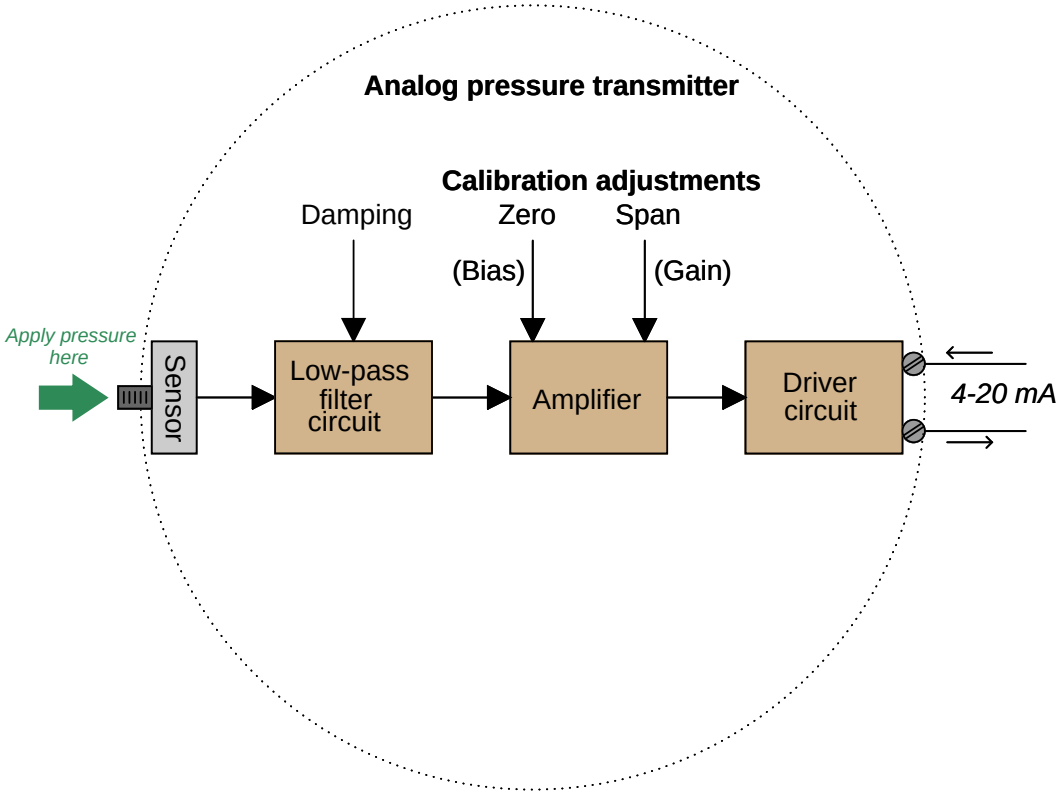


As you can see, the new instrument (PV_501) does not offer nearly as many function blocks as the original FF instrument (PT_501). The number of Fieldbus function blocks offered by any FF instrument is a function of that instrument's computational ability, internal task loading, and the whim of its designers. Obviously, this is an important factor to consider when designing a FF segment: being sure to include instruments that contain all the necessary function blocks to execute the desired control scheme. This may also become an issue if one of the FF instruments in a control scheme is replaced with one of a different manufacturer or model, having fewer available function blocks. If one or more mission-critical function blocks is not available in the replacement instrument, a different replacement must be sought.

16.5.3 Calibration and ranging

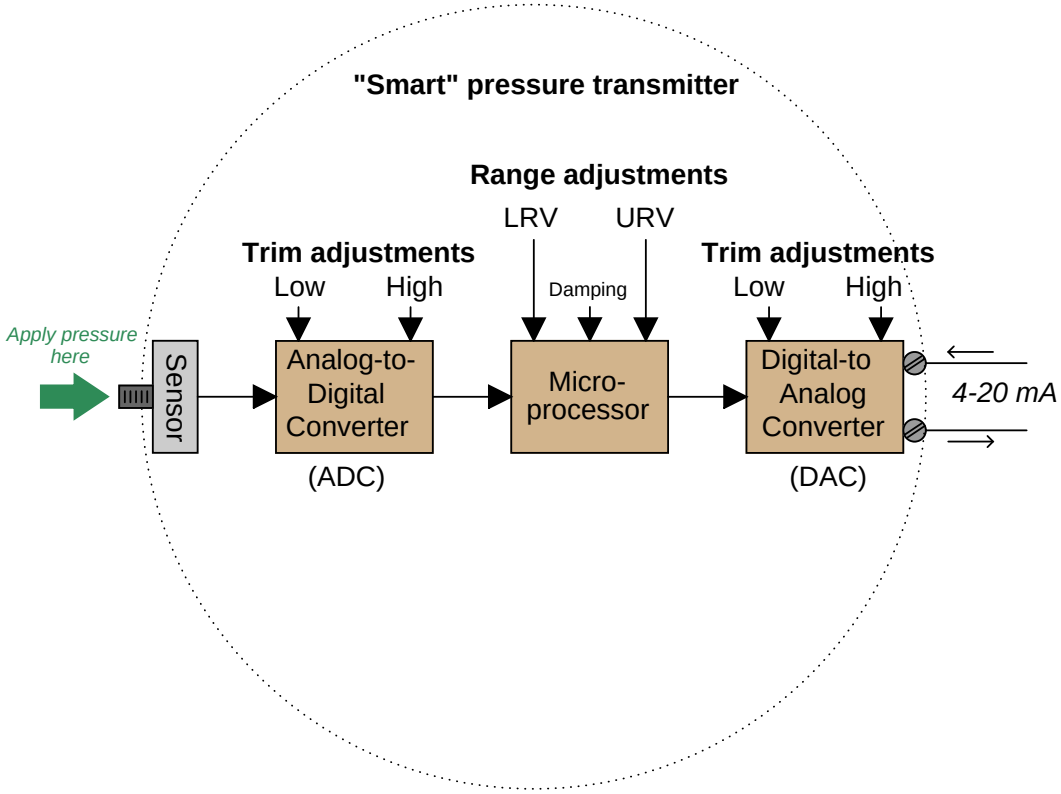
Calibration and ranging for a FF device is similar in principle to any other "smart" measurement instrument. Unlike analog instruments, where the "zero" and "span" adjustments completely define with the instrument's calibration and range, calibration and ranging are two completely different functions in a digital instrument.

A block diagram of an analog pressure transmitter shows the zero and span adjustments:



The "zero" and "span" adjustments together define the mathematical relationship between sensed pressure and current output. Calibration of an analog transmitter consists of applying known (reference standard) input stimuli to the instrument, and adjusting the "zero" and "span" settings until the desired current output values are achieved.

A "smart" (digital) transmitter equipped with an analog 4-20 mA current output distinctly separates the calibration and range functions:

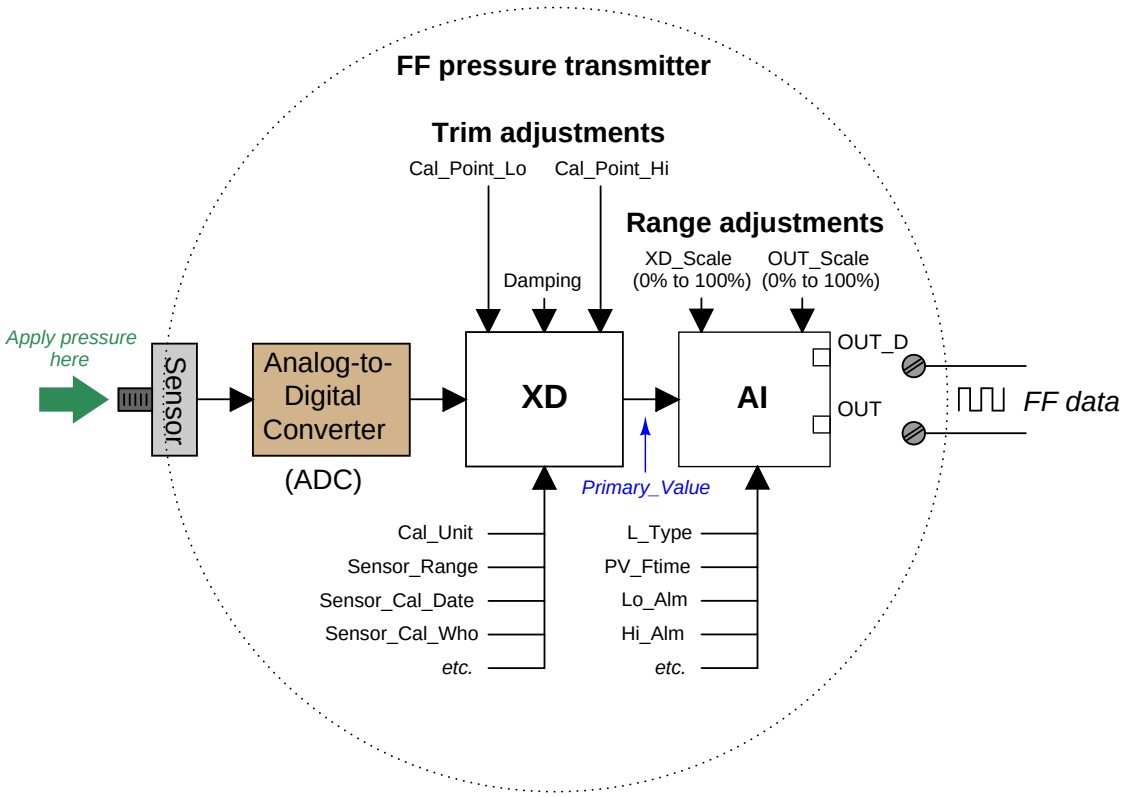


Calibration of a "smart" transmitter consists of applying known (reference standard) input stimuli to the instrument and engaging the "trim" functions until the instrument accurately registers the input stimuli. Ranging, by contrast, establishes the mathematical relationship between the registered input value and the output current value. To illustrate the difference between calibration and ranging, consider a case where a pressure transmitter is used to measure water flow through a venturi tube. Suppose the transmitter's pressure range of 0 to 100 inches water column translates to a venturi tube flow range of 0 to 250 gallons per minute. If we desired to re-range an analog pressure transmitter to measure a lower range of flow (say, 0 to 130 gallons per minute), we would have to re-calculate the new pressure range (0 to 27.04 inches water column, abiding by the quadratic behavior of the venturi tube) and then subject the analog pressure transmitter to a new (standard) pressure of 27.04 inches water column while we re-adjusted the transmitter's zero and span so it accurately represented the new pressure range. The only way we can re-range an analog transmitter is to completely re-calibrate it.

In a "smart" (digital) measuring instrument, however, calibration against a known (standard) source need only be done at the specified intervals to ensure accuracy despite the instrument's inevitable drift. If our hypothetical transmitter were calibrated accurately against a known pressure

standard and relied upon not to have drifted since the last calibration cycle, we could re-range it by programming it with the new LRV (lower range value) and URV (upper range value) parameters so that 27.04 inches water column now drives its current output to 20 mA instead of 100 inches water column pressure as was required before. Digital instrumentation allows us to re-range without re-calibrating, representing a tremendous savings in technician time and effort.

Fieldbus instruments, of course, are "smart" in the same way, and their internal block diagrams look much the same as the "smart" transmitters with analog current output, albeit with a far greater number of parameters within each block. The rectangle labeled "XD" in the following diagram is the Transducer block, while the rectangle labeled "AI" is the Analog Input block:



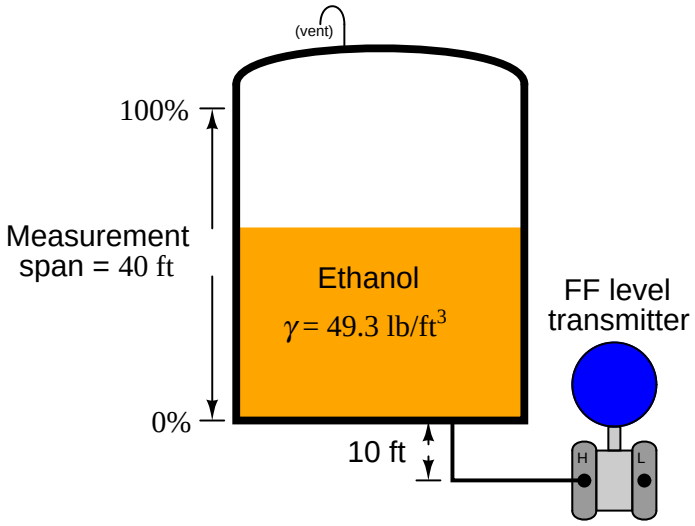
Calibration (trim) values are set in the transducer block along with the engineering unit, making the output of the transducer block a digital value scaled in real units of measurement (e.g. PSI, kPa, bar, mm Hg, etc.) rather than some raw ADC "count" value. The analog input function block receives this pre-scaled "Primary Value" and translates it to another scaled value based on a proportionality between transducer scale values (`XD_Scale` high and low) and output scale values (`OUT_Scale` high and low). The `L_Type` parameter residing in the analog input block determines whether the ranging is direct (output value equals primary input value), indirect (proportionately scaled), or indirect with square-root characterization (useful for translating a differential pressure measurement across a flow element into an actual fluid flow rate).

To calibrate such a transmitter, the transducer block should first be placed in *Out Of Service* (OOS) mode using a handheld FF communicator or the Fieldbus host system. Next, a standard (calibration-grade) fluid pressure is applied to the transmitter's sensor and the `Cal_Point_Lo` parameter is set to equal this applied pressure. After that, a greater pressure is applied to the sensor and the `Cal_Point_Hi` parameter is set to equal this applied pressure. After setting the various

calibration record-keeping parameters (e.g. `Sensor_Cal_Date`, `Sensor_Cal_Who`), the transducer block's mode may be returned to *Auto* and the transmitter used once again.

To range such a transmitter, a correspondence between sensed pressure and the process variable must be determined and entered into the analog input function block's `XD_Scale` and `OUT_Scale` parameters. If the pressure transmitter is being used to indirectly measure something other than pressure, these range parameters will become very useful, not only proportioning the numerical values of the measurement, but also casting the final digital output value into the desired "engineering units" (units of measurement).

The concept of ranging a FF transmitter makes more sense viewed in the context of a real application. Consider this example, where a pressure transmitter is being used to measure the level of ethanol (ethyl alcohol) stored in a 40 foot high tank. The transmitter connects to the bottom of the tank by a tube, and is situated 10 feet below the tank bottom:



Hydrostatic pressure exerted on the transmitter's sensing element is the product of liquid density (γ) and vertical liquid column height (h). When the tank is empty, there will still be a vertical column of ethanol 10 feet high applying pressure to the transmitter's "high" pressure port. Therefore, the pressure seen by the transmitter in an "empty" condition is equal to:

$$P_{empty} = \gamma h_{empty} = (49.3 \text{ lb/ft}^3)(10 \text{ ft})$$

$$P_{empty} = 493 \text{ lb/ft}^2 = 3.424 \text{ PSI}$$

When the tank is completely full (40 feet), the transmitter sees a vertical column of ethanol 50 feet high (the tank's 40 foot height plus the suppression height of 10 feet created by the transmitter's location below the tank bottom). Therefore, the pressure seen by the transmitter in a "full" condition is equal to:

$$P_{full} = \gamma h_{full} = (49.3 \text{ lb/ft}^3)(50 \text{ ft})$$

$$P_{full} = 2465 \text{ lb/ft}^2 = 17.12 \text{ PSI}$$

The control system does not "care" about the transmitter's 10-foot suppression, though. All it needs to know is where the ethanol level is in relation to the tank bottom (relative to an "empty" condition). Therefore, when we range this transmitter for the application, we will set the analog input block's range parameters as follows[18]:

AI block parameter	Range values
XD_Scale	3.424 PSI to 17.12 PSI
OUT_Scale	0 feet to 40 feet
L_Type	Indirect

Now, the ethanol tank's level will be accurately represented by the FF transmitter's output, both in numeric value and measurement unit. An empty tank generating a pressure of 3.424 PSI causes the transmitter to output a "0 feet" digital signal value, while a full tank generating 17.12 PSI of pressure causes the transmitter to output a "40 feet" digital signal value. Any ethanol levels between 0 and 40 feet will likewise be represented proportionally by the transmitter.

If at some later time the decision is made to re-locate the transmitter so it no longer has a 10 foot "suppression" with regard to the tank bottom, the XD_Scale parameters may be adjusted to reflect the corresponding shift in pressure range, and the transmitter will still accurately represent ethanol level from 0 feet to 40 feet, without adjusting or re-calibrating anything else in the transmitter.

16.6 H1 FF segment troubleshooting

Feedback obtained from industrial users of FF reveal a common pattern: Fieldbus is a wonderful technology, but only if it is properly installed. Poor installations, usually driven by a desire to minimize capital expenses, will cause numerous problems during commissioning and operation.

One relatively easy way to avoid problems caused by short-circuits in FF wiring is to use coupling devices with built-in short-circuit protection. This feature does not add significant cost to the coupling device, and it will prevent the entire segment from failing due to a short-circuit on a single spur cable or within a device. Use coupling devices with indicator LEDs as well, since these give easy visual verification of network power which may greatly accelerate FF segment troubleshooting when the need arises.

[18] When configuring the XD_Scale high and low range values, be sure to maintain consistency with the transducer block's Primary_Value_Range parameter unit. Errors may result from mis-matched measurement units between the transducer block's measurement channel and the analog input block's XD_Scale parameter.

16.6.1 Cable resistance

A simple check of an H1 segment's cabling consists of a series of resistance measurements performed with the segment unpowered (as is standard with any electrical resistance check), with all FF devices disconnected, and with the cable entirely disconnected (all three conductors) at the host end. The following table shows guidelines published by the Fieldbus Foundation for H1 segment cable resistance measurements:

Measurement points	Expected resistance
Between (+) and (-) conductors	> 50 kΩ, increasing over time
Between (+) conductor and shield (ground)	> 20 MΩ
Between (-) conductor and shield (ground)	> 20 MΩ
Between shield conductor and earth ground	> 20 MΩ

The last resistance check shown in the table checks for the presence of ground connections in the shield conductor *other than* the one ground connection at the host end (which has been disconnected for the purposes of the test). Since the shield should only be grounded at one point[19] (to avoid ground loops), and this one point has been disconnected, the shield conductor should register no continuity with earth ground during the test.

The necessity of disconnecting all FF devices and host system interfaces is essential so that the resistance measurements reflect the health of the cable and nothing else. The presence of any FF devices on the segment would substantially affect the resistance measurements, particularly resistance between the signal (+ and -) conductors.

16.6.2 Signal strength

The Fieldbus Foundation specifies a signal voltage (peak-to-peak) range of 350 mV to 700 mV for a healthy FF segment. Excessive signal voltage levels point to a lack of terminator resistor(s), while insufficient voltage levels point to an over-abundance of terminators (or perhaps even a device short):

Signal voltage (pk-pk)	Interpretation
800 mV or more	Possibly missing terminator resistor
350 mV to 700 mV	Good signal strength
150 mV to 350 mV	Marginally low signal – possible extra terminator resistor(s)
150 mV or less	Too little signal to function

[19] An alternative method of shield grounding is to directly connect it to earth ground at one end, and then capacitively couple it to ground at other points along the segment length. The capacitor(s) provide an AC path to ground for "bleeding off" any induced AC noise without providing a DC path which would cause a ground loop.

16.6.3 Electrical noise

FF, like all digital networks, are unaffected by noise voltage below a certain threshold. If noise voltage is present in excessive quantity, though, it may cause bits to be misinterpreted, causing data errors. The Fieldbus Foundation gives the following recommendations[20] for noise voltage levels on a FF segment:

Noise voltage (pk-pk)	Interpretation
25 mV or less	Excellent
25 mV to 50 mV	Okay
50 mV to 100 mV	Marginal
100 mV or more	Poor

16.6.4 Using an oscilloscope on H1 segments

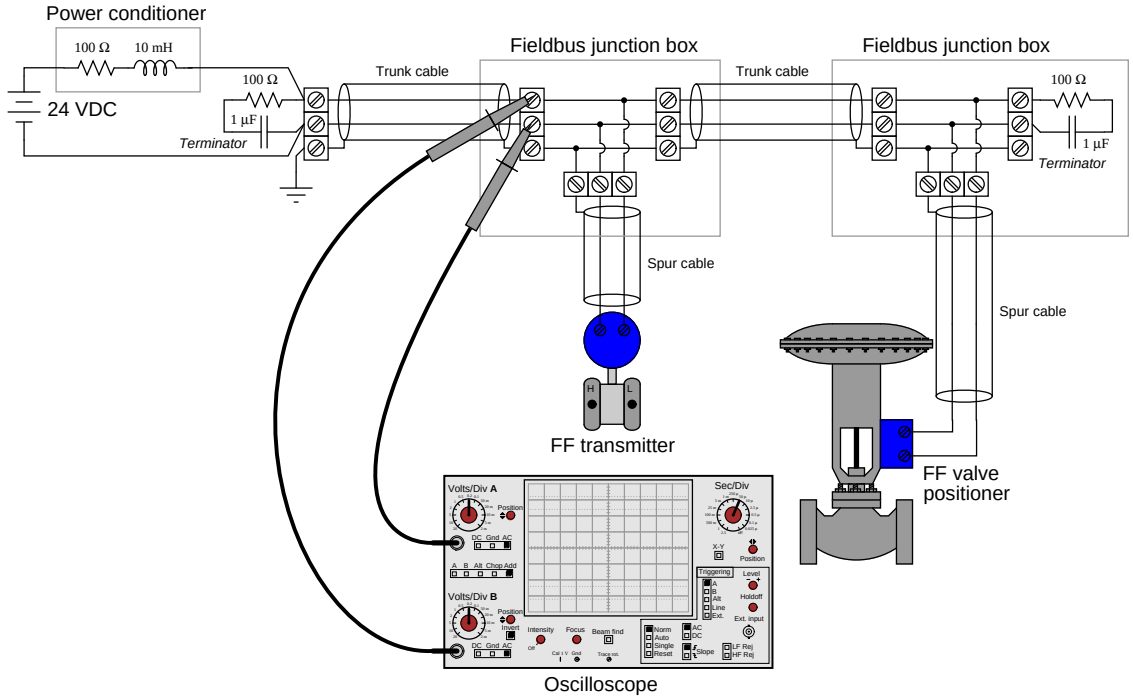
A tool available in most instrument shops is a digital-storage oscilloscope, which may be used to measure and display FF H1 signal waveforms for analysis of problems. Analog oscilloscopes are also useful for network troubleshooting, but to a lesser degree[21].

When using an oscilloscope to measure FF H1 signals, it is very important not to connect either of the FF segment conductors to earth ground through the oscilloscope. Doing so will almost certainly *cause* network problems, in addition to whatever problems already exist that compel you to diagnose with an oscilloscope. If a single channel of the oscilloscope is connected across the segment wires, the "ground" clip of the probe will force one of those conductors to earth ground potential via the metal chassis of the oscilloscope which is grounded through the third prong of the power plug for safety. An exception to this rule is if the oscilloscope itself is battery-powered and has an insulated case where no ground connection is made through the surface it sits on or the human hand that holds it. Otherwise, using a single channel on a line-powered oscilloscope to measure network signals is inviting trouble.

[20] Bear in mind the tolerable level for noise will vary with signal voltage level as well. All other factors being equal, a strong signal is less affected by the presence of noise than a weak signal (i.e. the signal-to-noise ratio, or *SNR*, is crucial).

[21] It is impossible to "lock in" (trigger) non-periodic waveforms on an analog oscilloscope, and so most network communications will appear as an incomprehensible blur when viewed on this kind of test instrument. Digital oscilloscopes have the ability to "capture" and display momentary pulse streams, making it possible to "freeze" any portion of a network signal for visual analysis.

If a line-powered oscilloscope must be used, the proper way to configure it is for *differential channel* measurement. In this mode, the oscilloscope will register the voltage *between* two probe tips, rather than register the voltage between a single probe tip and earth ground.



Configuring a dual-trace oscilloscope for differential mode is quite simple. On the front panel of the oscilloscope, you must set the multi-trace controls to the *Add* mode, where one trace on the screen represents the instantaneous sum of the two inputs (channels "A" and "B"). The volts per division "sensitivity" of both channels should be set to exactly the same value. Also, the *Invert* control must be engaged for the second input channel, forcing that channel's signal to be inverted (register upside-down on the screen). The summation of channel "A" and an inverted channel "B" is equivalent to the mathematical difference (subtraction) between "A" and "B," which means the single trace on the screen now represents the difference of potential between the two probe tips. The oscilloscope now behaves as an ungrounded voltmeter, where neither of the test leads is referenced to earth ground.

16.6.5 Message re-transmissions

Aside from voltage parameters (signal strength, noise amplitude), another good indicator of FF segment health is the number of message *re-transmissions* over time. Certain types of communication on an H1 segment require verification of a received signal (particularly client/server VCRs such as those used to communicate operator setpoint changes and diagnostic messages). If the signal received by the client FF device appears corrupted, the device will request a *re-transmission* of the message from the server device. Re-transmission events, therefore, are an indication of how often messages are getting corrupted, which is a direct function of signal integrity in a Fieldbus segment.

Most host systems provide re-transmission statistics in much the same way that computers communicating via TCP/IP protocol have the ability to display the number of "lost" data packets over time. Since nearly all FF segments function with a host system connected, this becomes a built-in diagnostic tool for technicians to troubleshoot FF network segments.

Hand-held diagnostic tools are also manufactured to detect signal voltage levels, noise voltage levels, and message re-transmissions. Relcom manufactures both the model FBT-3 and model FBT-6 hand-held Fieldbus testers at the time of this writing (2009), the FBT-6 being the more capable of the two test devices.

References

"Fieldbus Book – A Tutorial" (TI 38K02A01-01E) 1st Edition , Yokogawa Electric Corporation, Tokyo, Japan, 2001.

"FOUNDATION Fieldbus Application Guide – 31.25 kbit/s Intrinsically Safe Systems" (AG 163) Revision 2.0, The Fieldbus Foundation, Austin, TX, 2004.

"FOUNDATION Fieldbus Blocks" (00809-0100-4783) Revision BA, Rosemount, Inc., Chanhassen, MN, 2000.

"FOUNDATION Fieldbus System Engineering Guidelines" (AG 181) Revision 2.0, The Fieldbus Foundation, Austin, TX, 2004.

"FOUNDATION Specification System Architecture" (FF 581) Revision FS 1.1, The Fieldbus Foundation, Austin, TX, 2000.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Software and Digital Networks*, Third Edition, CRC Press, New York, NY, 2002.

"Model 3051 Transmitter with FOUNDATION Fieldbus" (00809-0100-4774) Revision AA, Rosemount, Inc., Chanhassen, MN, 1999.

"RSFieldbus – Configuring and Programming Foundation Fieldbus Devices Application Guide" (RSFBUS-AT001A-EN-E), Rockwell Software, Inc., Milwaukee, WI, 2004.

Smith, John I., *Modern Operational Circuit Design*, Wiley-Interscience, John Wiley & Sons, Inc., New York, NY, 1971.

Park, John; Mackay, Steve; Wright, Edwin; *Practical Data Communications for Instrumentation and Control*, IDC Technologies, published by Newnes (an imprint of Elsevier), Oxford, England, 2003.

"Rosemount 3095 MultiVariable Mass Flow Transmitter with HART or FOUNDATION Fieldbus Protocol" (00809-0100-4716) Revision JA, Rosemount, Inc., Chanhassen, MN, 2008.

"The FOUNDATION Fieldbus Primer" Revision 1.1, Fieldbus Inc., Austin, TX, 2001.

"Wiring and Installation 31.25 kbit/s, Voltage Mode, Wire Medium Application Guide" (AG-140) Revision 1.0, Fieldbus Foundation, Austin, TX, 2000.

Chapter 17

Instrument calibration

Every instrument has at least one *input* and one *output*. For a pressure sensor, the input would be some fluid pressure and the output would (most likely) be an electronic signal. For a loop indicator, the input would be a 4-20 mA current signal and the output would be a human-readable display. For a variable-speed motor drive, the input would be an electronic signal and the output would be electric power to the motor.

Calibration and *ranging* are two tasks associated with establishing an accurate correspondence between any instrument's input signal and its output signal.

17.1 Calibration versus re-ranging

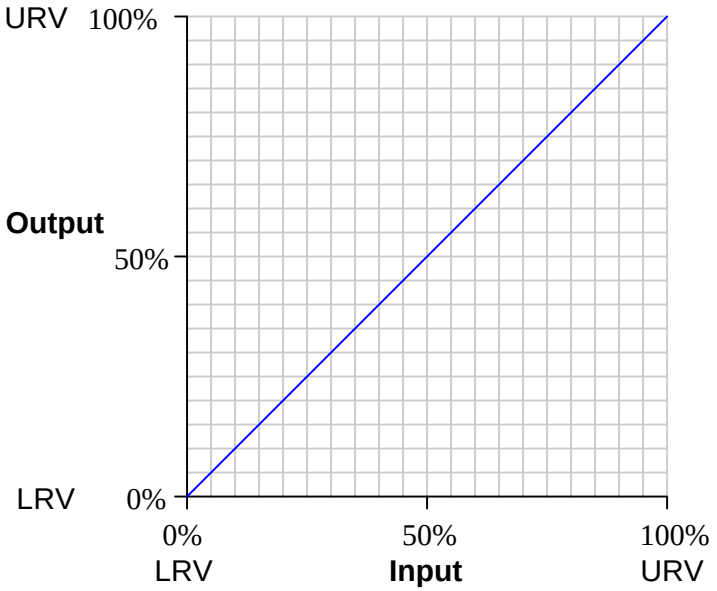
To *calibrate* an instrument means to check and adjust (if necessary) its response so the output accurately corresponds to its input throughout a specified range. In order to do this, one must expose the instrument to an actual input stimulus of precisely known quantity. For a pressure gauge, indicator, or transmitter, this would mean subjecting the pressure instrument to known fluid pressures and comparing the instrument response against those known pressure quantities. One cannot perform a true calibration without comparing an instrument's response to known, physical stimuli.

To *range* an instrument means to set the lower and upper range values so it responds with the desired sensitivity to changes in input. For example, a pressure transmitter set to a range of 0 to 200 PSI (0 PSI = 4 mA output ; 200 PSI = 20 mA output) could be re-ranged to respond on a scale of 0 to 150 PSI (0 PSI = 4 mA ; 150 PSI = 20 mA).

In analog instruments, re-ranging could (usually) only be accomplished by re-calibration, since the same adjustments were used to achieve both purposes. In digital instruments, calibration and ranging are typically separate adjustments (i.e. it is possible to re-range a digital transmitter without having to perform a complete recalibration), so it is important to understand the difference.

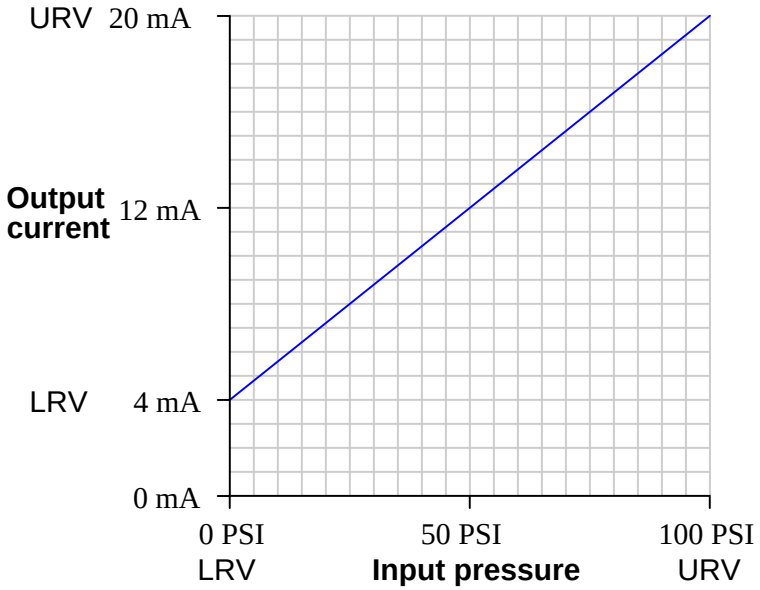
17.2 Zero and span adjustments (analog transmitters)

The purpose of *calibration* is to ensure the input and output of an instrument correspond to one another predictably throughout the entire range of operation. We may express this expectation in the form of a graph, showing how the input and output of an instrument should relate:



This graph shows how any given percentage of input should correspond to the same percentage of output, all the way from 0% to 100%.

Things become more complicated when the input and output axes are represented by units of measurement other than "percent." Take for instance a pressure *transmitter*, a device designed to sense a fluid pressure and output an electronic signal corresponding to that pressure. Here is a graph for a pressure transmitter with an input range of 0 to 100 pounds per square inch (PSI) and an electronic output signal range of 4 to 20 milliamps (mA) electric current:



Although the graph is still linear, zero pressure does not equate to zero current. This is called a *live zero*, because the 0% point of measurement (0 PSI fluid pressure) corresponds to a non-zero ("live") electronic signal. 0 PSI pressure may be the LRV (Lower Range Value) of the transmitter's input, but the LRV of the transmitter's output is 4 mA, not 0 mA.

Any linear, mathematical function may be expressed in "slope-intercept" equation form:

$$y = mx + b$$

Where,

y = Vertical position on graph

x = Horizontal position on graph

m = Slope of line

b = Point of intersection between the line and the vertical (y) axis

This instrument's calibration is no different. If we let x represent the input pressure in units of PSI and y represent the output current in units of milliamps, we may write an equation for this instrument as follows:

$$y = 0.16x + 4$$

On the actual instrument (the pressure transmitter), there are two adjustments which let us match the instrument's behavior to the ideal equation. One adjustment is called the *zero* while

the other is called the *span*. These two adjustments correspond exactly to the b and m terms of the linear function, respectively: the "zero" adjustment shifts the instrument's function vertically on the graph, while the "span" adjustment changes the slope of the function on the graph. By adjusting both zero and span, we may set the instrument for any range of measurement within the manufacturer's limits.

The relation of the slope-intercept line equation to an instrument's zero and span adjustments reveals something about how those adjustments are actually achieved in any instrument. A "zero" adjustment is always achieved by *adding* or *subtracting* some quantity, just like the y-intercept term b adds or subtracts to the product mx. A "span" adjustment is always achieved by *multiplying* or *dividing* some quantity, just like the slope m forms a product with our input variable x.

Zero adjustments typically take one or more of the following forms in an instrument:

- Bias force (spring or mass force applied to a mechanism)
- Mechanical offset (adding or subtracting a certain amount of motion)
- Bias voltage (adding or subtracting a certain amount of potential)

Span adjustments typically take one of these forms:

- Fulcrum position for a lever (changing the force or motion multiplication)
- Amplifier gain (multiplying or dividing a voltage signal)
- Spring rate (changing the force per unit distance of stretch)

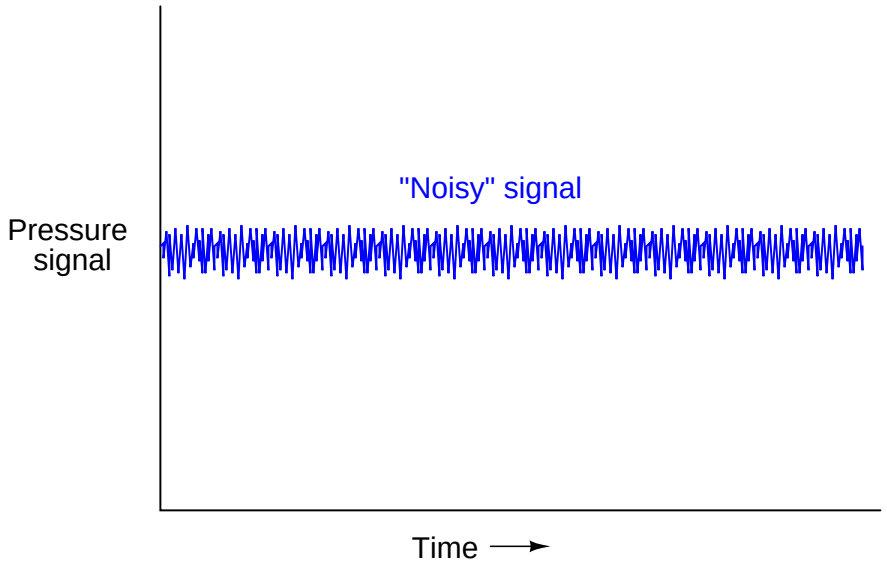
It should be noted that for most analog instruments, zero and span adjustments are *interactive*. That is, adjusting one has an effect on the other. Specifically, changes made to the span adjustment almost always alter the instrument's zero point[1]. An instrument with interactive zero and span adjustments requires much more effort to accurately calibrate, as one must switch back and forth between the lower- and upper-range points repeatedly to adjust for accuracy.

[1] It is actually quite rare to find an instrument where a change to the zero adjustment affects the instrument's span.

17.3 Damping adjustments

The vast majority of modern process transmitters (both analog and digital) come equipped with a feature known as *damping*. This feature is essentially a low-pass filter function placed in-line with the signal, reducing the amount of process "noise" reported by the transmitter.

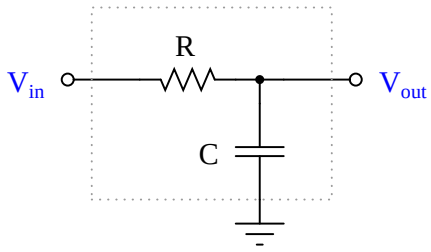
Imagine a pressure transmitter sensing water pressure at the outlet of a large pump. The flow of water exiting a pump tends to be extremely turbulent, and any pressure-sensing device connected to the immediate discharge port of a pump will interpret this turbulence as violent fluctuations in pressure. This means the pressure signal output by the transmitter will fluctuate as well, causing any indicator or control system connected to that transmitter to register a very "noisy" water pressure:



Such "noise" wreaks havoc with most forms of feedback control, since the control system will interpret these rapid fluctuations as real pressure changes requiring corrective action. Although it is possible to configure some control systems to ignore such noise, the best solution is to correct the problem at the source either by relocating the pressure transmitter's impulse line tap to a place where it does not sense as great an amount of fluid turbulence, or somehow prevent that sensed turbulence from being represented in the transmitter's signal.

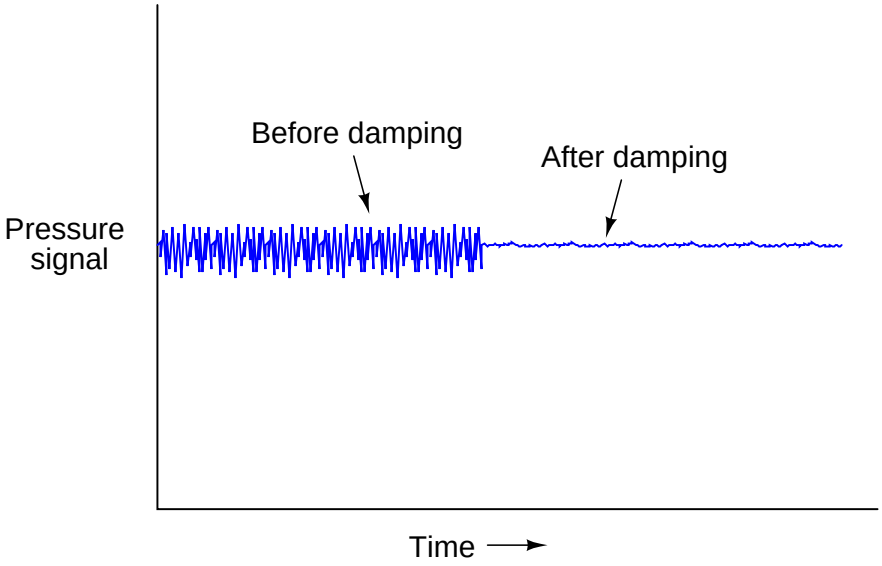
Since this noise is of a much greater frequency than the normal cycles of pressure in a process system, it is relatively easy to reduce the amount of noise in the transmitter signal simply by filtering that electronic signal using a low-pass filter circuit.

The simplest low-pass filter circuit is nothing more than a resistor and capacitor:



Low-frequency voltage signals applied to this circuit emerge at the output terminal relatively unattenuated, because the reactance of the capacitor is quite large at low frequencies. High-frequency signals applied to the same circuit become attenuated by the capacitor, which tends to "short" those signals to ground with its low reactance to high frequencies. The performance of such a filter circuit is primarily characterized by its *cutoff frequency*, mathematically defined as $f = \frac{1}{2\pi RC}$. The cutoff frequency is the point at which only 70.7% of the input signal appears at the output (a -3 dB attenuation in voltage).

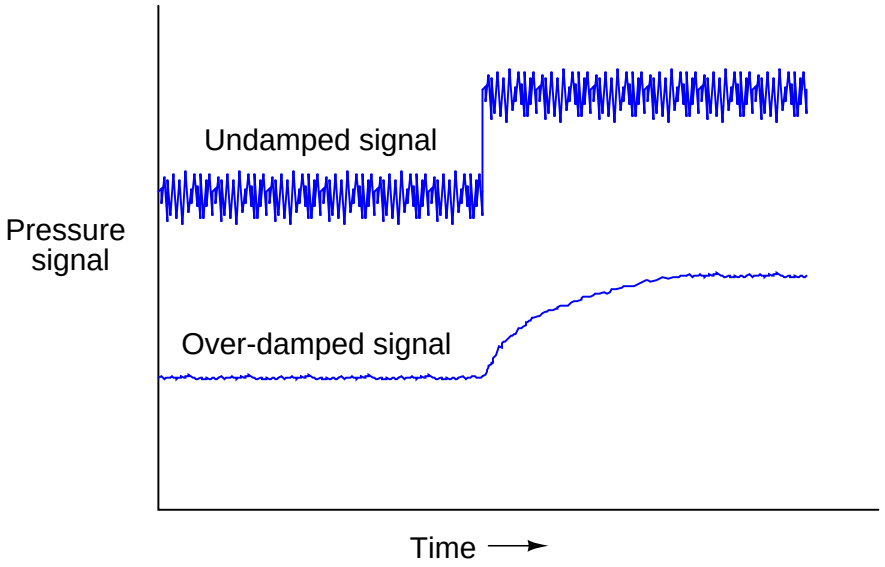
If successfully applied to a process transmitter, such low-pass filtering has the effect of "quieting" an otherwise noisy signal so only the real process pressure changes are seen, while the effect of turbulence (or whatever else was causing the noise) becomes minimal. In the world of process control, the intentional low-pass filtering of process measurement signals is often referred to as *damping* because its effect is to "damp" (turn down) the effects of process noise:



In order for damping to be a useful tool for the technician in mitigating measurement noise, it must be adjustable. In the case of the RC filter circuit, the degree of damping (cutoff frequency) may be adjusted by changing the value or either R or C, with R being the easier component to adjust. In digital transmitters where the damping is performed by a digital algorithm (either a sophisticated

digital filtering routine or something as simple as successive averaging of buffered signal values in a first-in-first-out shift register), damping may be adjusted by setting a constant value. In pneumatic transmitters, damping could be implemented by installing viscous elements to the mechanism, or more simply by adding volume to the signal line (e.g. excess tubing length, larger tubing diameter, or even "capacity tanks" connected to the tube for increased volume).

The key question for the technician then becomes, "how much damping do I use?" Insufficient damping will allow too much noise to reach the control system (causing "noisy" trends, indications, and erratic control), while excessive damping will cause the transmitter to understate the significance of sudden (real) process changes. In my experience there is a bad tendency for instrument technicians to apply excessive damping in transmitters. A transmitter with too much damping (i.e. cutoff frequency set too low, or time constant value set too high) causes the trend graph to be very smooth, which at first appears to be a good thing. After all, the whole point of a control system is to hold the process variable tightly to setpoint, so the appearance of a "flat line" process variable trend is enticing indeed. However, the problem with excessive damping is that the transmitter gives a sluggish response to any sudden changes in the real process variable. A dual-trend graph of a pressure transmitter experiencing a sudden increase in process pressure shows this principle, where the undamped transmitter signal is shown in the upper portion and the over-damped signal in the lower portion (please note the vertical offset between these two trends is shown only for your convenience in comparing the two trend shapes):



In summary, excessive damping causes the transmitter to "lie" to the control system by reporting a pressure that changes much slower than it actually does. The degree to which this "lie" adversely affects the control system (and/or the human operator's judgment in manually responding to the change in pressure) depends greatly on the nature of the control system and its importance to the overall plant operation. If any rule may be given as to how much damping to use in any transmitter, it is this: use as *little* as necessary to achieve good control.

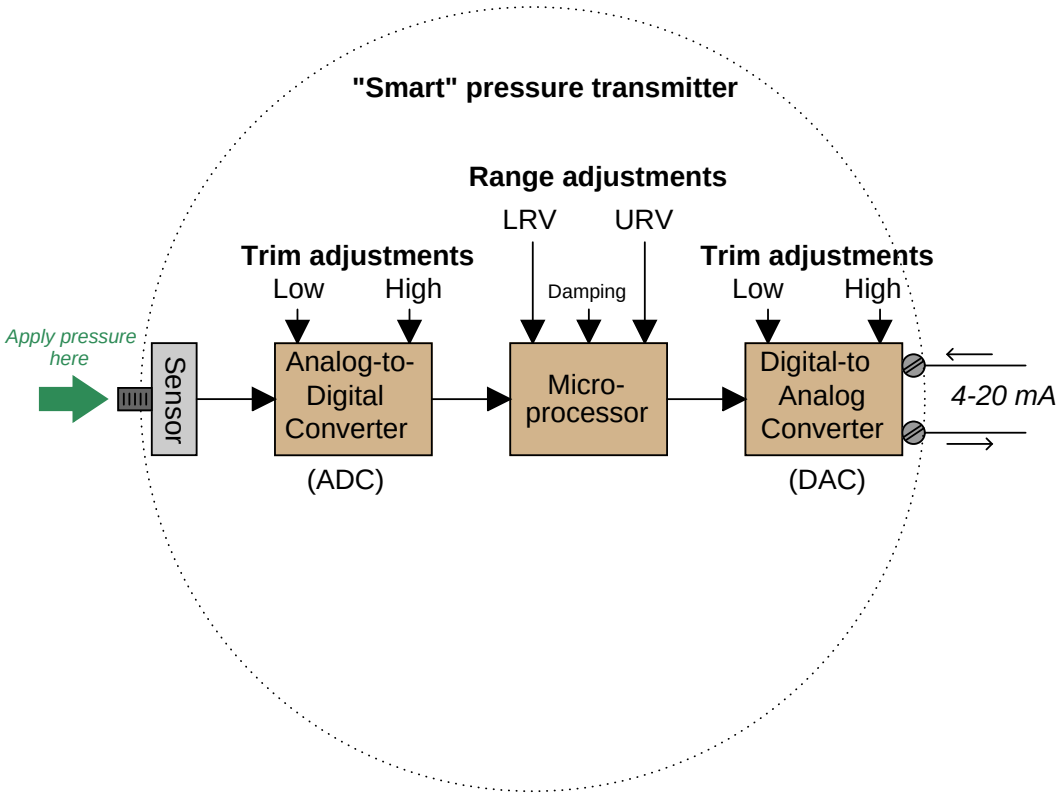
When calibrating a transmitter in a shop environment, the damping adjustment should be set to

its absolute minimum, so the results of applying stimuli to the transmitter are immediately seen by the technician. Any amount of damping in a transmitter being calibrated serves only to slow down the calibration procedure without benefit.

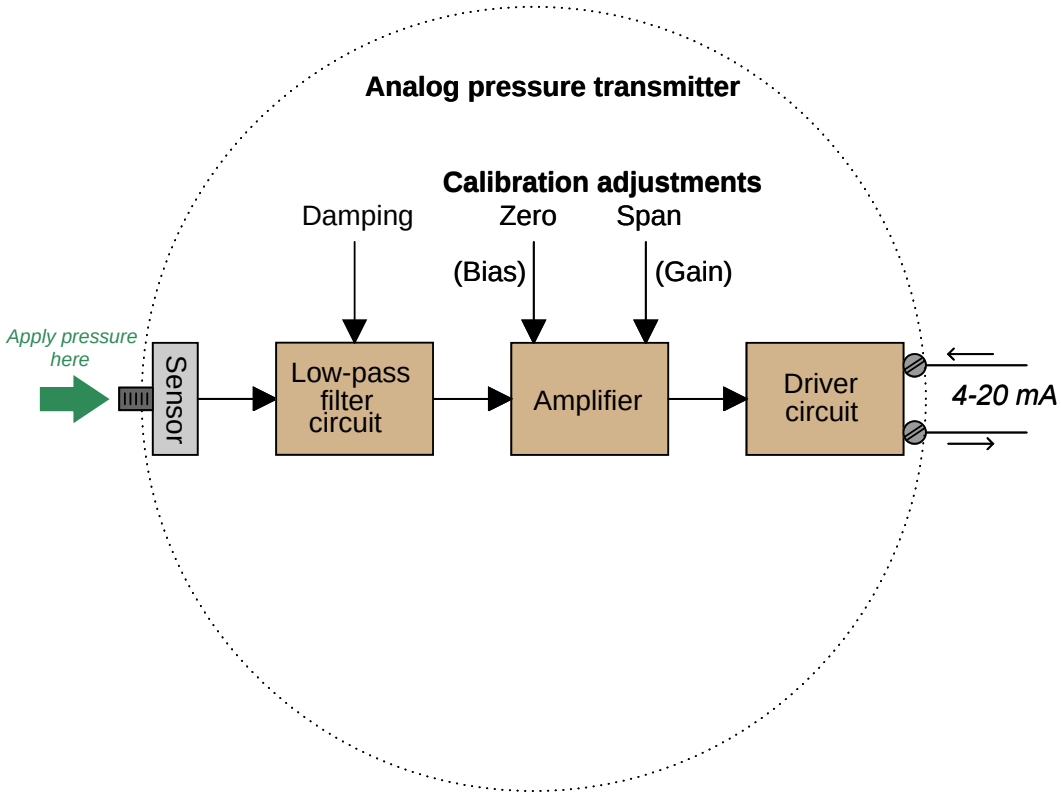
17.4 LRV and URV settings, digital trim (digital transmitters)

The advent of "smart" field instruments containing microprocessors has been a great advance for industrial instrumentation. These devices have built-in diagnostic ability, greater accuracy (due to digital compensation of sensor nonlinearities), and the ability to communicate digitally with host devices for reporting of various parameters.

A simplified block diagram of a "smart" pressure transmitter looks something like this:



It is important to note all the adjustments within this device, and how this compares to the relative simplicity of an all-analog pressure transmitter:



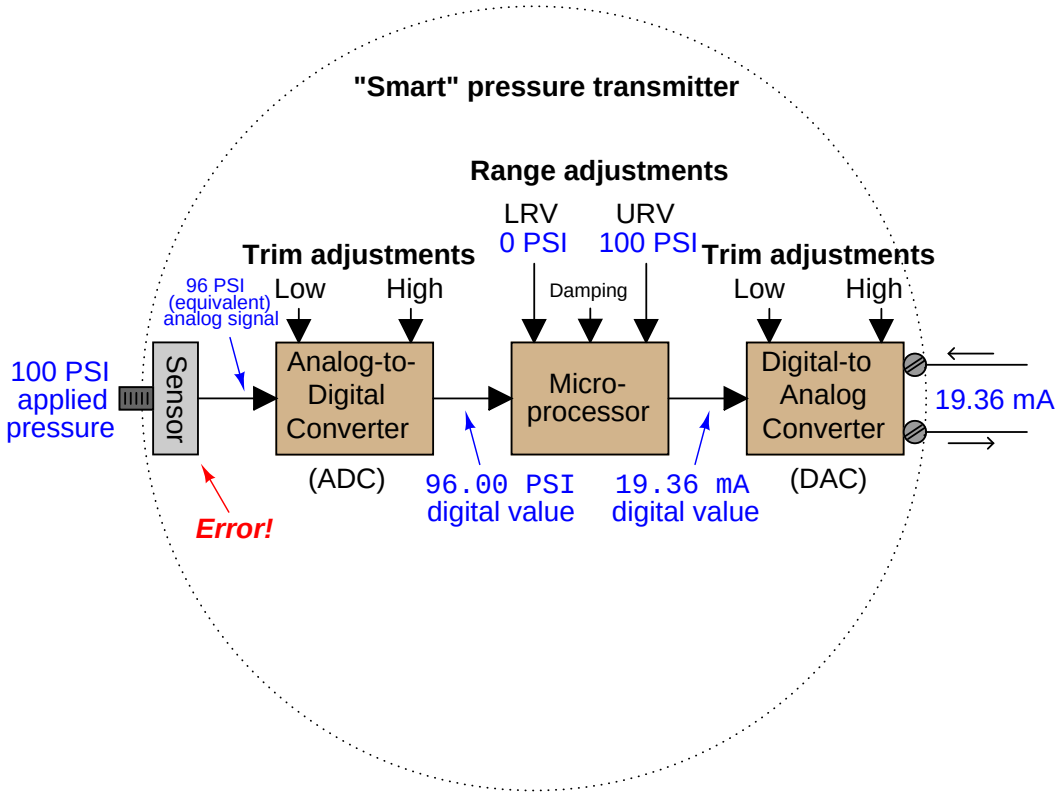
Note how the only calibration adjustments available in the analog transmitter are the "zero" and "span" settings. This is clearly not the case with smart transmitters. Not only can we set lower- and upper-range values (LRV and URV) in a smart transmitter, but it is also possible to calibrate the analog-to-digital and digital-to-analog converter circuits independently of each other. What this means for the calibration technician is that a full calibration procedure on a smart transmitter potentially requires more work and a greater number of adjustments than an all-analog transmitter[2].

A common mistake made among students and experienced technicians alike is to confuse the range settings (LRV and URV) for actual calibration adjustments. Just because you digitally set the LRV of a pressure transmitter to 0.00 PSI and the URV to 100.00 PSI does not necessarily mean it will register accurately at points within that range! The following example will illustrate this fallacy.

Suppose we have a smart pressure transmitter ranged for 0 to 100 PSI with an analog output range of 4 to 20 mA, but this transmitter's pressure sensor is fatigued from years of use such that an

[2] Although those adjustments made on a digital transmitter tend to be easier to perform than repeated zero-and-span adjustments on analog transmitters due to the inevitable "interaction" between analog zero and span adjustments requiring repeated checking and re-adjustment during the calibration period.

actual applied pressure of 100 PSI generates a signal that the analog-to-digital converter interprets as only 96 PSI[3]. Assuming everything else in the transmitter is in perfect condition, with perfect calibration, the output signal will still be in error:



As the saying goes, "a chain is only as strong as its weakest link." Here we see how the calibration of the most sophisticated pressure transmitter may be corrupted despite perfect calibration of both analog/digital converter circuits, and perfect range settings in the microprocessor. The microprocessor "thinks" the applied pressure is only 96 PSI, and it responds accordingly with a 19.36 mA output signal. *The only way anyone would ever know this transmitter was inaccurate at 100 PSI is to actually apply a known value of 100 PSI fluid pressure to the sensor and note the incorrect response.* The lesson here should be clear: digitally setting a smart instrument's LRV and URV points does *not* constitute a legitimate calibration of the instrument.

For this reason, smart instruments always provide a means to perform what is called a *digital trim* on both the ADC and DAC circuits, to ensure the microprocessor "sees" the correct representation of the applied stimulus and to ensure the microprocessor's output signal gets accurately converted into a DC current, respectively.

[3] A 4% calibration error caused by sensor aging is enormous for any modern digital transmitter, and should be understood as an exaggeration presented only for the sake of illustrating how sensor error affects overall calibration in a smart transmitter. A more realistic amount of sensor error due to aging would be expressed in small fractions of a percent.

I have witnessed some technicians use the LRV and URV settings in a manner not unlike the zero and span adjustments on an analog transmitter to correct errors such as this. Following this methodology, we would have to set the URV of the fatigued transmitter to 96 PSI instead of 100 PSI, so an applied pressure of 100 PSI would give us the 20 mA output signal we desire. In other words, we would let the microprocessor "think" it was only seeing 96 PSI, then skew the URV so it outputs the correct signal anyway. Such an approach will work to an extent, but any digital queries to the transmitter (e.g. using a digital-over-analog protocol such as HART) will result in conflicting information, as the current signal represents full scale (100 PSI) while the digital register inside the transmitter shows 96 PSI. The only comprehensive solution to this problem is to "trim" the analog-to-digital converter so the transmitter's microprocessor "knows" the actual pressure value applied to the sensor.

Once digital trims have been performed on both input and output converters, of course, the technician is free to re-range the microprocessor as many times as desired without re-calibration. This capability is particularly useful when re-ranging is desired for special conditions, such as process start-up and shut-down when certain process variables drift into uncommon regions. An instrument technician may use a hand-held digital "communicator" device to re-set the LRV and URV range values to whatever new values are desired by operations staff without having to re-check calibration by applying known physical stimuli to the instrument. So long as the ADC and DAC trims are both fine, the overall accuracy of the instrument will still be good with the new range. With analog instruments, the only way to switch to a different measurement range was to change the zero and span adjustments, which *necessitated* the re-application of physical stimuli to the device (a full re-calibration). Here and here alone we see where calibration is not necessary for a smart instrument. If overall measurement accuracy must be verified, however, there is no substitute for an actual physical calibration, and this entails both ADC and DAC "trim" procedures for a smart instrument.

Completely digital ("Fieldbus") transmitters are similar to "smart" analog-output transmitters with respect to distinct trim and range adjustments. For an explanation of calibration and ranging on FOUNDATION Fieldbus transmitters, refer to section 16.5.3 beginning on page 718.

17.5 Calibration procedures

As described earlier in this chapter, *calibration* refers to the adjustment of an instrument so its output accurately corresponds to its input throughout a specified range. This definition specifies the outcome of a calibration process, but not the procedure. It is the purpose of this section to describe procedures for efficiently calibrating different types of instruments.

17.5.1 Linear instruments

The simplest calibration procedure for an analog, linear instrument is the so-called *zero-and-span* method. The method is as follows:

1. Apply the lower-range value stimulus to the instrument, wait for it to stabilize
2. Move the "zero" adjustment until the instrument registers accurately at this point
3. Apply the upper-range value stimulus to the instrument, wait for it to stabilize
4. Move the "span" adjustment until the instrument registers accurately at this point
5. Repeat steps 1 through 4 as necessary to achieve good accuracy at both ends of the range

An improvement over this crude procedure is to check the instrument's response at several points between the lower- and upper-range values. A common example of this is the so-called *five-point calibration* where the instrument is checked at 0% (LRV), 25%, 50%, 75%, and 100% (URV) of range. A variation on this theme is to check at the five points of 10%, 25%, 50%, 75%, and 90%, while still making zero and span adjustments at 0% and 100%. Regardless of the specific percentage points chosen for checking, the goal is to ensure that we achieve (at least) the minimum necessary accuracy at all points along the scale, so the instrument's response may be trusted when placed into service.

Yet another improvement over the basic five-point test is to check the instrument's response at five calibration points *decreasing* as well as *increasing*. Such tests are often referred to as *Up-down* calibrations. The purpose of such a test is to determine if the instrument has any significant *hysteresis*: a lack of responsiveness to a change in direction.

Some analog instruments provide a means to adjust linearity. This adjustment should be moved only if absolutely necessary! Quite often, these linearity adjustments are very sensitive, and prone to over-adjustment by zealous fingers. The linearity adjustment of an instrument should be changed only if the required accuracy cannot be achieved across the full range of the instrument. Otherwise, it is advisable to adjust the zero and span controls to "split" the error between the highest and lowest points on the scale, and leave linearity alone.

The procedure for calibrating a "smart" digital transmitter – also known as *trimming* – is a bit different. Unlike the zero and span adjustments of an analog instrument, the "low" and "high" trim functions of a digital instrument are typically non-interactive. This means you should only have to apply the low- and high-level stimuli *once* during a calibration procedure. Trimming the sensor of a "smart" instrument consists of these four general steps:

1. Apply the lower-range value stimulus to the instrument, wait for it to stabilize
2. Execute the "low" sensor trim function
3. Apply the upper-range value stimulus to the instrument, wait for it to stabilize
4. Execute the "high" sensor trim function

Likewise, trimming the output (Digital-to-Analog Converter, or DAC) of a "smart" instrument consists of these six general steps:

1. Execute the "low" output trim test function
2. Measure the output signal with a precision milliammeter, noting the value after it stabilizes
3. Enter this measured current value when prompted by the instrument
4. Execute the "high" output trim test function
5. Measure the output signal with a precision milliammeter, noting the value after it stabilizes
6. Enter this measured current value when prompted by the instrument

After both the input and output (ADC and DAC) of a smart transmitter have been trimmed (i.e. calibrated against standard references known to be accurate), the lower- and upper-range values may be set. In fact, once the trim procedures are complete, the transmitter may be ranged and ranged again as many times as desired. The only reason for re-trimming a smart transmitter is to ensure accuracy over long periods of time where the sensor and/or the converter circuitry may have drifted out of acceptable limits. This stands in stark contrast to analog transmitter technology, where re-ranging *necessitates* re-calibration.

17.5.2 Nonlinear instruments

The calibration of inherently nonlinear instruments is much more challenging than for linear instruments. No longer are two adjustments (zero and span) sufficient, because more than two points are necessary to define a curve.

Examples of nonlinear instruments include expanded-scale electrical meters, square root characterizers, and position-characterized control valves.

Every nonlinear instrument will have its own recommended calibration procedure, so I will defer you to the manufacturer's literature for your specific instrument. I will, however, offer one piece of advice. When calibrating a nonlinear instrument, document all the adjustments you make (e.g. how many turns on each calibration screw) just in case you find the need to "re-set" the instrument back to its original condition. More than once I have struggled to calibrate a nonlinear instrument only to find myself further away from good calibration than where I originally started. In times like these, it is good to know you can always reverse your steps and start over!

17.5.3 Discrete instruments

The word "discrete" means *individual* or *distinct*. In engineering, a "discrete" variable or measurement refers to a true-or-false condition. Thus, a discrete sensor is one that is only able to indicate whether the measured variable is above or below a specified setpoint.

Examples of discrete instruments are *process switches* designed to turn on and off at certain values. A pressure switch, for example, used to turn an air compressor on if the air pressure ever falls below 85 PSI, is an example of a discrete instrument.

Discrete instruments require periodic calibration just like continuous instruments. Most discrete instruments have but one calibration adjustment: the *set-point* or *trip-point*. Some process switches have two adjustments: the set-point as well as a *deadband* adjustment. The purpose of a deadband adjustment is to provide an adjustable buffer range that must be traversed before the switch changes state. To use our 85 PSI low air pressure switch as an example, the set-point would be 85 PSI, but if the deadband were 5 PSI it would mean the switch would not change state until the pressure rose above 90 PSI (85 PSI + 5 PSI).

When calibrating a discrete instrument, you must be sure to check the accuracy of the set-point *in the proper direction of stimulus change*. For our air pressure switch example, this would mean checking to see that the switch changed states at 85 PSI *falling*, not 85 PSI *rising*. If it were not for the existence of deadband, it would not matter which way the applied pressure changed during the calibration test. However, deadband will always be present in a discrete instrument, whether that deadband is adjustable or not. Given a deadband of 5 PSI for this example switch, the difference between verifying a change of state at 85 PSI falling versus 85 PSI rising would mean the difference between the air compressor turning on if the pressure fell below 85 PSI versus turning on if the pressure fell below 80 PSI.

A procedure to efficiently calibrate a discrete instrument without too many trial-and-error attempts is to set the stimulus at the desired value (e.g. 85 PSI for our hypothetical low-pressure switch) and then move the set-point adjustment in the *opposite* direction as the intended direction of the stimulus (in this case, *increasing* the set-point value until the switch changes states). The basis for this technique is the realization that most comparison mechanisms cannot tell the difference between a rising process variable and a falling setpoint (or visa-versa). Thus, a falling pressure may be simulated by a rising set-point adjustment. You should still perform an actual changing-stimulus test to ensure the instrument responds properly under realistic circumstances, but this "trick" will help you achieve good calibration in less time.

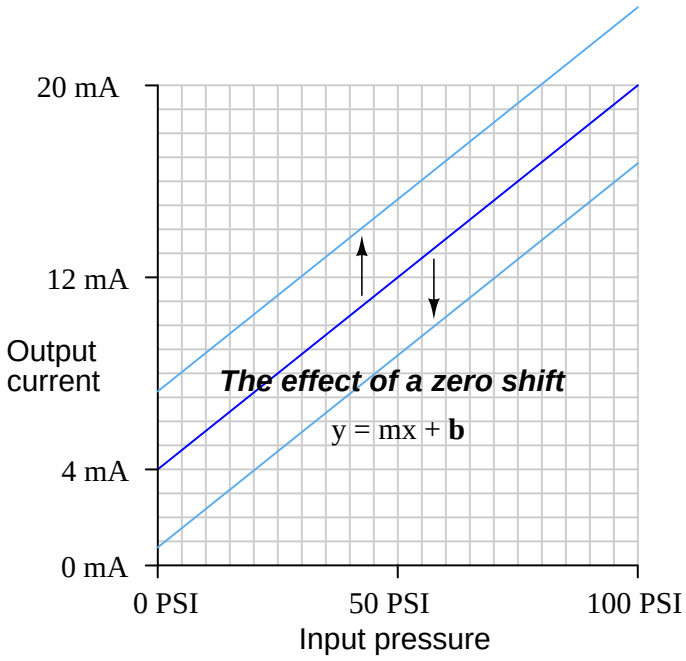
17.6 Typical calibration errors

Recall that the slope-intercept form of a linear equation describes the response of a linear instrument:

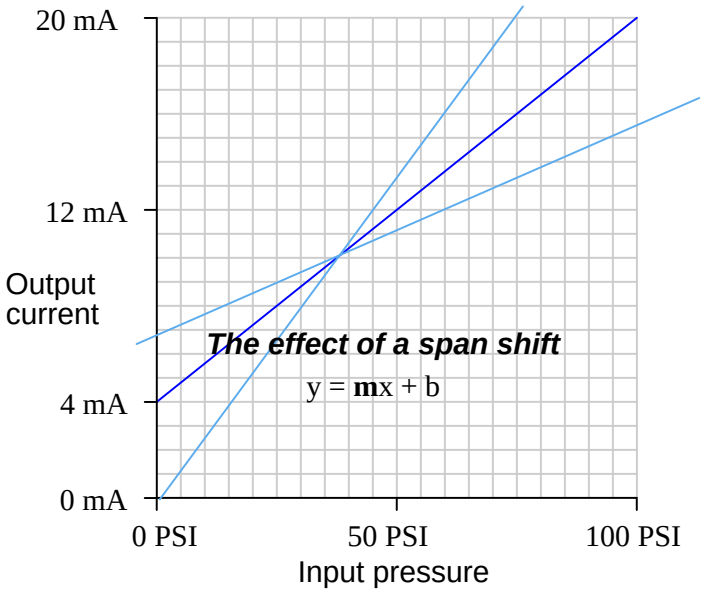
$$y = mx + b$$

Where,
y = Output
m = Span adjustment
x = Input
b = Zero adjustment

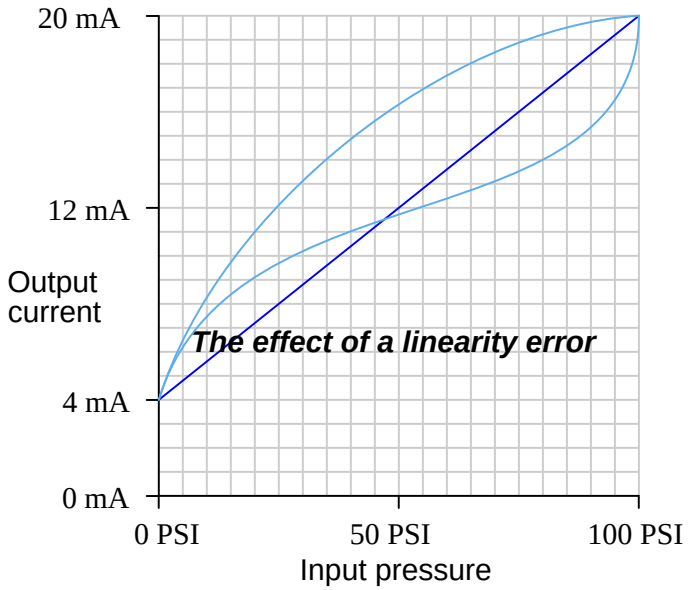
A *zero shift* calibration error shifts the function vertically on the graph. This error affects *all* calibration points equally, creating the same percentage of error across the entire range:



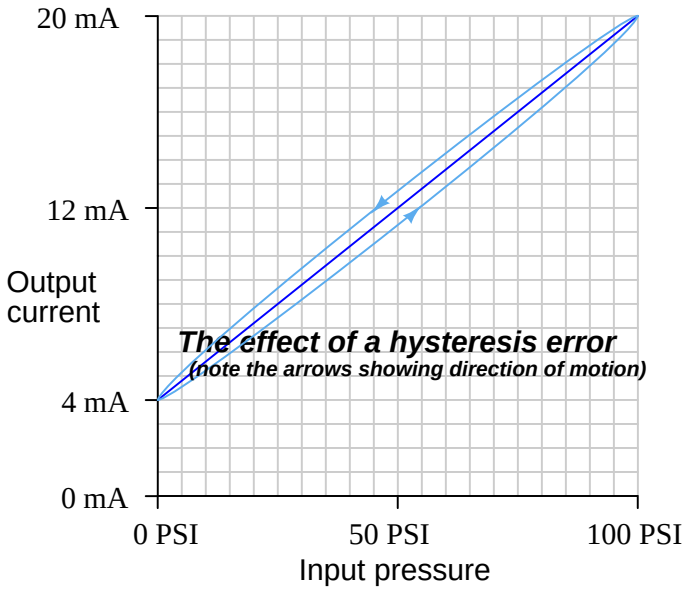
A *span shift* calibration error shifts the slope of the function. This error's effect is unequal at different points throughout the range:



A *linearity* calibration error causes the function to deviate from a straight line. This type of error does not directly relate to a shift in either zero (b) or span (m) because the slope-intercept equation only describes straight lines. If an instrument does not provide a linearity adjustment, the best you can do for this type of error is "split the error" between high and low extremes, so the maximum absolute error at any point in the range is minimized:



A *hysteresis* calibration error occurs when the instrument responds differently to an increasing input compared to a decreasing input. The only way to detect this type of error is to do an *up-down* calibration test, checking for instrument response at the same calibration points going down as going up:



Hysteresis errors are almost always caused by mechanical friction on some moving element (and/or a loose coupling between mechanical elements) such as bourdon tubes, bellows, diaphragms, pivots, levers, or gear sets. Flexible metal strips called *flexures* – which are designed to serve as frictionless pivot points in mechanical instruments – may also cause hysteresis errors if cracked or bent.

In practice, most calibration errors are some combination of zero, span, linearity, and hysteresis problems.

17.6.1 As-found and as-left documentation

An important principle in calibration practice is to document every instrument's calibration as it was found *and* as it was left after adjustments were made. The purpose for documenting both conditions is to make data available for calculating instrument *drift* over time. If only one of these conditions is documented during each calibration event, it will be difficult to determine how well an instrument is holding its calibration over long periods of time. Excessive drift is often an indicator of impending failure, which is vital for any program of predictive maintenance or quality control.

Typically, the format for documenting both As-Found and As-Left data is a simple table showing the points of calibration, the ideal instrument responses, the actual instrument responses, and the calculated error at each point. The following table is an example for a pressure transmitter with a range of 0 to 200 PSI over a five-point scale:

Percent of range	Input pressure	Output current (ideal)	Output current (measured)	Error (percent of span)
0%	0 PSI	4.00 mA		
25%	50 PSI	8.00 mA		
50%	100 PSI	12.00 mA		
75%	150 PSI	16.00 mA		
100%	200 PSI	20.00 mA		

17.6.2 Up-tests and Down-tests

It is not uncommon for calibration tables to show multiple calibration points going *up* as well as going *down*, for the purpose of documenting hysteresis and deadband errors. Note the following example, showing a transmitter with a maximum hysteresis of 0.313 % (the offending data points are shown in bold-faced type):

Percent of range	Input pressure	Output current (ideal)	Output current (measured)	Error (percent of span)
0%	0 PSI	4.00 mA	3.99 mA	-0.0625 %
25% ↑	50 PSI	8.00 mA	**7.98 mA**	**-0.125 %**
50% ↑	100 PSI	12.00 mA	11.99 mA	-0.0625 %
75% ↑	150 PSI	16.00 mA	15.99 mA	-0.0625 %
100% ↑	200 PSI	20.00 mA	20.00 mA	0 %
75% ↓	150 PSI	16.00 mA	16.01 mA	+0.0625 %
50% ↓	100 PSI	12.00 mA	12.02 mA	+0.125 %
25% ↓	50 PSI	8.00 mA	**8.03 mA**	**+0.188 %**
0% ↓	0 PSI	4.00 mA	4.01 mA	+0.0625 %

In the course of performing such a directional calibration test, it is important not to overshoot any of the test points. If you do happen to overshoot a test point in setting up one of the input conditions for the instrument, simply "back up" the test stimulus and re-approach the test point from the same direction as before. Unless each test point's value is approached from the proper direction, the data cannot be used to determine hysteresis/deadband error.

17.7 NIST traceability

As defined previously, *calibration* means the comparison and adjustment (if necessary) of an instrument's response to a stimulus of precisely known quantity, to ensure operational accuracy. In order to perform a calibration, one must be reasonably sure that the physical quantity used to stimulate the instrument is accurate in itself. For example, if I try calibrating a pressure gauge to read accurately at an applied pressure of 200 PSI, I must be reasonably sure that the pressure I am using to stimulate the gauge is actually 200 PSI. If it is not 200 PSI, then all I am doing is adjusting the pressure gauge to register 200 PSI when in fact it is sensing something different.

Ultimately, this is a philosophical question of *epistemology*: how do we know what is true? There are no easy answers here, but teams of scientists and engineers known as *metrologists* devote their professional lives to the study of calibration standards to ensure we have access to the best approximation of "truth" for our calibration purposes. *Metrology* is the science of measurement, and the central repository of expertise on this science within the United States of America is the *National Institute of Standards and Technology*, or the *NIST* (formerly known as the *National Bureau of Standards*, or *NBS*).

Experts at the NIST work to ensure we have means of tracing measurement accuracy back to *intrinsic standards*, which are quantities inherently fixed (as far as anyone knows). The vibrational frequency of an isolated cesium atom when stimulated by radio energy, for example, is an intrinsic standard used for the measurement of time (forming the basis of the so-called *atomic clock*). So far as anyone knows, this frequency is fixed in nature and cannot vary. Intrinsic standards therefore serve as absolute references which we may calibrate certain instruments against.

The machinery necessary to replicate intrinsic standards for practical use are quite expensive and usually delicate. This means the average metrologist (let alone the average industrial instrument technician) simply will never have access to one. In order for these intrinsic standards to be useful within the industrial world, we use them to calibrate other instruments, which are used to calibrate other instruments, and so on until we arrive at the instrument we intend to calibrate for field service in a process. So long as this "chain" of instruments is calibrated against each other regularly enough to ensure good accuracy at the end-point, we may calibrate our field instruments with confidence. The documented confidence is known as *NIST traceability*: that the accuracy of the field instrument we calibrate is ultimately ensured by a trail of documentation leading to intrinsic standards maintained by the NIST.

17.8 Instrument turndown

An important performance parameter for transmitter instruments is something often referred to as *turndown* or *rangedown*. "Turndown" is defined as the ratio of maximum allowable span to the minimum allowable span for a particular instrument.

Suppose a pressure transmitter has a maximum calibration range of 0 to 300 pounds per square inch (PSI), and a turndown of 20:1. This means that a technician may adjust the span anywhere between 300 PSI and 15 PSI. This is important to know in order to select the proper transmitter for any given measurement application. The odds of you finding a transmitter with just the perfect factory-calibrated range for your measurement application may be quite small, meaning you will have to adjust its range to fit your needs. The turndown ratio tells you how far you will be able to practically adjust your instrument's range.

17.9 Practical calibration standards

Within the context of a calibration shop environment, where accurate calibrations are important yet intrinsic standards are not readily accessible, we must do what we can to maintain a workable degree of accuracy in the calibration equipment used to calibrate field instruments.

It is important that the degree of uncertainty in the accuracy of a test instrument is *significantly less* than the degree of uncertainty we hope to achieve in the instruments we calibrate. Otherwise, calibration becomes an exercise in futility. This ratio of uncertainties is called the *Test Uncertainty Ratio*, or *TUR*. A good rule-of-thumb is to maintain a TUR of at least 4:1 (ideally 10:1 or better), the test equipment being many times more accurate (less uncertain) than the field instruments we calibrate with them.

I have personally witnessed the confusion and wasted time that results from trying to calibrate a field instrument to a tighter tolerance than what the calibrating equipment is capable of. In one case, an instrument technician attempted to calibrate a pneumatic pressure transmitter to a tolerance of +/- 0.5% of span using a test gauge that was only good for +/- 1% of the same span. This poor technician kept going back and forth, adjusting zero and span over and over again, trying to stay within the stated specification of 0.5%. After giving up, he tested the test gauges by comparing three of them, one against the other. When it was realized no two test gauges would agree with each other to within the tolerance he was trying to achieve in calibrating the transmitter, it became clear what the problem was.

The lesson to be learned here is to always ensure the equipment used to calibrate industrial instruments is reliably accurate (enough). No piece of test equipment will ever be *perfectly* accurate, but perfection is not what we need. Our goal is to be *accurate enough* that the final calibration will be reliable within specified boundaries.

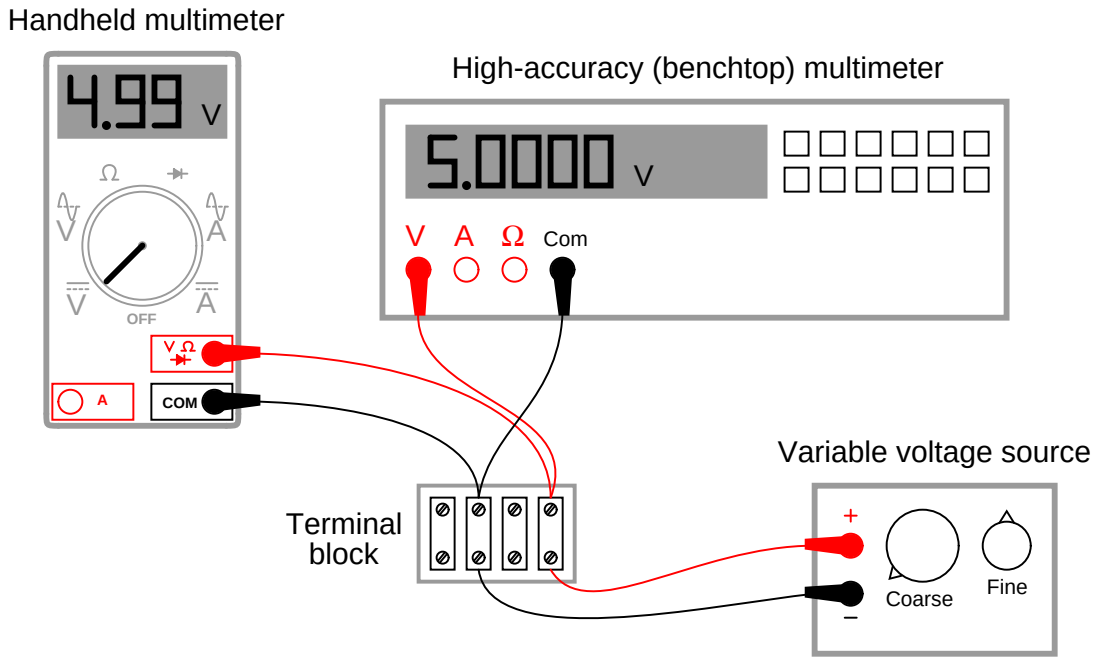
The next few subsections describe various standards used in instrument shops to calibrate industrial instruments.

17.9.1 Electrical standards

Electrical calibration equipment – used to calibrate instruments measuring voltage, current, and resistance – must be periodically calibrated against higher-tier standards maintained by outside laboratories. In years past, instrument shops would often maintain their own *standard cell* batteries (often called *Weston* cells) as a primary voltage reference. These special-purpose batteries produced 1.0183 volts DC at room temperature with low uncertainty and drift, but were sensitive to vibration and non-trivial to actually use. Now, electronic voltage references have all but displaced standard cells in calibration shops and laboratories, but these references must be checked and adjusted for drift in order to maintain their NIST traceability.

One enormous benefit of electronic calibration references is that they are able to generate accurate currents and resistances in addition to voltage (and not just voltage at one fixed value, either!). Modern electronic references are digitally-controlled as well, which lends themselves well to automated testing in assembly-line environments, and/or programmed multi-point calibrations with automatic documentation of as-found and as-left calibration data.

If a shop cannot afford one of these versatile references for benchtop calibration use, an acceptable alternative in some cases is to purchase a high-accuracy multimeter and equip the calibration bench with adjustable voltage, current, and resistance sources. These sources will be simultaneously connected to the high-accuracy multimeter and the instrument under test, and adjusted until the high-accuracy meter registers the desired value. The measurement shown by the instrument under test is then compared against the reference meter and adjusted until matching (to within the required tolerance). The following illustration shows how a high-accuracy voltmeter could be used to calibrate a handheld voltmeter in this fashion:



It should be noted that the variable voltage source shown in this test arrangement need not be sophisticated. It simply needs to be *variable* (to allow precise adjustment until the high-accuracy voltmeter registers the desired voltage value) and *stable* (so the adjustment will not drift appreciably over time).

17.9.2 Temperature standards

The most common technologies for industrial temperature measurement are electronic in nature: RTDs and thermocouples. As such, the standards used to calibrate such devices are the same standards used to calibrate electrical instruments such as digital multimeters (DMMs). For RTDs, this means a precision resistance standard such as a *decade box* used to precisely set known quantities of electrical resistance. For thermocouples, this means a *precision potentiometer* used to generate precise quantities of low DC voltage (in the millivolt range, with microvolt resolution). Modern, electronic calibrators are also available now for RTD and thermocouple instrument calibration, able to generate accurate quantities of electrical resistance and DC millivoltage for the simulation of RTD and thermocouple elements, respectively.

However, there are some temperature-measuring instruments that are not electrical in nature. This category includes bimetallic thermometers, filled-bulb temperature systems, and optical pyrometers. In order to calibrate these types of instruments, we must accurately create the calibration temperatures in the instrument shop. In other words, the instrument to be calibrated must be subjected to an actual temperature of accurately known value.

Even with RTDs and thermocouples – where the sensor signal may be easily simulated using electronic test equipment – there is merit in using an actual source of precise temperature to calibrate the temperature instrument. Simulating the voltage produced by a thermocouple at a precise temperature, for example, is fine for calibrating the instrument normally receiving the millivoltage signal from the thermocouple, but this calibration test does nothing to validate the accuracy of the thermocouple element itself! The *best* type of calibration for any temperature-measuring instrument, from the perspective of overall integrity, is to actually subject the sensing element to a precisely known temperature. For this we need special calibration equipment designed to produce accurate temperature samples on demand.

A time-honored standard for low-temperature industrial calibrations is pure water, specifically the freezing and boiling points of water. Pure water at sea level (full atmospheric pressure) freezes at 32 degrees Fahrenheit (0 degrees Celsius) and boils at 212 degrees Fahrenheit (100 degrees Celsius). In fact, the Celsius temperature scale is *defined* by these two points of phase change for water at sea level[4].

To use water as a temperature calibration standard, simply prepare a vessel for one of two conditions: thermal equilibrium at freezing or thermal equilibrium at boiling. "Thermal equilibrium" in this context simply means equal temperature throughout the mixed-phase sample. In the case of freezing, this means a well-mixed sample of solid ice and liquid water. In the case of boiling, this means a pot of water at a steady boil (vaporous steam and liquid water in direct contact). What you are trying to achieve here is ample contact between the two phases (either solid and liquid; or liquid and vapor) to eliminate hot or cold spots. When the entire water sample is homogeneous in temperature and changing phase (either freezing or boiling), the sample will have only one degree of thermodynamic freedom: its temperature is an exclusive function of atmospheric pressure. Since atmospheric pressure is relatively stable and well-known, this fixes the temperature at a constant value. For ultra-precise temperature calibrations in laboratories, the *triple point* of water is used

[4]The Celsius scale used to be called the *Centigrade* scale, which literally means "100 steps." I personally prefer "Centigrade" to "Celsius" because it actually describes something about the unit of measurement. In the same vein, I also prefer the older label "Cycles Per Second" (cps) to "Hertz" as the unit of measurement for frequency. You may have noticed by now that the instrumentation world does not yield to my opinions, much to my chagrin.

as the reference. When water is brought to its triple point, the sample will have *zero degrees* of thermodynamic freedom, which means both its temperature and its pressure will become locked at stable values: pressure at 0.006 atmospheres, and temperature at 0.01 degrees Celsius.

The major limitation of water as a temperature calibration standard is it only provides two points of calibration: 0 ^{o}C and 100 ^{o}C, with the latter[5] being strongly pressure-dependent. If other reference temperatures are required for a calibration, some substance other than water must be used.

A variety of substances with known phase-change points have been standardized as fixed points on the International Practical Temperature Scale (ITS-90). The following list is a sample of some of these substances and their respective phase states and temperatures[6]:

- Neon (triple point) = -248.6 ^{o}C
- Oxygen (triple point) = -218.8 ^{o}C
- Mercury (triple point) = -38.83 ^{o}C
- Tin (freezing point) = 231.93 ^{o}C
- Zinc (freezing point) = 419.53 ^{o}C
- Aluminum (freezing point) = 660.32 ^{o}C
- Copper (freezing point) = 1084.62 ^{o}C

Substances at the triple point must be in thermal equilibrium with solid, liquid, and vaporous phases co-existing. Substances at the freezing point must be a two-phase mixture of solid and liquid (i.e. a liquid *in the process of freezing*, neither a completely liquid nor a completely solid sample). The physical principle at work in all of these examples is that of *latent heat*: the thermal energy exchange required to change the phase of a substance. So long as the minimum heat exchange requirement for complete phase change is not met, a substance in the midst of phase transition will exhibit a fixed temperature, and therefore behave as a temperature *standard*. Small amounts of heat gain or loss to such a sample will merely change the proportion of one phase to another (e.g. how much solid versus how much liquid), but the temperature will remain locked at a constant value until the sample becomes a single phase.

One major disadvantage of using phase changes to produce accurate temperatures in the shop is the limited availability of temperatures. If you need to create some other temperature for calibration purposes, you either need to find a suitable material with a phase change happening at that exact same temperature (good luck!) or you need to find a finely adjustable temperature source and use an accurate thermometer to compare your instrument under test against. The latter scenario is

[5] Pressure does have some influence on the freezing point of most substances as well, but not nearly to the degree it has on the boiling point. For a comparison between the pressure-dependence of freezing versus boiling points, consult a phase diagram for the substance in question, and observe the slopes of the solid-liquid phase line and liquid-vapor phase line. A nearly-vertical solid-liquid phase line shows a weak pressure dependence, while the liquid-vapor phase lines are typically much closer to horizontal.

[6] For each of these examples, the assumptions of a 100% pure sample and an airless testing environment are made. Impurities in the initial sample and/or resulting from chemical reactions with air at elevated temperatures, may introduce serious errors.

analogous to the use of a high-accuracy voltmeter and an adjustable voltage source to calibrate a voltage instrument: comparing one instrument (trusted to be accurate) against another (under test).

Laboratory-grade thermometers are relatively easy to secure. Variable temperature sources suitable for calibration use include *oil bath* and *sand bath* calibrators. These devices are exactly what they sound like: small pots filled with either oil or sand, containing an electric heating element and a temperature control system using a laboratory-grade (NIST-traceable) thermal sensor. In the case of sand baths, a small amount of compressed air is introduced at the bottom of the vessel to "fluidize" the sand so the grains move around much like the molecules of a liquid, helping the system reach thermal equilibrium. To use a bath-type calibrator, place the temperature instrument to be calibrated such the sensing element dips into the bath, then wait for the bath to reach the desired temperature.

An oil bath temperature calibrator is shown in the following photograph, with sockets to accept seven temperature probes into the heated oil reservoir:

Dry-block temperature calibrators also exist for creating accurate calibration temperatures in the instrument shop environment. Instead of a fluid (or fluidized powder) bath as the thermal medium, these devices use metal blocks with blind (dead-end) holes drilled for the insertion of temperature-sensing instruments.

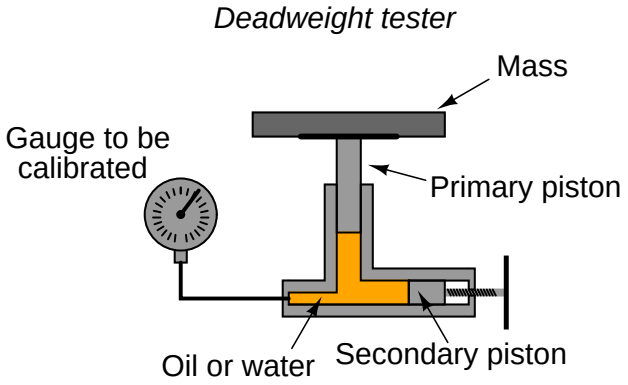
An inexpensive dry-block temperature calibrator intended for bench-top service is shown in this photograph:

Optical temperature instruments require a different sort of calibration tool: one that emits radiation equivalent to that of the process object at certain specified temperatures. This type of calibration tool is called a *blackbody calibrator*, having a target area where the optical instrument may be aimed. Like oil and sand bath calibrators, a blackbody calibrator relies on an internal temperature sensing element as a reference, to control the optical emissions of the blackbody target at any specified temperature within a practical range.

17.9.3 Pressure standards

In order to accurately calibrate a pressure instrument in a shop environment, we must create fluid pressures of known magnitude against which we compare the instrument being calibrated. As with other types of physical calibrations, our choices of instruments falls into two broad categories: devices that inherently *produce* known pressures versus devices that accurately measure pressures created by some (other) adjustable source.

A *deadweight tester* (sometimes referred to as a *dead-test* calibrator) is an example in the former category. These devices *create* accurately known pressures by means of precise masses and pistons of precise area:



After connecting the gauge (or other pressure instrument) to be calibrated, the technician adjusts the secondary piston to cause the primary piston to lift off its resting position and be suspended by oil pressure alone. So long as the mass placed on the primary piston is precisely known, Earth's gravitational field is constant, and the piston is perfectly vertical, the fluid pressure applied to the instrument under test *must* be equal to the value described by the following equation:

$$P = \frac{F}{A}$$

Where,

P = Fluid pressure

F = Force exerted by the action of gravity on the mass ($F_{weight} = mg$)

A = Area of piston

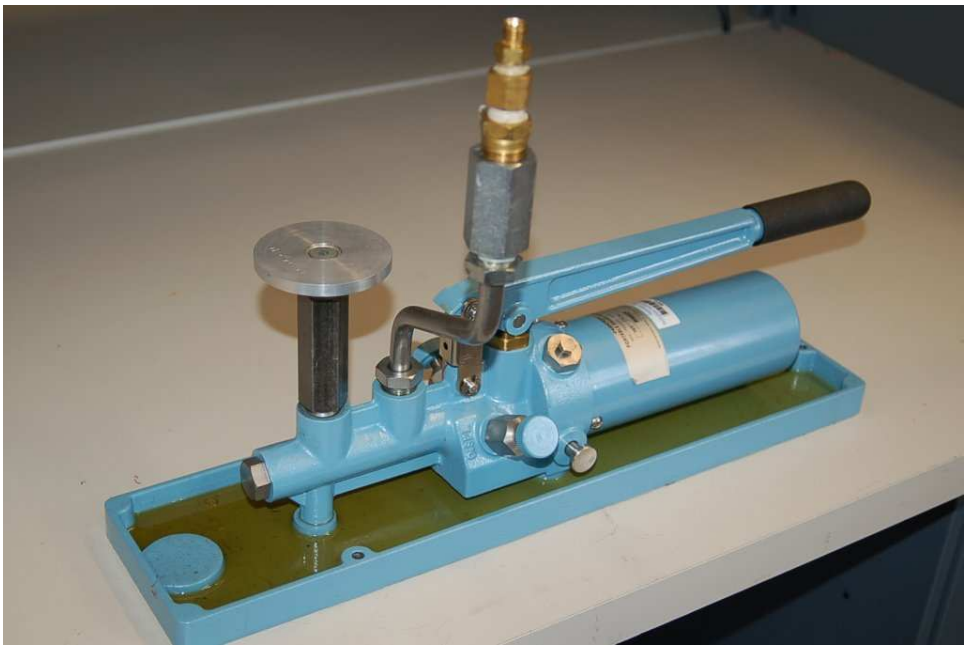
The primary piston area, of course, is precisely set at the time of the deadweight tester's manufacture and does not change appreciably throughout the life of the device.

A very simple deadweight tester unit appears in the next photograph, mounted to a yellow wooden base:

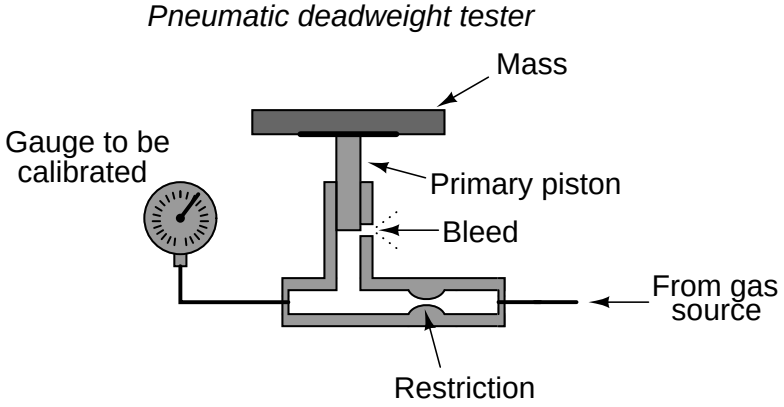
When sufficient pressure has been accumulated inside the tester to overcome the weight on the piston, the piston rises off its rest and "floats" on the pressurized oil, as shown in this close-up photograph:

A common operating practice for any deadweight tester is to gently spin the mass during testing so the primary piston continually rotates within its cylinder. Any motion will prevent static friction from taking hold, helping to ensure the only force on the primary piston is the force of the fluid within the deadweight tester.

Most modern deadweight testers include extra features such as hand pumps and bleed valves in addition to secondary pistons, to facilitate both rapid and precise operation. The next photograph shows a newer deadweight tester, with these extra features:



There is also such a thing as a *pneumatic* deadweight tester. In these devices, a constant flow of gas such as compressed air or bottled nitrogen vents through a bleed port operated by the primary piston. The piston moves as necessary to maintain just enough gas pressure inside the unit to suspend the mass(es) against gravity. This gas pressure passes on to the instrument under test, just as liquid pressure in a hydraulic deadweight tester passes to the test instrument for comparison:

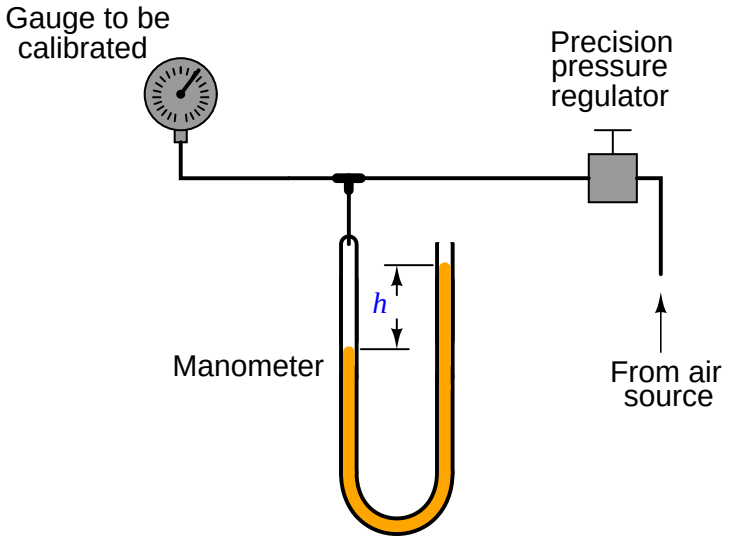


In fact, the construction and operation of a pneumatic deadweight tester is quite similar to a self-balancing (force-balance) pneumatic instrument mechanism with a baffle/nozzle assembly. A

moving element opens or closes a variable restriction downstream of a fixed restriction to generate a varying pressure. In this case, that pressure directly operates the bleed vent to self-regulate gas pressure at whatever value is necessary to suspend the mass against gravity.

Deadweight testers (both hydraulic and pneumatic) lend themselves well to relatively high pressures, owing to the practical limitations of mass and piston area. You could use a deadweight tester to calibrate a 100 PSI pressure gauge used for measuring water mains pressure, for example, but you could not use a deadweight tester to calibrate a 0 to 1 "W.C. (zero to one inch water column) pressure gauge used to measure draft pressure in a furnace flue.

For low-pressure calibrations, the simple *manometer* is a much more practical standard. Manometers, of course, do not generate pressure on their own. In order to use a manometer to calibrate a pressure instrument, you must connect both devices to a source of variable fluid pressure, typically instrument air through a precision pressure regulator:



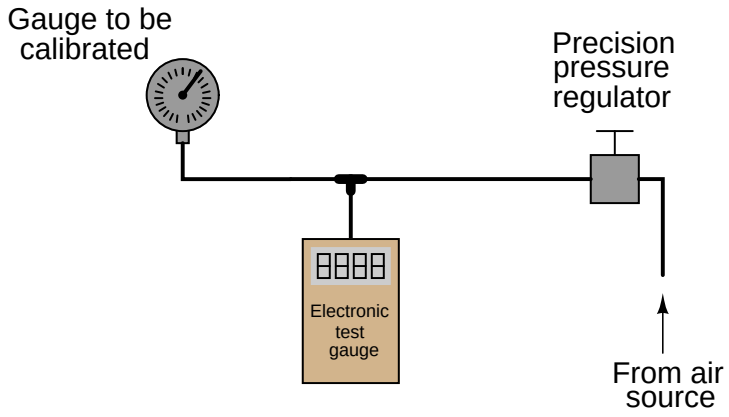
The difference in liquid column heights (h) within the manometer shows the pressure applied to the gauge. As with the deadweight tester, the accuracy of this pressure measurement is bound by just a few physical constants, none of which are liable to spurious change. So long as the manometer's liquid density is precisely known, Earth's gravitational field is constant, and the manometer tubes are perfectly vertical, the fluid pressure indicated by the manometer *must* be equal to the value described by the following equation (two different forms given):

$$P = \rho g h \qquad \text{(or)} \qquad P = \gamma h$$

Where,

P = Fluid pressure
ρ = Mass density of fluid
γ = Weight density of fluid
g = Acceleration of gravity
h = Height difference between manometer liquid columns

Of course, with pressure-measuring test instruments of suitable accuracy (preferably NIST-traceable), the same sort of calibration jig may be used for virtually any desired range of pressures:



When the electronic test gauge is designed for very low pressures (inches of water column), they are sometimes referred to as *electronic manometers*.

Instrument calibrations performed in the field (i.e. in locations near or at the intended point of use rather than in a professionally-equipped shop) are almost always done this way: a pressure-generating source is connected to both the instrument under test and a trusted calibration gauge ("test gauge"), and the two indications are compared at several points along the calibrated range. Test equipment suitable for field pressure calibrations include *slack-tube manometers* made from flexible plastic tubing hung from any available anchor point near eye level, and *test gauges* typically of the helical bourdon tube variety. Portable electronic test gauges are also available for field use, many with built-in hand pumps for generating precise air pressures.

A noteworthy example of a pneumatic pressure calibrator for field use was a device manufactured by the Wallace & Tiernan corporation, affectionately called a *Wally box* by at least one generation of instrument technicians. A "Wally box" consisted of a large dial pressure gauge (several inches in diameter) with a multi-turn needle and a very fine scale, connected to a network of valves and regulators which were used to set different air pressures from any common compressed air source. The entire mechanism was housed in an impact-resistance case for ruggedness. One of the many nice features of this calibration instrument was a selector valve allowing the technician to switch between two different pressures output by independent pressure regulators. Once the two pressure regulator values were set to the instrument's lower- and upper-range values (LRV and URV), it was possible to switch back and forth between those two pressures at will, making the task of adjusting an analog instrument with interactive zero and span adjustments much easier than it would have been to precisely adjust a single pressure regulator again and again.

17.9.4 Flow standards

Most forms of continuous flow measurement are inferential; that is, we measure flow indirectly by measuring some other variable (such as pressure, voltage, or frequency) directly. With this in mind, we may usually achieve reasonable calibration accuracy simply by calibrating the primary sensor and replacing the flow element (if inspection proves necessary). In the case of an orifice plate used to measure fluid flow rate, this would mean calibrating the differential pressure transmitter to measure pressure accurately and replacing the orifice plate if it shows signs of wear.

In some cases, though, direct validation of flow measurement accuracy is needed. Most techniques of flow rate validation take the form of measuring accumulated fluid volume over time. This may prove to be complicated, especially if the fluids in question are hazardous in any way, and/or the flow rates are large, and/or the fluid is a gas or vapor.

For simple validation of liquid flow rates, the flow may be diverted from its normal path in the process and into a container where either accumulated volume or accumulated weight may be measured over time. If the rate of flow into this container is constant, the accumulated volume (or weight) should increase linearly over time. The actual flow rate may then be calculated by dividing the change in volume (ΔV) by the time interval over which the change in volume was measured (Δt). The resulting quotient is the average flow rate between those two points in time, which is an approximation of instantaneous flow rate:

$$\frac{\Delta V}{\Delta t} = \text{ Average flow}$$

$$\frac{\Delta V}{\Delta t} \approx \frac{dV}{dt} = \text{ Instantaneous flow}$$

If a suitable vessel exists in the process with level-measuring capability (e.g. a liquid storage vessel equipped with a level transmitter), you may apply the same mathematical technique: use that vessel as an accumulator for the flow in question, tracking the accumulated (or lost) volume over time and then calculating $\frac{\Delta V}{\Delta t}$. The accuracy of this technique rests on some additional factors, though:

- The accuracy of the level transmitter (as a *volume* measuring instrument!)
- The ability to ensure only *one* flow path in or out of that vessel

The first condition listed here places significant limitations on the flow calibration accuracy one can achieve with this method. In essence, you are using the level instrument as the "test gauge" for the flow instrument, so it needs to be high-accuracy in order to achieve even reasonable accuracy for the flowmeter being calibrated.

A more sophisticated approach for direct flow validation is the use of a device called a *flow prover*. A "flow prover" is a precision piston-and-cylinder mechanism used to precisely measure a quantity of liquid over time. Process flow is diverted through the prover, moving the piston over time. Sensors on the prover mechanism detect when the piston has reached certain positions, and time measurements taken at those different positions enable the calculation of average flow ($\frac{\Delta V}{\Delta t}$).

17.9.5 Analytical standards

An *analyzer* measures intrinsic properties of a substance sample such as its density, chemical content, or purity. Whereas the other types of instruments discussed in this chapter measure quantities incidental to the composition of a substance (pressure, level, temperature, and flow rate), an analyzer measures something related to the *nature* of substance being processed.

As previously defined, to *calibrate* an instrument means to check and adjust (if necessary) its response so the output accurately corresponds to its input throughout a specified range. In order to do this, one must expose the instrument to an actual input stimulus of precisely known quantity. This is no different for an analytical instrument. In order to calibrate an analyzer, we must exposed it to known quantities of substances with the desired range of properties (density, chemical composition, etc.).

A classic example of this is the calibration of a pH analyzer. pH is the measurement of hydrogen ion activity in an aqueous solution. The standard range of measurement is 0 pH to 14 pH, the number representing a negative power of 10 approximately describing the hydrogen ion molarity of the solution (how many moles of active hydrogen ions per liter of solution)[7].

The pH of a solution is typically measured with a pair of special electrodes immersed in the solution, which generate a voltage proportional to the pH of the solution. In order to calibrate a pH instrument, you must have a sample of liquid solution with a known pH value. For pH instrumentation, such calibration solutions are called *buffers*, because they are specially formulated to maintain stable pH values even in the face of (slight levels of) contamination.

pH buffers may be purchased in liquid form or in powder form. Liquid buffer solutions may be used directly out of the bottle, while powdered buffers must be dissolved in appropriate quantities of de-ionized water to generate a solution ready for calibration use. Pre-mixed liquid buffers are convenient to use, but have a fairly limited shelf life. Powdered buffer capsules are generally superior for long-term storage, and also enjoy the advantage of occupying less storage space in their dry state than a liquid buffer solution.

[7] For example, a solution with a pH value of 4.7 has a concentration of $10^{-4.7}$ moles of active hydrogen ions per liter. For more information on "moles" and solution concentration, see section 3.7, beginning on page 162.

The following photograph shows a few 7.00 pH (+/- 0.02 pH) buffer capsules ready to be mixed with water to form a usable buffer solution:

After preparing the buffer solution in a cup, the pH probe is inserted into the buffer solution and given time to stabilize. One stabilized, the pH instrument may be adjusted to register the proper pH value. Buffer solutions should not be exposed to ambient air for any longer than necessary (especially alkaline buffers such as 10.0 pH) due to contamination[8]. Pre-mixed liquid buffer storage containers should be capped immediately after pouring into working cups. Used buffer solution should be discarded rather than re-used at a later date.

Analyzers designed to measure the concentration of certain gases in air must be calibrated in a similar manner. Oxygen analyzers, for example, used to measure the concentration of free oxygen in the exhaust gases of furnaces, engines, and other combustion processes must be calibrated against known standards of oxygen concentration. An oxygen analyzer designed to measure oxygen concentration over a range of ambient (20.9% oxygen) to 0% oxygen may be calibrated with ambient

[8]Carbon dioxide gas in ambient air will cause carbonic acid to form in an aqueous solution. This has an especially rapid effect on high-pH (alkaline) buffers.

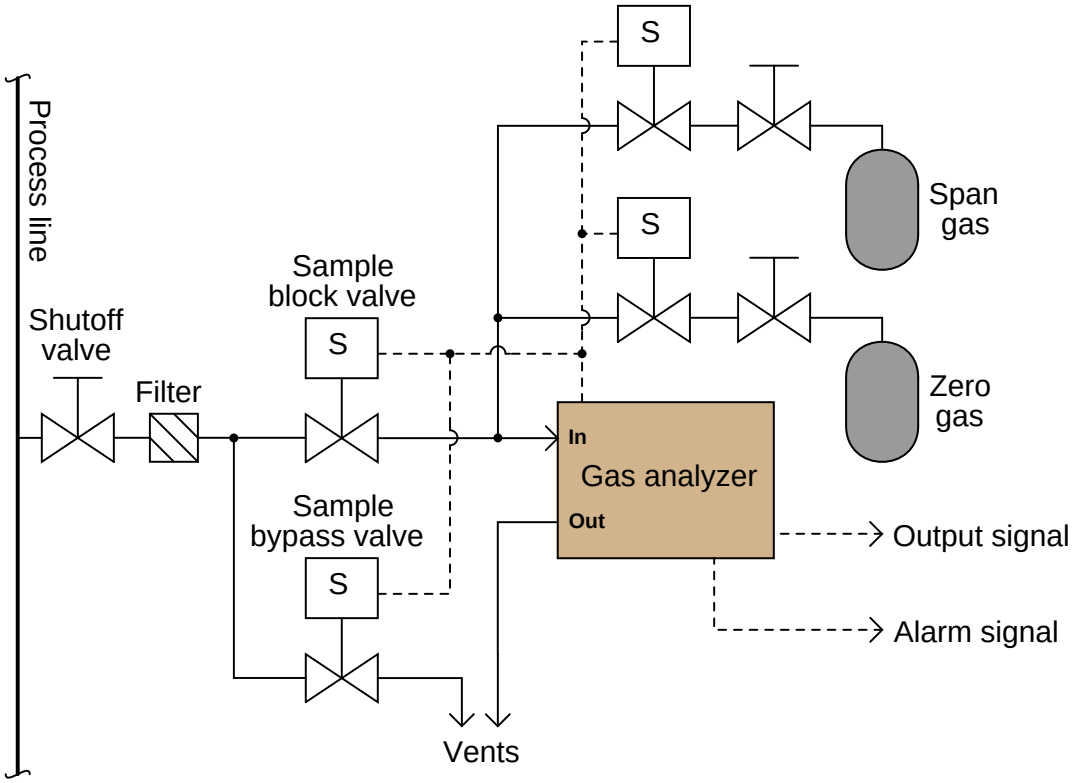
air as one of the standard values[9], and a sample of pure nitrogen gas (containing 0% oxygen) as the other standard value. An oxygen analyzer intended for the measurement of oxygen concentrations in excess of ambient air would require a different standard, most likely a sample of 100% pure oxygen, as a calibration reference.

An analyzer designed to measure the concentration of hydrogen sulfide (H_2S), a toxic gas produced by anaerobic bacterial decomposition of organic matter, will require a sample of gas with a precisely known concentration of hydrogen sulfide mixed in it as a calibration reference. A typical reference gas concentration might be 25 or 50 parts per million (ppm). Gas mixtures with such precise concentration values as this may be purchased from chemical laboratories for the purpose of calibrating concentration analyzers, and are often referred to as *span gases* because they are used to set the span of analyzer instruments.

Analytical instruments are generally subject to greater drifting over time than instruments that measure incidental quantities such as pressure, level, temperature, or flow rate. It is not uncommon for instrument technicians to be tasked with *daily* calibration checks of certain instruments responsible for monitoring atmospheric or water emissions at industrial facilities. For this reason, it is often practical to equip such critical analyzers with *self-calibration* systems. A self-calibration system is a system of solenoid (electrically controlled on-off) valves and reference gas bottles set up in such a way that a computer is able to switch the analyzer off-line and subject it to standard reference gases on a regular schedule to check calibration. Many analyzers are programmed to automatically calibrate themselves against these reference gases, thus eliminating tedious work for the instrument technician.

[9] It is assumed that the concentration of oxygen in ambient air is a stable enough quantity to serve as a calibration standard for most industrial applications. It is certainly an *accessible* standard!

A typical self-calibration system for a gas analyzer might look like this:



The gas analyzer is equipped with its own auto-calibration controls and programming, allowing it to periodically shut off the process sample and switch to known reference gases for "zero" and "span" calibration checks. If these checks indicate excessive drift or any other questionable results, the analyzer has the ability to flag a maintenance alarm to alert an instrument technician to a potential problem that may require servicing. This sort of self-calibration and self-diagnostic capability saves the instrument technician from having to spend substantial time running manual calibration checks, yet alerts the technician if anything is in need of actual repair. Barring any component failures within this system, the only maintenance this system will need is periodic replacement of the calibration gas bottles.

References

Calibration: Philosophy In Practice, Second Edition, Fluke Corporation, Everett, WA, 1994.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Measurement and Analysis Volume I*, Fourth Edition, CRC Press, New York, NY, 2003.

Chapter 18

Continuous pressure measurement

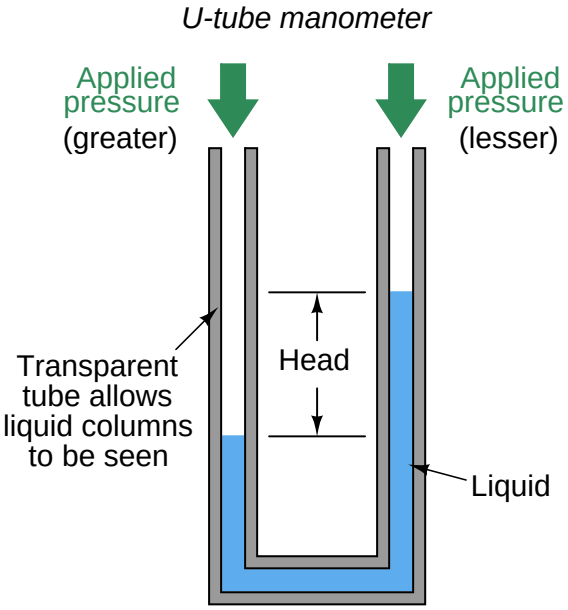
In many ways, pressure is the primary variable for a wide range of process measurements. Many types of industrial measurements are actually inferred from pressure, such as:

- Flow (measuring the pressure dropped across a restriction)
- Liquid level (measuring the pressure created by a vertical liquid column)
- Liquid density (measuring the pressure difference across a fixed-height liquid column)
- Weight (hydraulic load cell)

Even temperature may be inferred from pressure measurement, as in the case of a fluid-filled chamber where fluid pressure and fluid temperature are directly related. As such, pressure is a very important quantity to measure, and measure accurately. This section describes different technologies for the measurement of pressure.

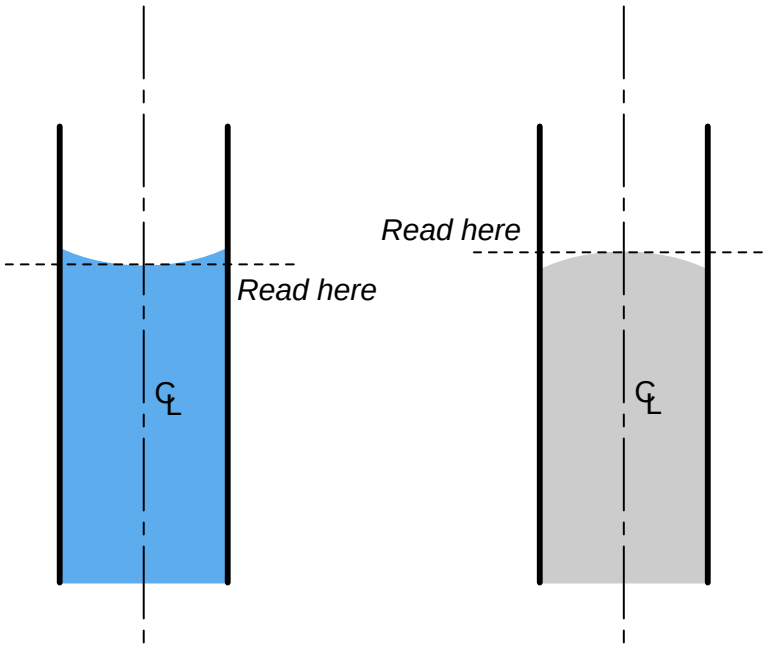
18.1 Manometers

A very simple device used to measure pressure is the *manometer*: a fluid-filled tube where an applied gas pressure causes the fluid height to shift proportionately. This is why pressure is often measured in units of liquid height (e.g. inches of water, inches of mercury). As you can see, a manometer is fundamentally an instrument of *differential* pressure measurement, indicating the difference between two pressures by a shift in liquid column height:

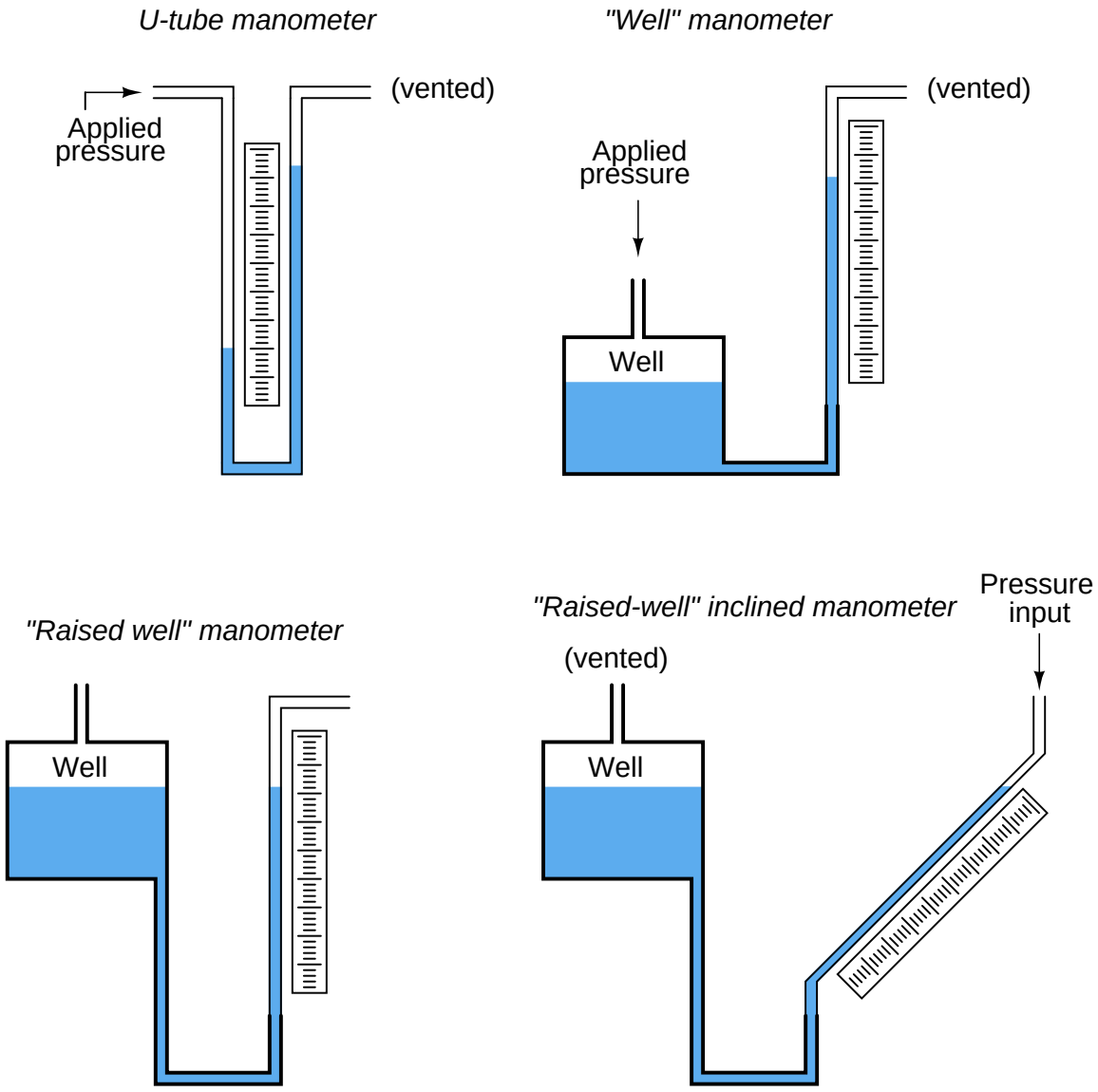


Of course, it is entirely acceptable to simply vent one tube of a manometer and use it as a *gauge* pressure instrument, comparing the applied pressure at one tube against atmospheric pressure in the other.

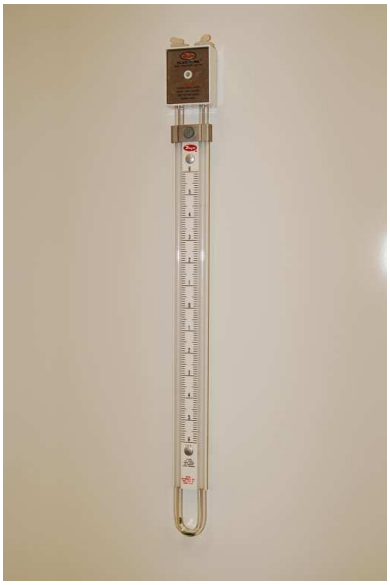
Liquid column height in a manometer should always be interpreted at the centerline of the liquid column, regardless of the shape of the liquid's meniscus (the curved air/liquid interface):



Manometers come in a variety of forms, the most common being the *U-tube*, *well* (sometimes called a *cistern*), *raised well*, and *inclined*:



U-tube manometers are very inexpensive, and are generally made from clear plastic (see the left-hand photo). Cistern-style manometers are the norm for calibration bench work, and are typically constructed from metal cisterns and glass tubes (see the right-hand photo):



Inclined manometers are used to measure very low pressures, owing to their exceptional sensitivity (note the fractional scale for inches of water column in the following photograph, extending from 0 to 1.5 inches on the scale, reading left to right):

Note that venting one side of a manometer is standard practice when using is as a *gauge pressure* indicator (responding to pressure in excess of atmospheric). Both pressure ports will be used if the manometer is applied to the measurement of differential pressure, just as in the case of the U-tube manometer first shown in this section. Absolute pressure may also be measured by a manometer, if one of the pressure ports connects to a sealed vacuum chamber. This is how a *mercury barometer* is constructed for the measurement of absolute ambient air pressure: by sealing off one side of a manometer and removing all the air in that side, such that the applied (atmospheric) pressure is always compared against a vacuum.

Manometers incorporating a "well" have the advantage of single-point reading: one need only compare the height of *one* liquid column, not the difference in height between *two* liquid columns. The cross-sectional area of the liquid column in the well is so much greater than that within the transparent manometer tube that the change in height within the well is usually negligible. In cases where the difference is significant, the spacing between divisions on the manometer scale may be skewed to compensate[1].

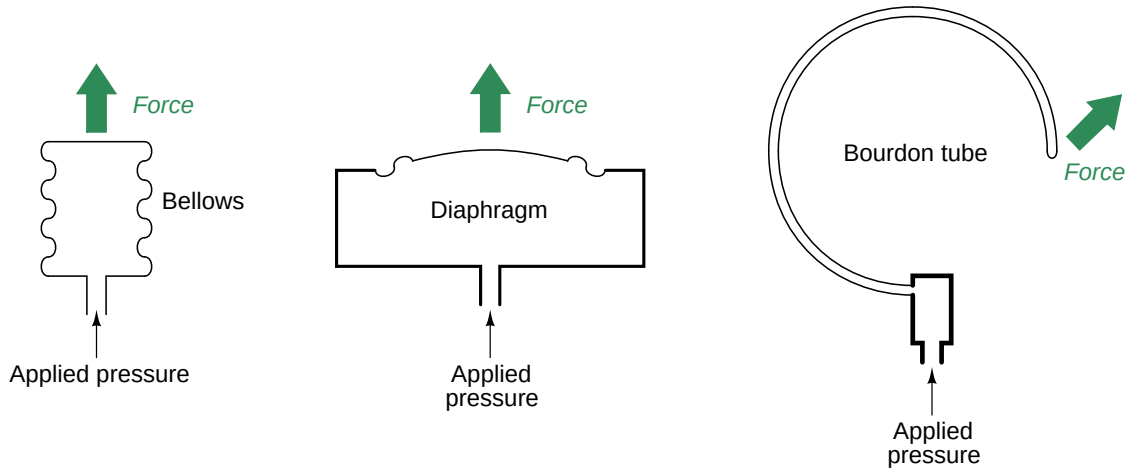
Inclined manometers enjoy the advantage of increased sensitivity. Since manometers fundamentally operate on the principle of pressure balanced by liquid height, and this liquid height is always measured parallel to the line of gravitational pull (perfectly vertical), inclining the manometer tube means that liquid must travel further along the tube to generate the same change in (purely)

[1] If you are having difficulty understanding this concept, imagine a simple U-tube manometer where one of the tubes is opaque, and therefore one of the two liquid columns cannot be seen. In order to be able to measure pressure just by looking at one liquid column height, we would have to make a custom scale where every inch of height registered as *two* inches of water column pressure, because for each inch of height change in the liquid column we can see, the liquid column we can't see also changes by an inch. A scale custom-made for a well-type manometer is just the same concept, only without such dramatic skewing of scales.

vertical height than it would in a vertical manometer tube. Thus, an inclined manometer tube causes an amplification in liquid motion for a given amount of pressure change, allowing measurements of greater resolution.

18.2 Mechanical pressure elements

Mechanical pressure-sensing elements include the *bellows*, the *diaphragm*, and the *bourdon tube*. Each of these devices converts a fluid pressure into a force. If unrestrained, the natural elastic properties of the element will produce a motion proportional to the applied pressure.

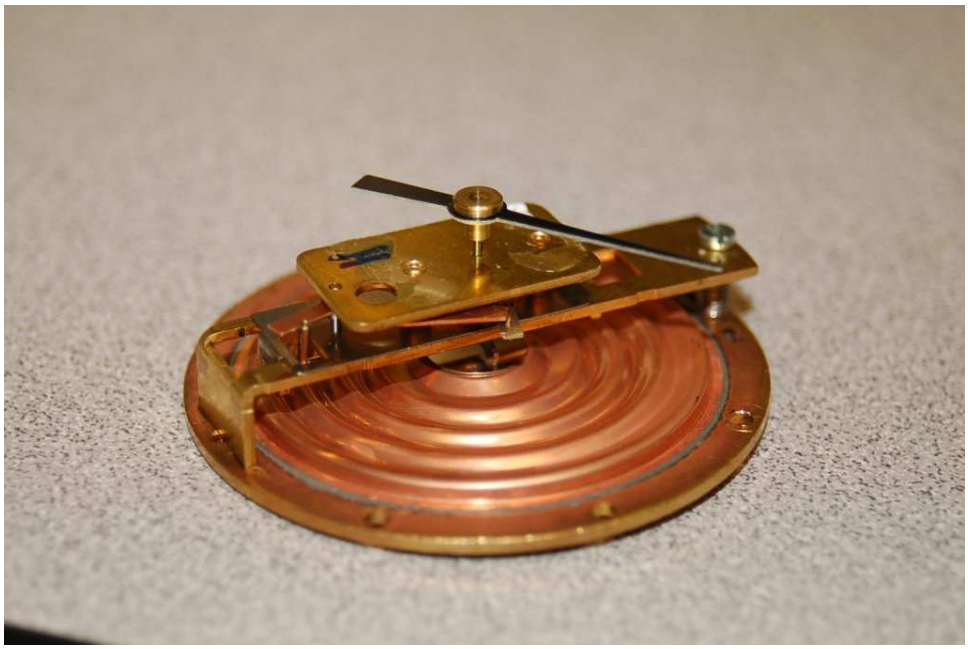


Bellows resemble an accordion constructed from metal instead of fabric. Increasing pressure inside a bellows unit causes it to elongate. A photograph of a bellows is shown here:

A diaphragm is nothing more than a thin disk of material which bows outward under the influence of a fluid pressure. Many diaphragms are constructed from metal, which gives them spring-like qualities. Some diaphragms are intentionally constructed out of materials with little strength, such that there is negligible spring effect. These are called *slack diaphragms*, and they are used

in conjunction with external mechanisms that produce the necessary restraining force to prevent damage from applied pressure.

The following photograph shows the mechanism of a small pressure gauge using a brass diaphragm as the sensing element:

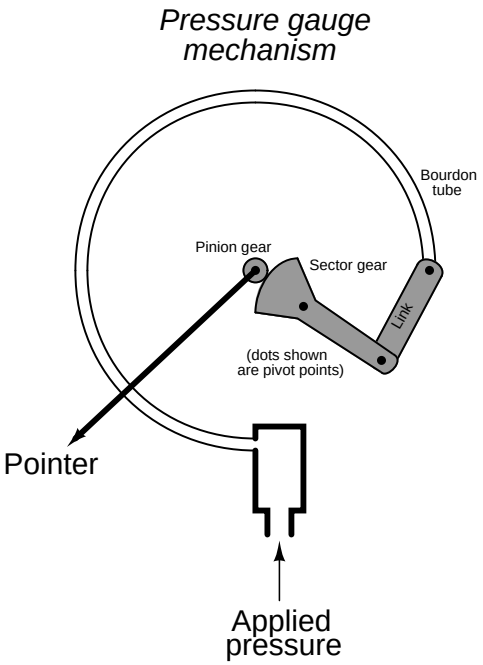


As pressure is applied to the rear of the diaphragm, it distends upward (away from the table on which it rests as shown in the photograph), causing a small shaft to twist in response. This twisting motion is transferred to a lever which pulls on a tiny link chain wrapped around the pointer shaft, causing it to rotate and move the pointer needle around the gauge scale. Both the needle and scale on this gauge mechanism have been removed for easier viewing of diaphragm and mechanism.

Bourdon tubes are made of spring-like metal alloys bent into a circular shape. Under the influence of internal pressure, a bourdon tube "tries" to straighten out into its original shape before being bent at the time of manufacture.

Most pressure gauges use a bourdon tube as their pressure-sensing element. Most pressure transmitters use a diaphragm as their pressure-sensing element. Bourdon tubes may be made in *spiral* or *helical* forms for greater motion (and therefore greater gauge resolution).

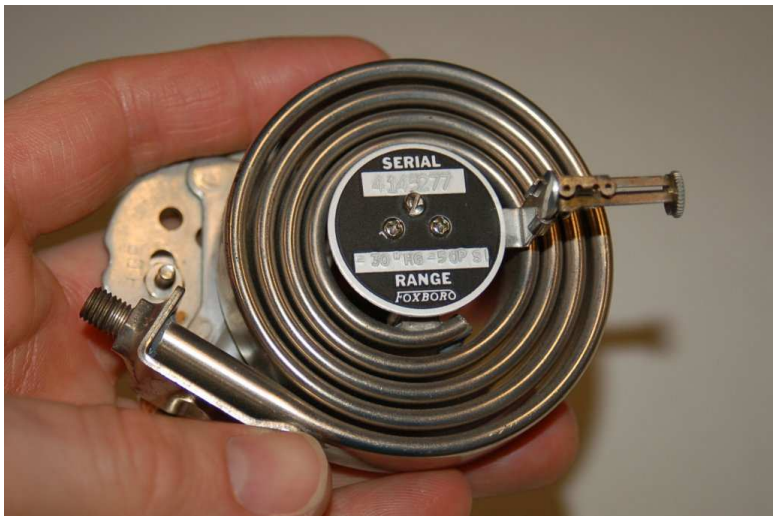
A typical C-shaped bourdon tube pressure gauge mechanism is shown in the following illustration:



A photograph of a C-tube pressure gauge mechanism (taken from the rear of the gauge, behind the pointer and scale) reveals its mechanical workings:

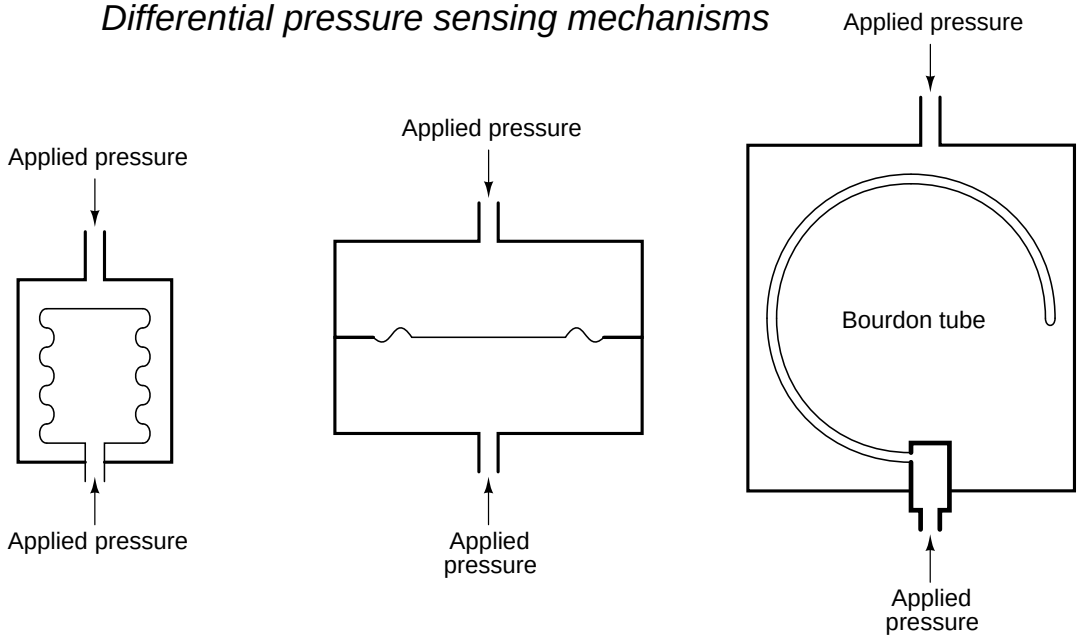
The dark, C-shaped tube is the bourdon tube sensing element, while the shiny metal parts are the linkage, lever, and gear assembly.

The next photograph shows a *spiral* bourdon tube, designed to produce a wider range of motion than a C-tube bourdon:



It should be noted that bellows, diaphragms, and bourdon tubes alike may all be used to measure differential and/or absolute pressure in addition to gauge pressure. All that is needed for these other functionalities is to subject the *other* side of each pressure-sensing element to either another applied pressure (in the case of differential measurement) or to a vacuum chamber (in the case of absolute pressure measurement).

This next set of illustrations shows how bellows, diaphragms, and bourdon tubes may be used as differential pressure-sensing elements:



The challenge in doing this, of course, is how to extract the mechanical motion of the pressure-sensing element to an external mechanism (such as a pointer) while maintaining a good pressure seal. In gauge pressure mechanisms, this is no problem because one side of the pressure-sensing element must be exposed to atmospheric pressure anyway, and so that side is always available for mechanical connection.

A differential pressure gauge is shown in the next photograph. The two pressure ports are clearly evident on either side of the gauge:

18.3 Electrical pressure elements

Several different technologies exist for the conversion of fluid pressure into an electrical signal response. These technologies form the basis of electronic *pressure transmitters*: devices designed to measure fluid pressure and transmit that information via electrical signals such as the 4-20 mA analog standard, or in digital form such as HART or FOUNDATION Fieldbus.

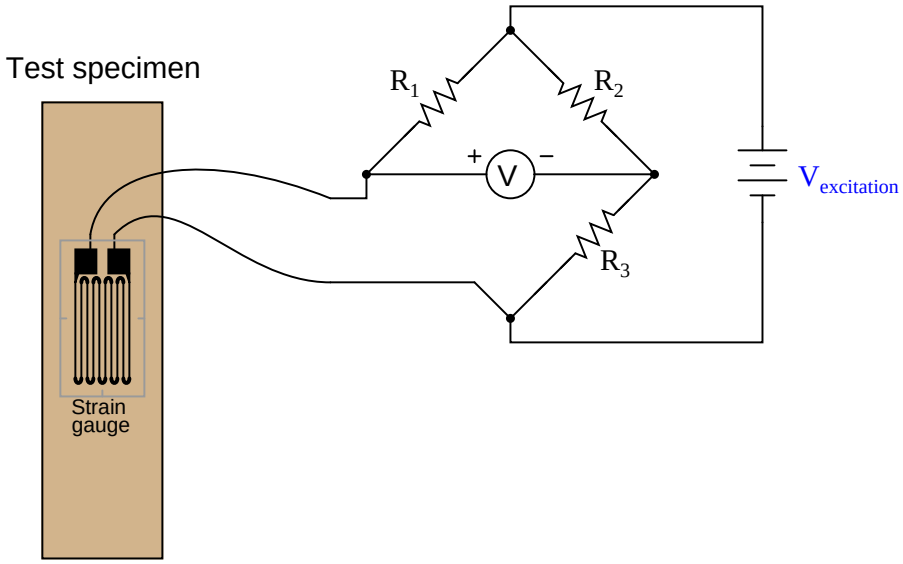
A brief survey of electronic pressure transmitters in contemporary[2] use reveals a diverse representation of electrical pressure-sensing elements:

Manufacturer	Model	Pressure sensor technology
ABB/Bailey	PTSD	Differential reluctance
ABB/Bailey	PTSP	Piezoresistive (strain gauge)
Foxboro	IDP10	Piezoresistive (strain gauge)
Honeywell	ST3000	Piezoresistive (strain gauge)
Rosemount	1151	Differential capacitance
Rosemount	3051	Differential capacitance
Rosemount	3095	Differential capacitance
Yokogawa	EJX series	Mechanical resonance

[2] As of this writing, 2008.

18.3.1 Piezoresistive (strain gauge) sensors

Piezoresistive means "pressure-sensitive resistance," or a resistance that changes value with applied pressure. The *strain gauge* is a classic example of a piezoresistive element:



As the test specimen is stretched or compressed by the application of force, the conductors of the strain gauge are similarly deformed. Electrical resistance of any conductor is proportional to the ratio of length over cross-sectional area ($R \propto \frac{l}{A}$), which means that tensile deformation (stretching) will increase electrical resistance by simultaneously increasing length and decreasing cross-sectional area while compressive deformation (squishing) will decrease electrical resistance by simultaneously decreasing length and increasing cross-sectional area.

Attaching a strain gauge to a diaphragm results in a device that changes resistance with applied pressure. Pressure forces the diaphragm to deform, which in turn causes the strain gauge to change resistance. By measuring this change in resistance, we can infer the amount of pressure applied to the diaphragm.

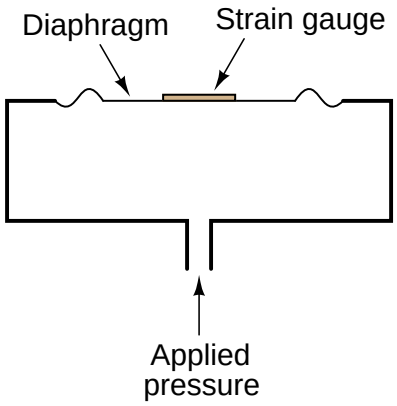
The classic strain gauge system represented in the previous illustration is made of metal (both the test specimen and the strain gauge itself). Within its elastic limits, many metals exhibit good spring characteristics. Metals, however, are subject to *fatigue* over repeated cycles of strain (tension and compression), and they will begin to "flow" if strained beyond their elastic limit. This is a common source of error in metallic piezoresistive pressure instruments: if overpressured, they tend to lose accuracy due to damage of the spring and strain gauge elements.[3]

Modern manufacturing techniques have made possible the construction of strain gauges made of silicon instead of metal. Silicon exhibits very linear spring characteristics over its narrow range of motion, and a high resistance to fatigue. When a silicon strain gauge is over-stressed, it fails

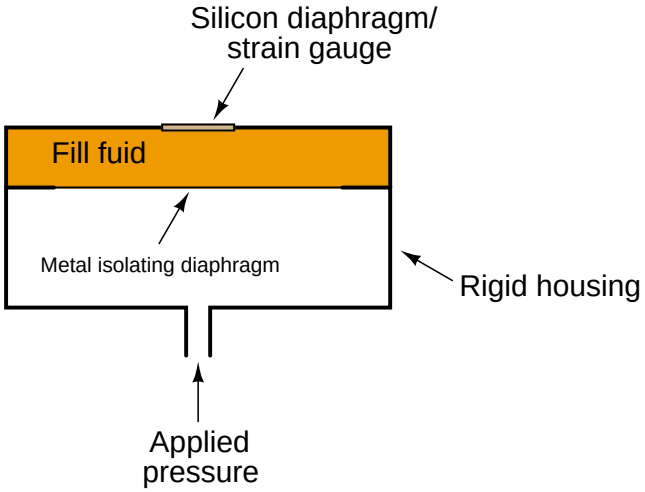
[3] For a simple demonstration of metal fatigue and metal "flow," simply take a metal paper clip and repeatedly bend it back and forth until you feel the metal wire weaken. Gentle force applied to the paper clip will cause it to deform in such a way that it returns to its original shape when the force is removed. Greater force, however, will exceed the paper clip's elastic limit, causing permanent deformation and also altering the spring characteristics of the clip.

completely rather than "flows" as is the case with metal strain gauges. This is generally considered a better result, as it clearly indicates the need for sensor replacement (whereas a metallic strain sensor may give the false impression of continued function after an over-stress event).

Thus, most modern piezoresistive-based pressure instruments use silicon strain gauge elements to sense deformation of a diaphragm due to applied fluid pressure. A simplified illustration of a diaphragm / strain gauge pressure sensor is shown here:



In some designs, a single silicon wafer serves as both the diaphragm and the strain gauge so as to fully exploit the excellent mechanical properties of silicon (high linearity and low fatigue). However, silicon is not chemically compatible with many process fluids, and so pressure must be transferred to the silicon diaphragm/sensor via a non-reactive *fill fluid* (commonly a silicone-based or fluorocarbon-based liquid). A metal *isolating diaphragm* transfers process fluid pressure to the fill fluid. Another simplified illustration shows how this works:



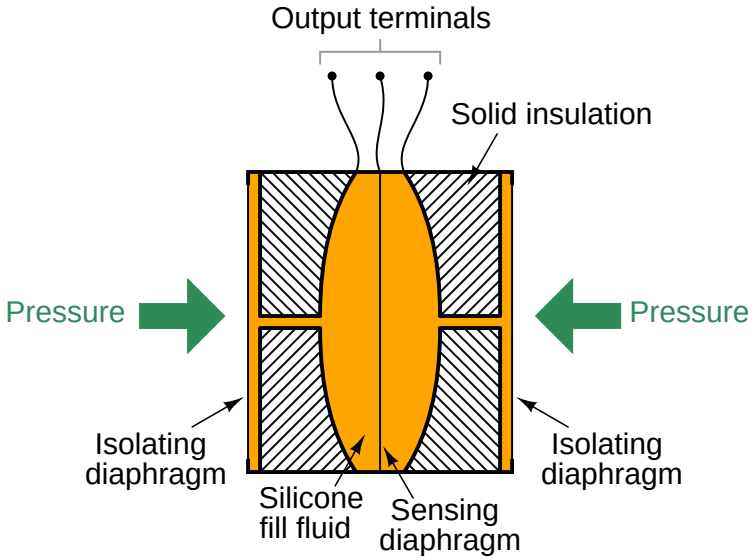
The isolating diaphragm is designed to be much more flexible (less rigid) than the silicon diaphragm, because its purpose is to seamlessly transfer fluid pressure from the process fluid to

the fill fluid, not to act as a spring element. In this way, the silicon sensor experiences the same pressure that it would if it were directly exposed to the process fluid, without having to contact the process fluid.

An example of a pressure instrument utilizing a silicon strain gauge element is the Foxboro model IDP10 differential pressure transmitter, shown in the following photograph:

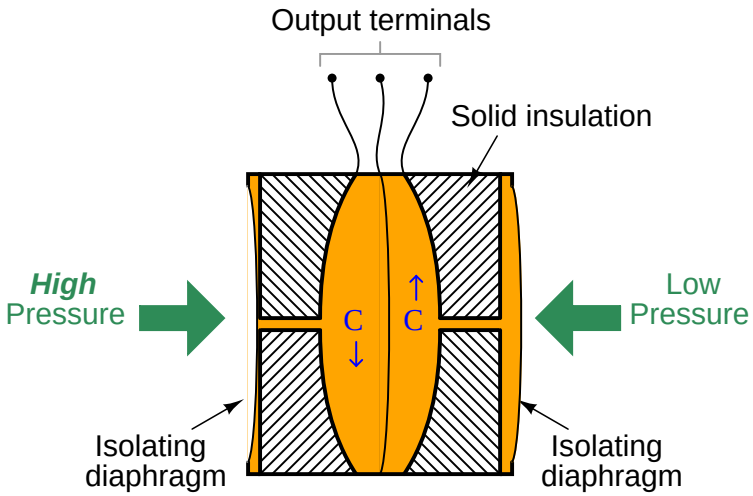
18.3.2 Differential capacitance sensors

Another common electrical pressure sensor design works on the principle of *differential capacitance*. In this design, the sensing element is a taut metal diaphragm located equidistant between two stationary metal surfaces, forming a complementary pair of capacitances. An electrically insulating fill fluid (usually a liquid silicone compound) transfers motion from the isolating diaphragms to the sensing diaphragm, and also doubles as an effective dielectric for the two capacitors:



Any difference of pressure across the cell will cause the diaphragm to flex in the direction of least pressure. The sensing diaphragm is a precision-manufactured spring element, meaning that its displacement is a predictable function of applied force. The applied force in this case can only be a function of differential pressure acting against the surface area of the diaphragm in accordance with the standard force-pressure-area equation $F = PA$. In this case, we have two forces caused by two fluid pressures working against each other, so our force-pressure-area equation may be re-written to describe *resultant* force as a function of differential pressure $(P_1 - P_2)$ and diaphragm area: $F = (P_1 - P_2)A$. Since diaphragm area is constant, and force is predictably related to diaphragm displacement, all we need now in order to infer differential pressure is to accurately measure displacement of the diaphragm.

The diaphragm's secondary function as one plate of two capacitors provides a convenient method for measuring displacement. Since capacitance between conductors is inversely proportional to the distance separating them, capacitance on the low-pressure side will increase while capacitance on the high-pressure side will decrease:

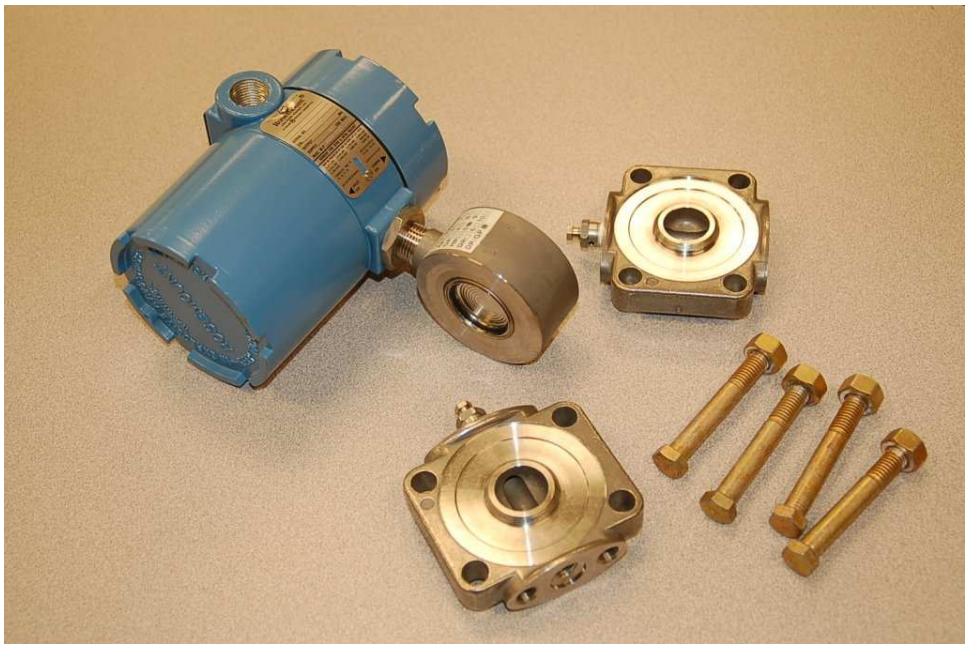


A capacitance detector circuit connected to this cell uses a high-frequency AC excitation signal to measure the different in capacitance between the two halves, translating that into a DC signal which ultimately becomes the signal output by the instrument representing pressure.

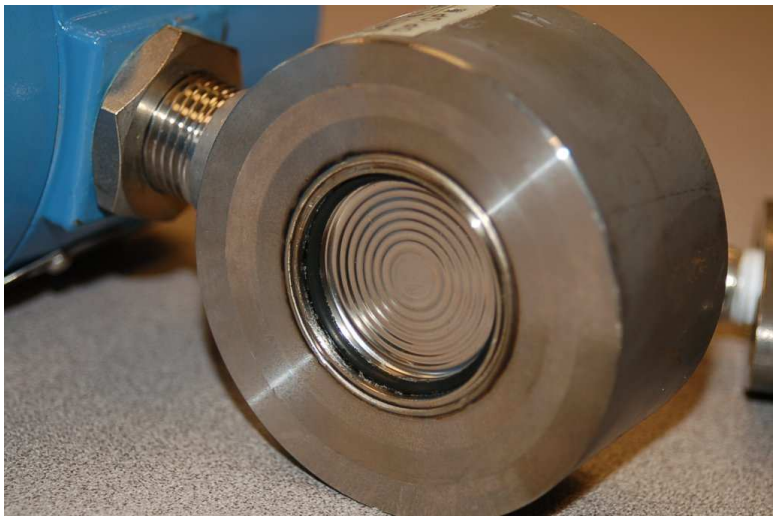
These pressure sensors are highly accurate, stable, and rugged. The solid frame bounds the motion of the two isolating diaphragms such that the sensing diaphragm cannot move past its elastic limit. This gives the differential capacitance excellent resistance to overpressure damage.

A classic example of a pressure instrument based on the differential capacitance sensor is the Rosemount model 1151 differential pressure transmitter, shown in assembled form in the following photograph:

By removing four bolts from the transmitter, we are able to remove two flanges from the pressure capsule, exposing the isolating diaphragms to plain view:



A close-up photograph shows the construction of one of the isolating diaphragms, which unlike the sensing diaphragm is designed to be very flexible. The concentric corrugations in the metal of the diaphragm allow it to easily flex with applied pressure, transmitting process fluid pressure through the silicone fill fluid to the taut sensing diaphragm inside the differential capacitance cell:



The differential capacitance sensor inherently measures *differences* in pressure applied between

its two sides. In keeping with this functionality, this pressure instrument has two threaded ports into which fluid pressure may be applied. A later section in this chapter will elaborate on the utility of differential pressure transmitters (section 18.5 beginning on page 797).

All the electronic circuitry necessary for converting the sensor's differential capacitance into an electronic signal representing pressure is housed in the blue-colored structure above the capsule and flanges.

A more modern realization of the differential capacitance pressure-sensing principle is the Rosemount model 3051 differential pressure transmitter:

As is the case for all differential pressure devices, this instrument has two ports through which fluid pressure may be applied to the sensor. The sensor, in turn, responds only to the *difference* in pressure between the ports.

The differential capacitance sensor construction is more complex in this particular pressure instrument, with the plane of the sensing diaphragm lying perpendicular to the plane of the two isolating diaphragms. This "coplanar" design is far more compact than the older style of sensor, and it isolates the sensing diaphragm from flange bolt stress – one of the main sources of error in the previous design[4].

[4]Not only did applied torque of the four capsule bolts affect measurement accuracy in the older 1151 model design, but changes in temperature resulting in changing bolt tension also had a detrimental impact on accuracy. Most modern differential pressure transmitter designs strive to isolate the sensing diaphragm assembly from flange bolt stress for these reasons.

18.3.3 Resonant element sensors

As any guitarist, violinist, or other stringed-instrument musician can tell you, the natural frequency of a tensed string increases with tension. This, in fact, is how stringed instruments are tuned: the tension on each string is precisely adjusted to achieve the desired resonant frequency.

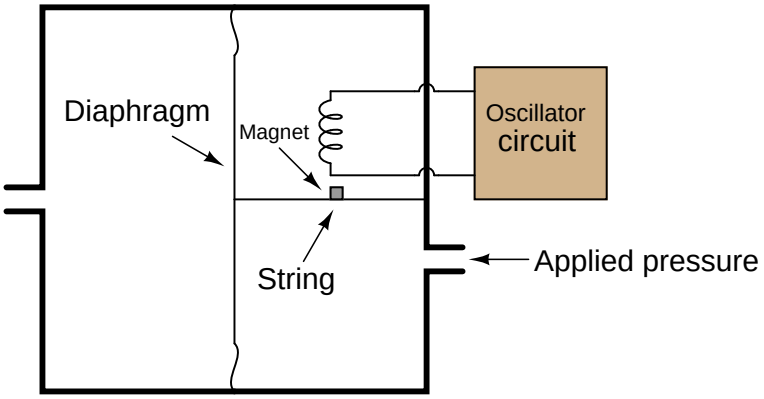
Mathematically, the resonant frequency of a string may be described by the following formula:

$$f = \frac{1}{2L}\sqrt{\frac{F_T}{\mu}}$$

Where,

f = Fundamental resonant frequency of string (Hertz)
L = String length (meters)
F_T = String tension (newtons)
μ = Unit mass of string (kilograms per meter)

It stands to reason, then, that a string may serve as a force sensor. All that is needed to complete the sensor is an oscillator circuit to keep the string vibrating at its resonant frequency, and that frequency becomes an indication of tension (force). If the force stems from pressure applied to some sensing element such as a bellows or diaphragm, the string's resonant frequency will indicate fluid pressure. A proof-of-concept device based on this principle might look like this:



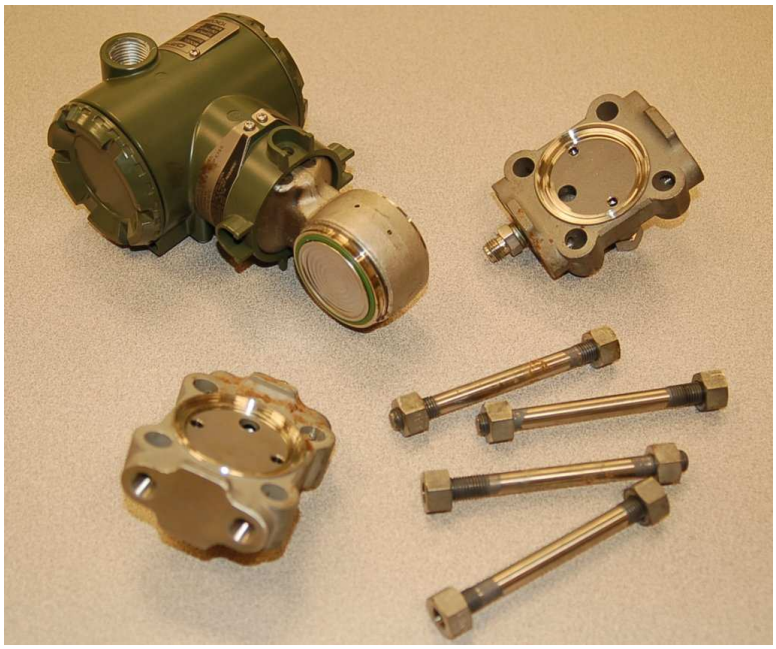
The Foxboro company pioneered this concept in an early *resonant wire* design of pressure transmitter. Later, the Yokogawa corporation of Japan applied the concept to a pair of micro-machined[5] silicon resonator structures, which became the basis for their successful line of "DPharp" pressure transmitters.

[5]This is an example of a micro-electro-mechanical system, or *MEMS*.

A photograph of a Yokogawa model EJA110 pressure transmitter with this technology is seen here:

Process pressure enters through ports in two flanges, presses against a pair of isolating diaphragms, transferring motion to the sensing diaphragm where the resonant elements change frequency with diaphragm strain. Electronic circuits within the upper housing measure the two resonant elements' frequencies and generate an output signal proportional to their frequency difference. This, of course, is a representation of applied differential pressure.

Even when disassembled, the transmitter does not look much different from the more common differential capacitance sensor design.



The important design differences are hidden from view, inside the sensing capsule. Functionally, though, this transmitter is much the same as its differential-capacitance cousin.

An interesting advantage of the resonant element pressure sensor is that the sensor signal is very easy to digitize. The vibration of each resonant element is sensed by the electronics package as an AC frequency. Any frequency signal may be easily "counted" over a given span of time and converted to a binary digital representation. Quartz crystal electronic oscillators are extremely precise, providing the stable frequency reference necessary for comparison in any frequency-based instrument.

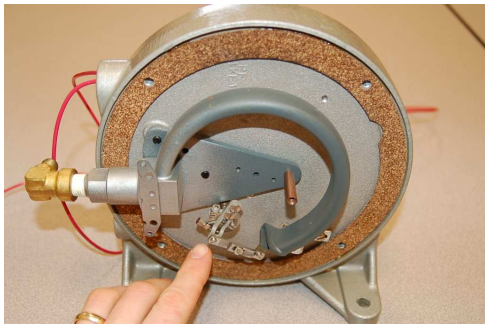
In the Yokogawa "DPharp" design, the two resonant elements oscillate at a nominal frequency of 90 kHz. As the sensing diaphragm deforms with applied differential pressure, one resonator experiences tension while the other experiences compression, causing the frequency of the former to shift up and the latter to shift down (as much as +/- 20 kHz). The signal conditioning electronics inside the transmitter measures this difference in resonator frequency to infer applied pressure.

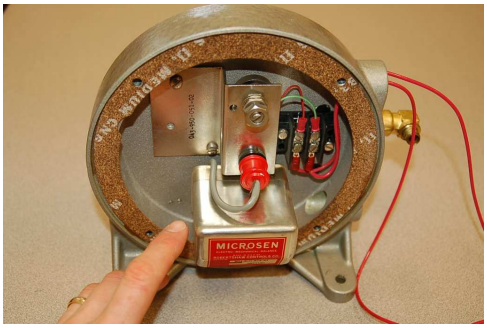
18.3.4 Mechanical adaptations

Most modern electronic pressure sensors convert very small diaphragm motions into electrical signals through the use of sensitive motion-sensing techniques (strain gauge sensors, differential capacitance cells, etc.). Diaphragms made from elastic materials behave as springs, but circular diaphragms exhibit very nonlinear behavior when significantly stretched unlike classic spring designs such as coil and leaf springs which exhibit linear behavior over a wide range of motion. Therefore, in order to yield a linear response to pressure, a diaphragm-based pressure sensor must be designed in such a way that the diaphragm stretches very little over the normal range of operation. Limiting the displacement of a diaphragm necessitates highly sensitive motion-detection techniques such as strain gauge sensors, differential capacitance cells, and mechanical resonance sensors to convert that diaphragm's very slight motion into an electronic signal.

An alternative approach to electronic pressure measurement is to use mechanical pressure-sensing elements with more linear pressure-displacement characteristics – such as bourdon tubes and spring-loaded bellows – and then detect the large-scale motion of the pressure element using a less-sophisticated electrical motion-sensing device such as a potentiometer, LVDT, or Hall Effect sensor. In other words, we take the sort of mechanism commonly found in a direct-reading pressure gauge and attach it to a potentiometer (or similar device) to derive an electrical signal from the pressure measurement.

The following photographs show front and rear views of an electronic pressure transmitter using a large C-shaped bourdon tube as the sensing element (seen in the left-hand photograph):



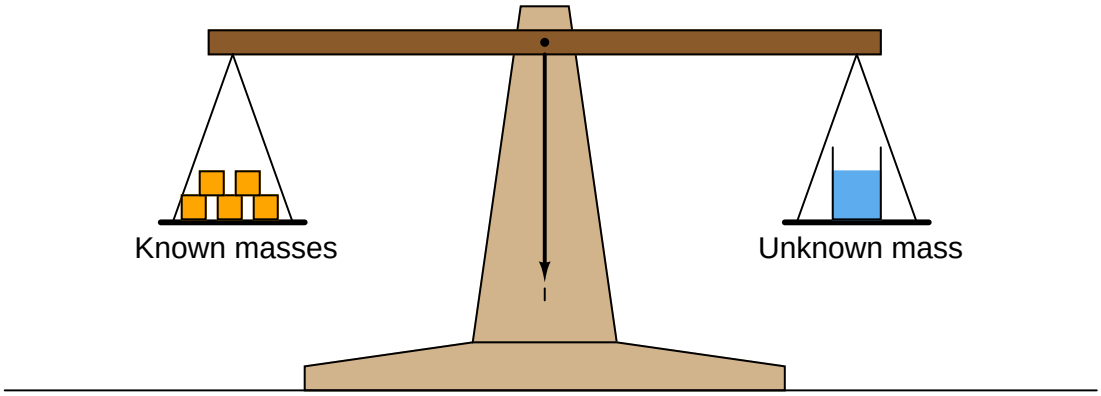


This alternative approach is undeniably simpler and less expensive to manufacture than the more sophisticated approaches used with diaphragm-based pressure instruments, but is prone to greater inaccuracies. Even bourdon tubes and bellows are not perfectly linear spring elements, and the substantial motions involved with using such pressure elements introduces the possibility of hysteresis errors (where the instrument does not respond accurately during reversals of pressure, where the mechanism changes direction of motion) due to mechanism friction, and deadband errors due to backlash (looseness) in mechanical connections.

You are likely to encounter this sort of pressure instrument design in direct-reading gauges equipped with electronic transmitting capability. An instrument manufacturer will take a proven product line of pressure gauge and add a motion-sensing device to it that generates an electric signal proportional to mechanical movement inside the gauge, resulting in an inexpensive pressure transmitter that happens to double as a direct-reading pressure gauge.

18.4 Force-balance pressure transmitters

An important legacy technology for all kinds of continuous measurement is the *self-balancing system*. A "self-balance" system continuously balances an adjustable quantity against a sensed quantity, the adjustable quantity becoming an indication of the sensed quantity once balance is achieved. A common manual-balance system is the type of scale used in laboratories to measure mass:

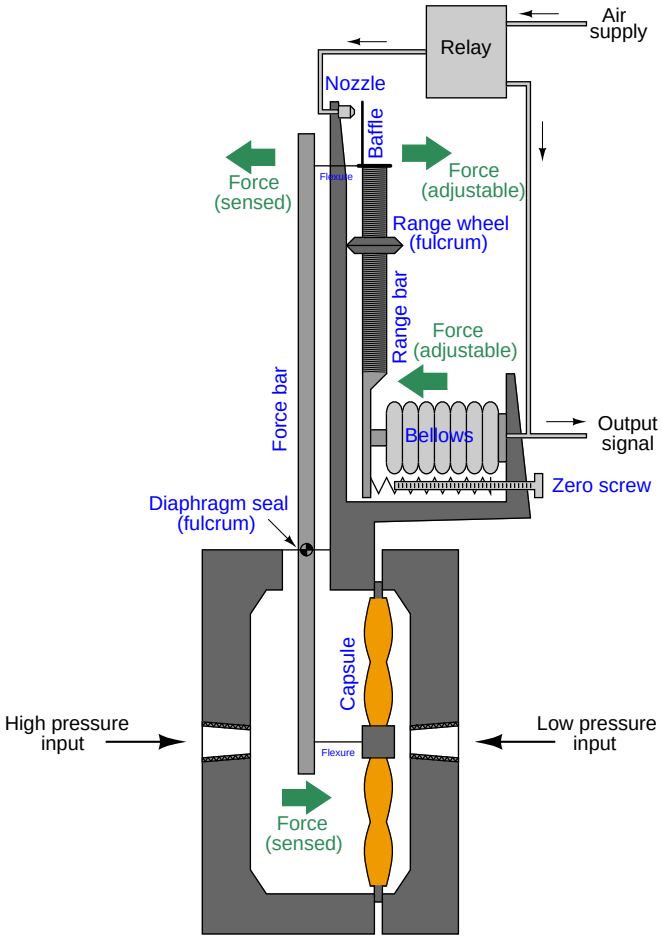


Here, the unknown mass is the sensed quantity, and the known masses are the adjustable quantity. A human lab technician applies as many masses to the left-hand side of the scale as needed to achieve balance, then counts up the sum total of those masses to determine the quantity of the unknown mass.

Such a system is perfectly linear, which is why these balance scales are popularly used for scientific work. The scale mechanism itself is the very model of simplicity, and the only thing the pointer needs to accurately sense is a condition of balance (equality between masses).

If the task of balancing is given to an automatic mechanism, the adjustable quantity will continuously change and adapt as needed to balance the sensed quantity, thereby becoming a representation of that sensed quantity. In the case of pressure instruments, pressure is easily converted into force by acting on the surface area of a sensing element such as a diaphragm or a bellows. A balancing force may be generated to exactly cancel the process pressure's force, making a *force-balance* pressure instrument. Like the laboratory balance scale, an industrial instrument built on the principle of balancing a sensed quantity with an adjustable quantity will be inherently linear, which is a tremendous advantage for measurement purposes.

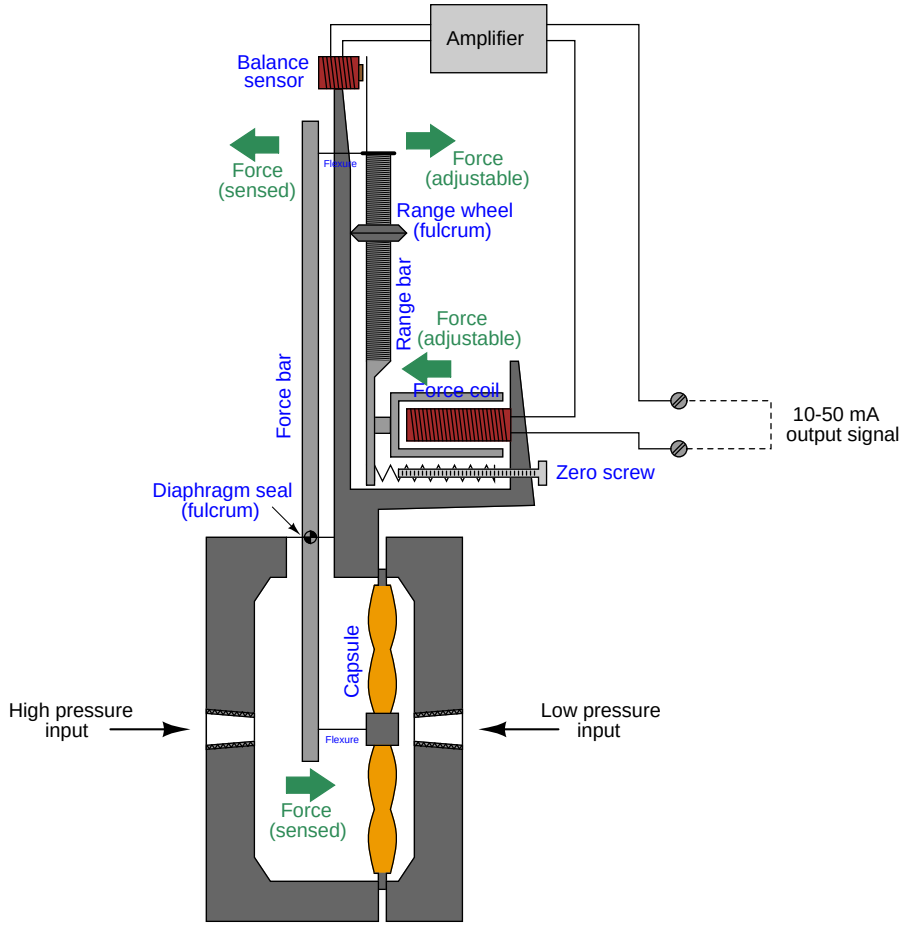
Here, we see a diagram of a force-balance pneumatic pressure transmitter[6], balancing a sensed differential pressure with an adjustable air pressure which becomes a pneumatic output signal:



Differential pressure is sensed by a liquid-filled diaphragm "capsule," which transmits force to a "force bar." If the force bar moves out of position due to this applied force, a highly sensitive "baffle" and "nozzle" mechanism senses it and causes a pneumatic amplifier (called a "relay") to send a different amount of air pressure to a bellows unit. The bellows presses against the "range bar" which pivots to counter-act the initial motion of the force bar. When the system returns to equilibrium, the air pressure inside the bellows will be a direct, linear representation of the process fluid pressure applied to the diaphragm capsule.

[6]Based on the design of Foxboro's popular model 13A pneumatic "DP cell" differential pressure transmitter.

With minor modifications to the design of this pressure transmitter[7], we may convert it from pneumatic to electronic force-balancing:



Differential pressure is sensed by the same type of liquid-filled diaphragm capsule, which transmits force to the force bar. If the force bar moves out of position due to this applied force, a highly sensitive electromagnetic sensor detects it and causes an electronic amplifier to send a different amount of electric current to a force coil. The force coil presses against the range bar which pivots to counteract the initial motion of the force bar. When the system returns to equilibrium, the milliampere current through the force coil will be a direct, linear representation of the process fluid pressure applied to the diaphragm capsule.

A distinct advantage of force-balance pressure instruments (besides their inherent linearity) is the constraining of sensing element motion. Unlike a modern diaphragm-based pressure transmitter which relies on the spring characteristics of the diaphragm to convert pressure into force and then

[7]Very loosely based on the design of Foxboro's now-obsolete E13 electronic "DP cell" differential pressure transmitter.

into motion (displacement) which is sensed and converted into an electronic signal, a force-balance transmitter works best when the diaphragm is slack and has no spring characteristics at all. Balance with the force of the process fluid pressure is achieved by the application of either an adjustable air pressure or an adjustable electric current, not by the natural tensing of a spring element. This makes a force-balance instrument far less susceptible to errors due to metal fatigue or any other degradation of spring characteristics.

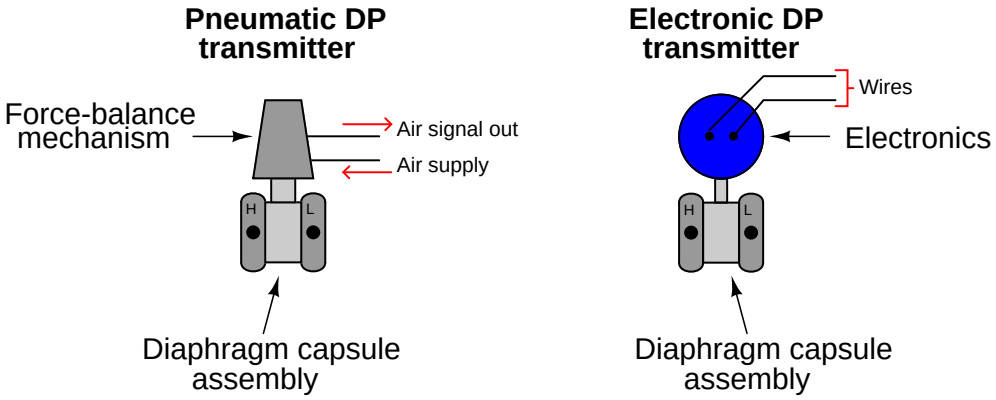
Unfortunately, force-balance instruments have significant disadvantages as well. Force-balance mechanisms tend to be bulky[8], and they translate external vibration into inertial force which adds "noise" to the output signal. Also, the amount of electrical power necessary to provide adequate balancing force in an electronic force-balance transmitter is such that it is nearly impossible to limit below the level necessary to ensure intrinsic safety (protection against the accidental ignition of explosive atmospheres by limiting the amount of energy the instrument could possibly discharge into a spark).

[8] One instrument technician I encountered referred to the Foxboro E13 differential pressure transmitter as "pig iron" after having to hoist it by hand to the top of a distillation column.

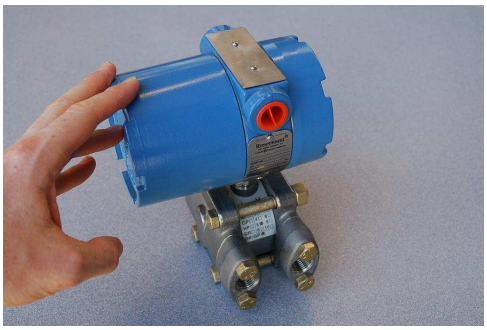
18.5 Differential pressure transmitters

One of the most common, and most useful, pressure measuring instruments in industry is the *differential pressure transmitter*. This device senses the difference in pressure between two ports and outputs a signal representing that pressure in relation to a calibrated range. Differential pressure transmitters may be based on any of the previously discussed pressure-sensing technologies, so this section focuses on application rather than theory.

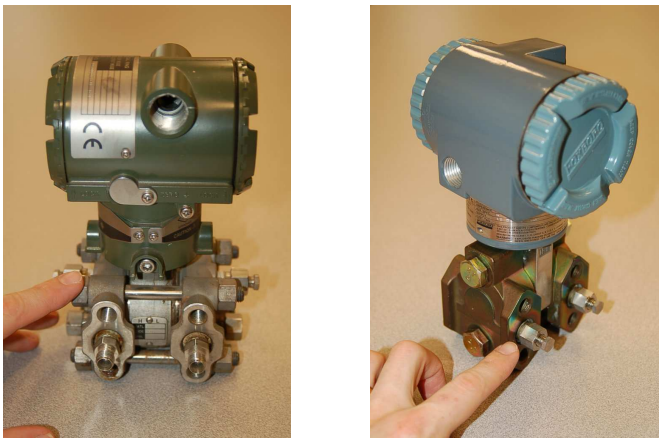
Differential pressure transmitters look something like this:



Two models of electronic differential pressure transmitter are shown here, the Rosemount model 1151 (left) and model 3051 (right):

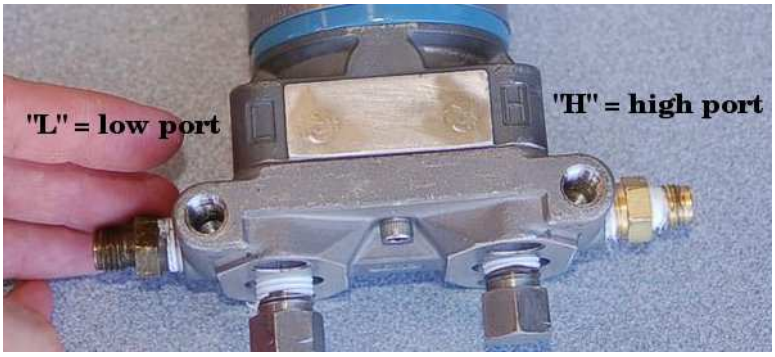


Two more models of electronic differential pressure transmitter are shown in the next photograph, the Yokogawa EJA110 (left) and the Foxboro IDP10 (right):



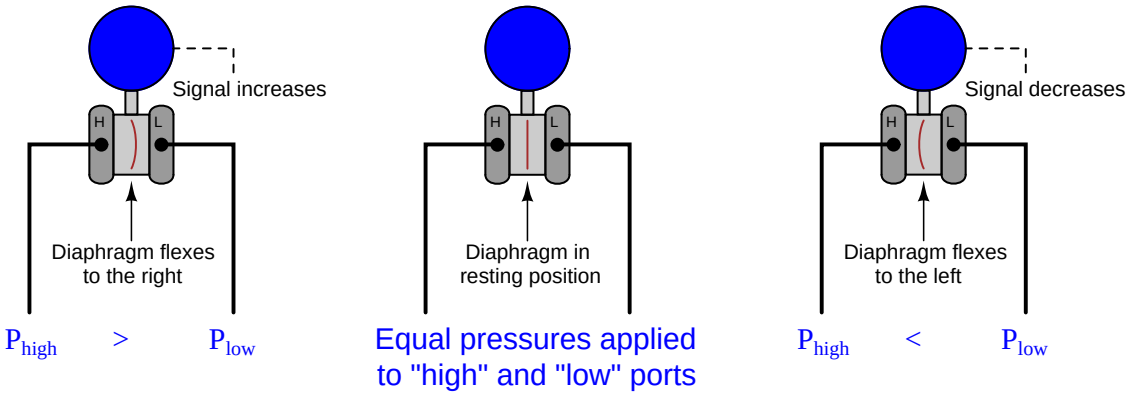
In each of these differential pressure transmitter examples, the pressure-sensing element is housed in the bottom half of the device (the forged-steel structure) while the electronics are housed in the top half (the colored, round, cast-aluminum structure).

Regardless of make or model, every differential pressure ("DP", "d/p", or ΔP)[9] transmitter has *two* pressure ports to sense different process fluid pressures. These ports typically have $\frac{1}{4}$ inch female NPT threads to readily accept connection to the process. One of these ports is labeled "high" and the other is labeled "low". This labeling does not necessarily mean that the "high" port must always be at a greater pressure than the "low" port. What these labels represent is the effect that a pressure at that point will have on the output signal.

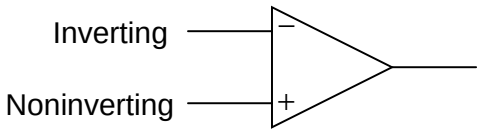


[9] As far as I have been able to determine, the labels "D/P" and "DP cell" were originally trademarks of the Foxboro Company. Those particular transmitter models became so popular that the term "DP cell" came to be applied to nearly *all* makes and models of differential pressure transmitter, much like the trademark "Vise-Grip" is often used to describe *any* self-locking pliers, or "Band-Aid" is often used to describe *any* form of self-adhesive bandage.

The most common sensing element used by modern DP transmitters is the diaphragm. One side of this diaphragm receives process fluid pressure from the "high" port, while the other receives process fluid pressure from the "low" port. Any difference of pressure between the two ports causes the diaphragm to flex from its normal resting (center) position. This flexing is then translated into an output signal by any number of different technologies, depending on the manufacturer and model of the transmitter:

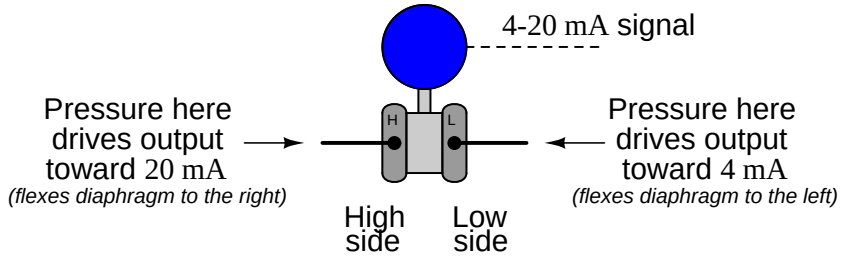


The concept of differential pressure instrument port labeling is very similar to the "inverting" and "noninverting" labels applied to operational amplifier input terminals:



The "+" and "-" symbols do not imply polarity of the input voltage(s). It is not as though the "+" input must be more positive than the "-" input. These symbols merely represent the different effects on the output signal that each input has. An increasing voltage applied to the "+" input drives the op-amp's output positive, while an increasing voltage applied to the "-" input drives the op-amp's output negative.

In a similar manner, an increasing pressure applied to the "high" port of a DP transmitter will drive the output signal to a greater level (up), while an increasing pressure applied to the "low" port of a DP transmitter will drive the output signal to a lesser level (down)[10]:

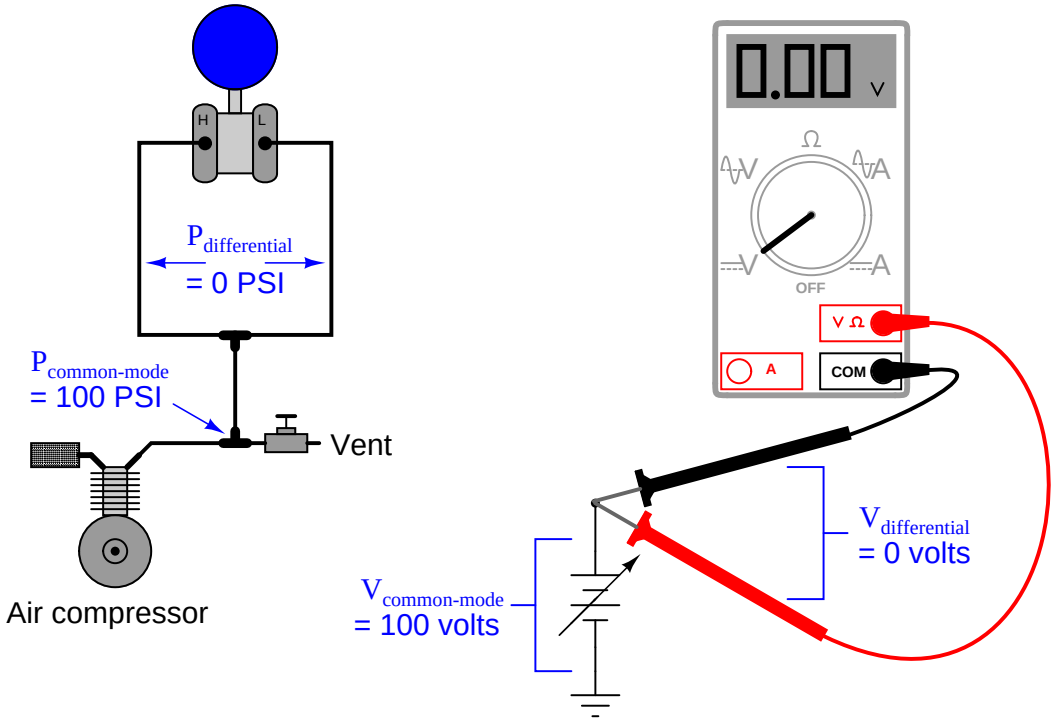


In the world of electronics, we refer to the ability of a differential voltage sensor (such as an operational amplifier) to sense small differences in voltage while ignoring large potentials measured with reference to ground by the phrase *common-mode rejection*. An ideal operational amplifier completely ignores the amount of voltage common to both input terminals, responding only to the *difference* in voltage *between* those terminals. This is precisely what a well-designed DP instrument does, except with fluid pressure instead of electrical voltage. A DP instrument ignores gauge pressure common to both ports, while responding only to *differences* in pressure *between* those two ports.

[10] One transmitter manufacturer I am aware of (ABB/Bailey) actually does use the "+" and "-" labels to denote high- and low-pressure ports rather than the more customary "H" and "L" labels found on other manufacturers' DP products.

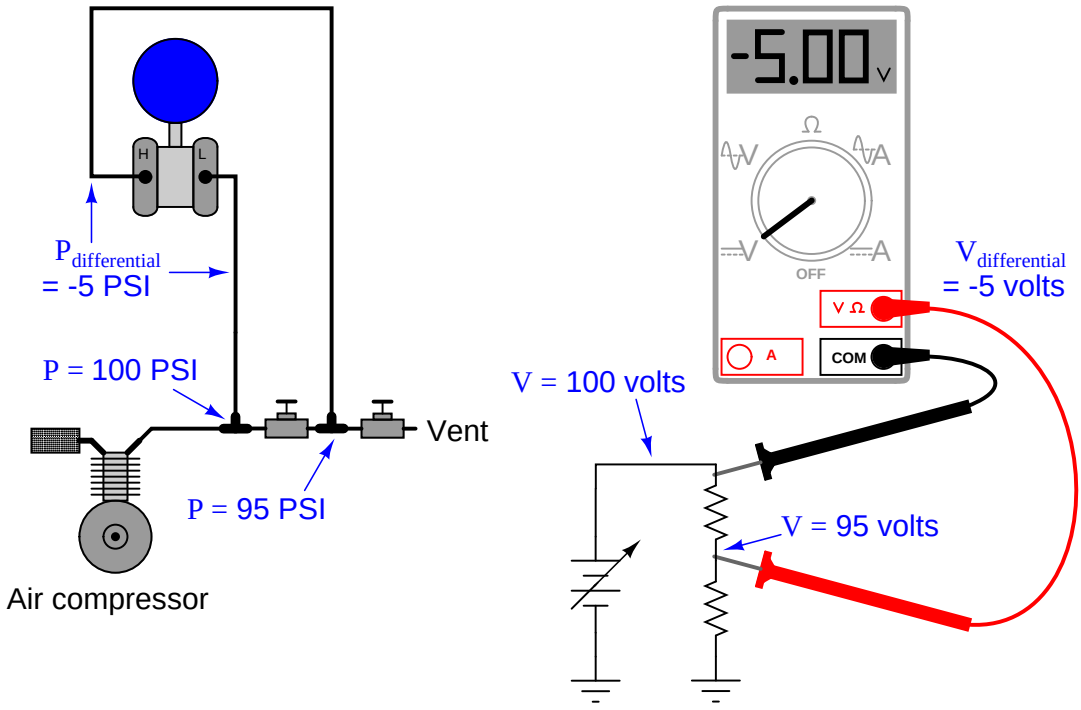
To illustrate, we may connect the "high" and "low" ports of a differential pressure transmitter together using pipe or tube, then expose both ports simultaneously to a source of fluid pressure such as pressurized air from an air compressor. If the transmitter is in good working order, it should continue to register zero differential pressure even as we vary the amount of static pressure applied to both ports. So long as the applied pressures to each port are equal, the transmitter's sensing diaphragm should experience zero net force pushing left or right. All force applied to the diaphragm from the "high" port's fluid pressure should be precisely countered (canceled) by force applied to the diaphragm from the "low" port's fluid pressure.

An electrical analogy to this would be connecting both red and block test leads of a voltmeter to a common point in an electrical circuit, then varying the amount of voltage between that point and earth ground. Since the voltmeter only registers *differences* of potential between its test leads, and those test leads are now electrically common to one another, the magnitude of common-mode voltage between that one point of the circuit and earth ground is irrelevant from the perspective of the voltmeter:



In each case the differential measurement device *rejects* the common-mode value, registering only the amount of difference (zero) between its sensing points.

The same common-mode rejection principle reveals itself in more complex fluid and electrical circuits. Consider the case of a DP transmitter and a voltmeter, both used to measure differential quantities in a "divider" circuit:



In each case the differential measurement device responds only to the difference between the two measurement points, rejecting the common-mode value (97.5 PSI for the pressure transmitter, 97.5 volts for the voltmeter). Just to make things interesting in this example, the "high" side of each measuring instrument connects to the point of lesser value, such that the measured difference is a negative quantity. Like digital voltmeters, modern DP transmitters are equally capable of accurately measuring negative pressure differences as well as positive pressure differences.

A vivid contrast between *differential* pressure and *common-mode* pressure for a DP instrument is seen in the pressure ratings shown on the nameplate of a Foxboro model 13A differential pressure transmitter:

This nameplate tells us that the transmitter has a calibrated differential pressure range of 50" H_2O (50 inches water column, which is only about 1.8 PSI). However, the nameplate also tells us that the transmitter has a *maximum working pressure* (MWP) of 1500 PSI. "Working pressure" refers to the amount of gauge pressure common to each port, not the differential pressure between ports. Taking these figures at face value means this transmitter will register zero (no differential pressure) even if the gauge pressure applied equally to both ports is a full 1500 PSI! In other words, this differential pressure transmitter will *reject* up to 1500 PSI of common-mode gauge pressure, and respond only to small differences in pressure between the ports (1.8 PSI differential being enough to stimulate the transmitter to full scale output).

18.5.1 Pressure measurement applications

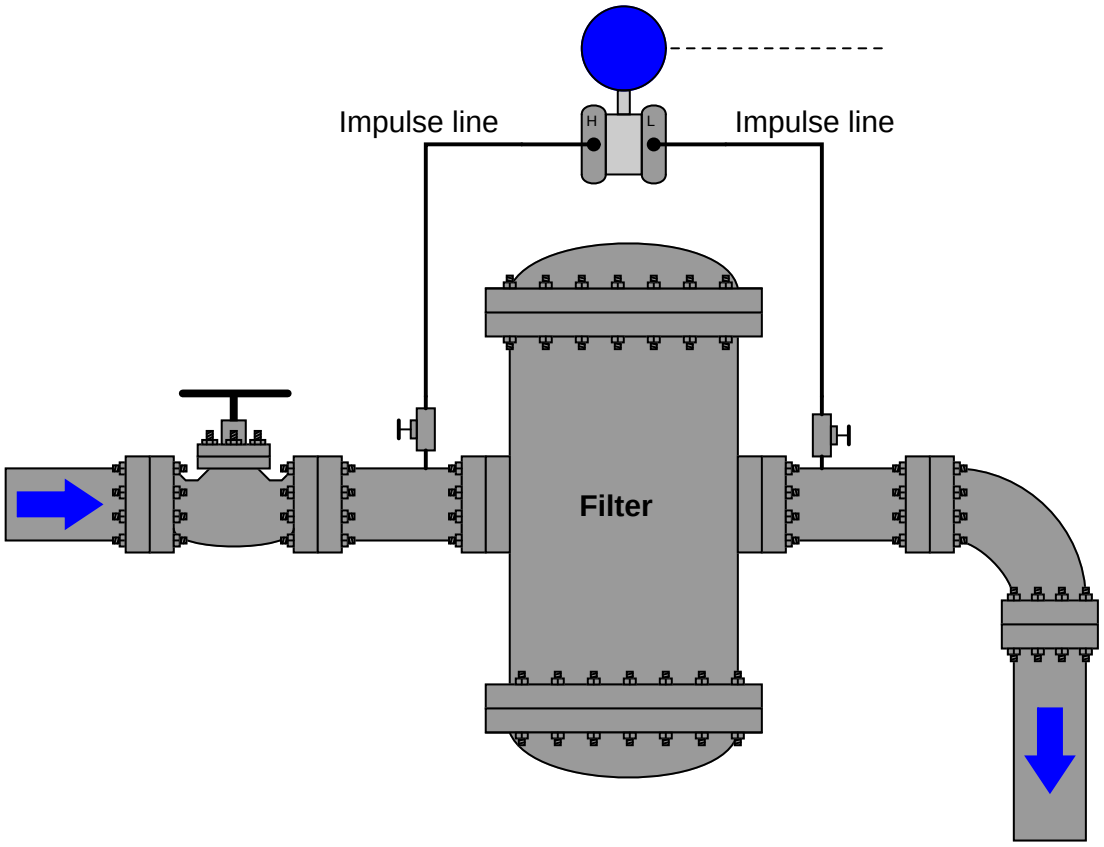
The combination of two differential pressure ports makes the DP transmitter very versatile as a pressure-measuring device. This one instrument may be used to measure pressure differences, positive (gauge) pressures, negative (vacuum) pressures, and even absolute pressures, just by connecting the "high" and "low" sensing ports differently.

In every DP transmitter application, there must be some means of connecting the transmitter's pressure-sensing ports to the points in a process. Metal or plastic tubes (or pipes) work well for this purpose, and are commonly called *impulse lines*, or *gauge lines*, or *sensing lines*[11]. This is equivalent to the test wires used to connect a voltmeter to points in a circuit for measuring voltage. Typically, these tubes are connected to the transmitter and to the process by means of *compression fittings* which allow for relatively easy disconnection and reconnection of tubes. For more information on instrument tube fittings, refer to section 8.2.1 beginning on page 329.

[11] Also called impulse *tubes*, gauge *tubes*, or sensing *tubes*.

Measuring process vessel clogging

We may use the DP transmitter to measure an actual difference of pressure across a process vessel such as a filter, a heat exchanger, or a chemical reactor. The following illustration shows how a differential pressure transmitter may be used to measure clogging of a water filter:

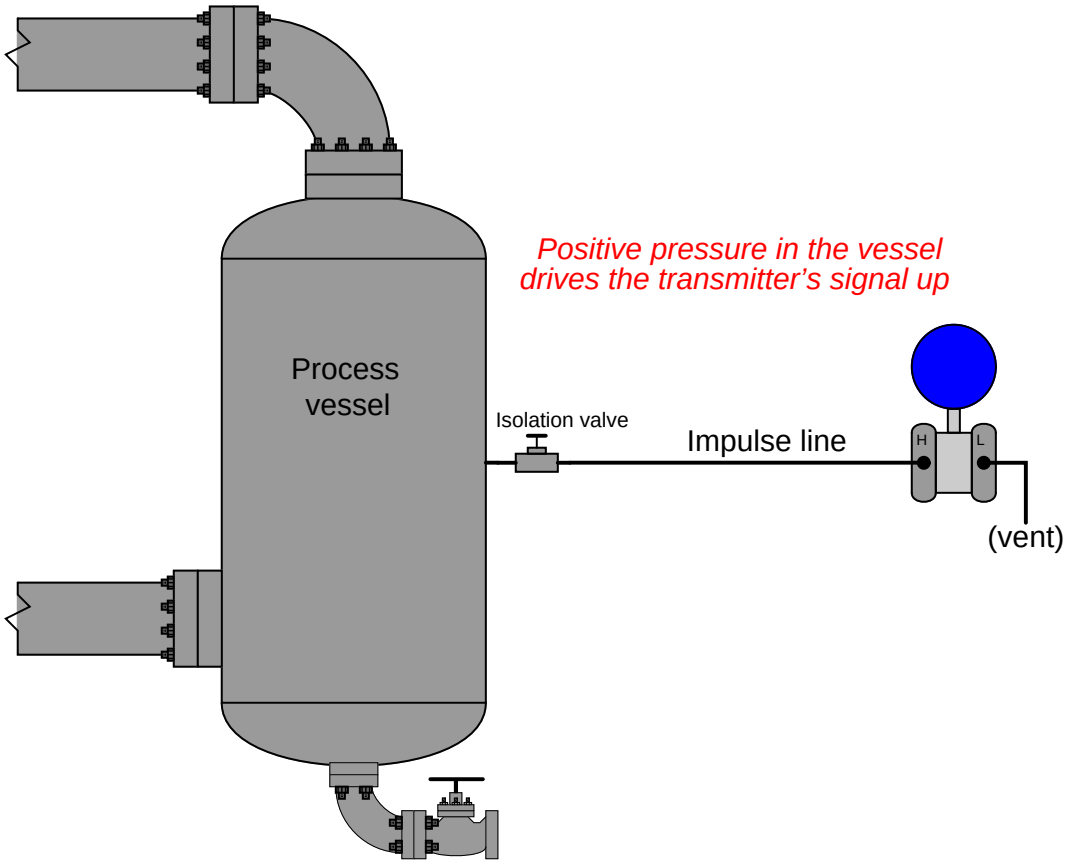


Note how the high side of the DP transmitter connects to the upstream side of the filter, and the low side of the transmitter to the downstream side of the filter. This way, increased filter clogging will result in an increased transmitter output. Since the transmitter's internal pressure-sensing diaphragm only responds to *differences* in pressure between the "high" and "low" ports, the pressure in the filter and pipe relative to the atmosphere is completely irrelevant to the transmitter's output signal. The filter could be operating at a line pressure of 10 PSI or 10,000 PSI – the only variable the DP transmitter measures is the pressure *drop* across the filter. If the upstream side is at 10 PSI and the downstream side is at 9 PSI, the differential pressure will be 1 PSI (sometimes labeled as PSID, "D" for *differential*). If the upstream pressure is 10,000 PSI and the downstream pressure is 9,999 PSI, the DP transmitter will still see a differential pressure of just 1 PSID. Likewise, the technician calibrating the DP transmitter on the workbench could use a precise air pressure of just 1 PSI (applied to the "high" port, with the "low" port vented to atmosphere) to simulate either

of these real-world conditions. The DP transmitter simply cannot tell the difference between these three scenarios, nor should it be able to tell the difference if its purpose is to exclusively measure differential pressure.

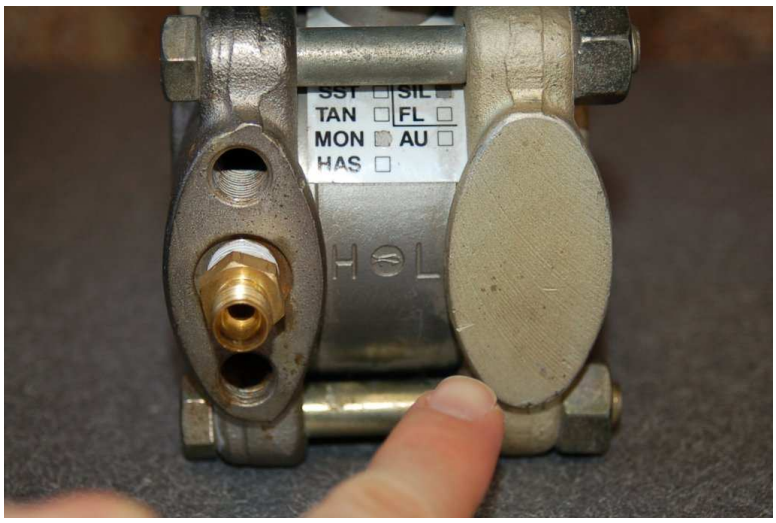
Measuring positive gauge pressure

DP instruments may also serve as simple *gauge pressure* instruments if needed, responding to pressures in excess of atmosphere. If we simply connect the "high" side of a DP instrument to a process vessel using an impulse tube, while leaving the "low" side vented to atmosphere, the instrument will interpret any positive pressure in the vessel as a positive *difference* between the vessel and atmosphere:



Although this may seem like a waste of the transmitter's abilities (why not just use a simpler gauge pressure transmitter with just one port?), it is actually a very common application for DP transmitters. This usage of a differential device may not actually be a "waste" if true-differential applications exist at the same facility for that pressure transmitter, which means only one spare transmitter need be stocked in the facility's warehouse instead of two spare transmitters (one of each type).

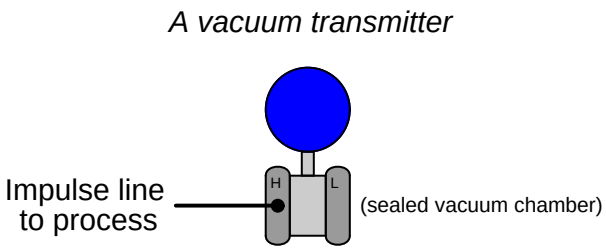
Most DP instrument manufacturers offer "gauge pressure" versions of their differential instruments, with the "high" side port open for connection to an impulse line and the "low" side of the sensing element capped off with a special vented flange, effectively performing the same function we see in the above example at a slightly lesser cost. A close-up photograph of a Rosemount model 1151GP gauge pressure transmitter shows the port-less flange on the "low" side of the pressure-sensing module. Only the "high" side of the sensor has a place for an impulse line to connect:



A closer look at this flange reveals a vent near the bottom, ensuring the "low" side of the pressure-sensing capsule always senses ambient (atmospheric) pressure:

Measuring absolute pressure

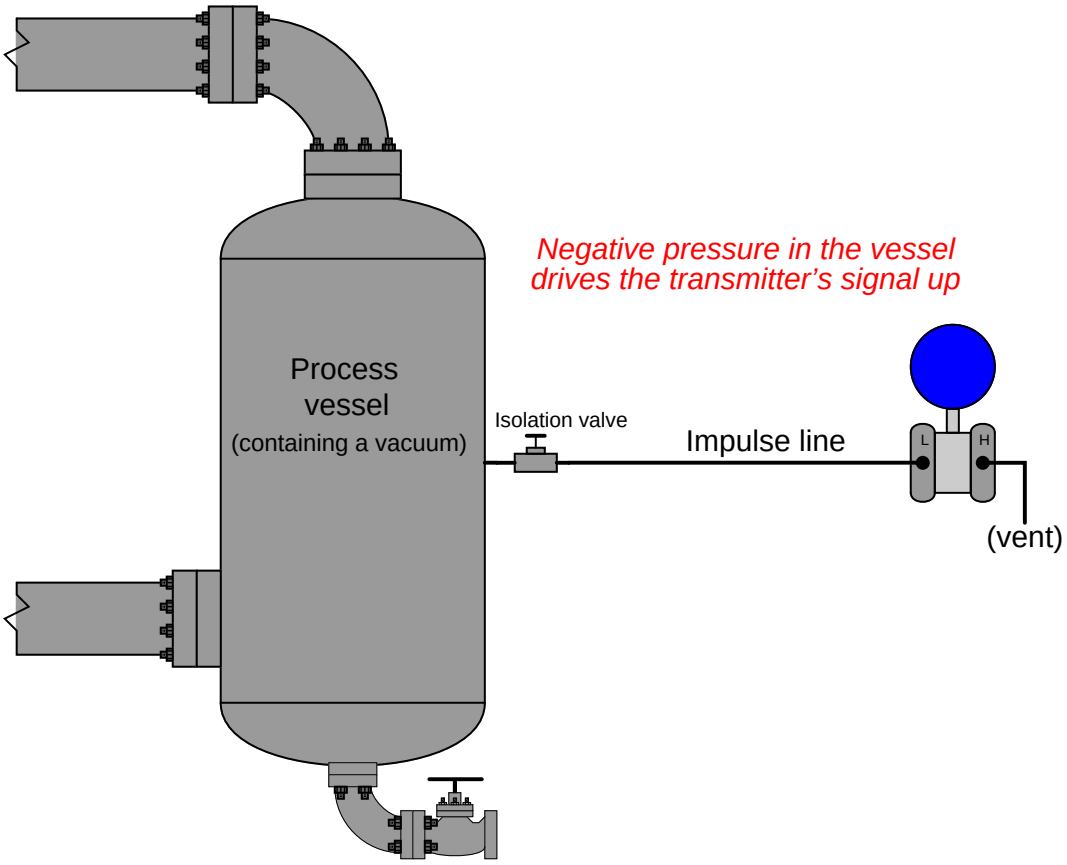
Absolute pressure is defined as the difference between a given fluid pressure and a perfect vacuum. We may build an absolute pressure sensing instrument by taking a DP instrument and sealing the "low" side of its pressure-sensing element in connection to a vacuum chamber. This way, any pressure greater than a perfect vacuum will register as a positive difference:



Most absolute pressure transmitters resemble "gauge pressure" adaptations of DP transmitters, with only one port available to connect an impulse line. Unlike gauge pressure transmitters, though, absolute pressure transmitters do *not* have vent holes on their "low" sides. The "low" side of an absolute pressure transmitter must be a sealed vacuum in order to accurately measure the "high" side fluid pressure in absolute terms.

Measuring vacuum

The same principle of connecting one port of a DP device to a process and venting the other works well as a means of measuring *vacuum* (pressures below that of atmosphere). All we need to do is connect the "low" side to the vacuum process and vent the "high" side to atmosphere:



Any pressure in the process vessel less than atmospheric will register to the DP transmitter as a *positive* difference (with P_{high} greater than P_{low}). Thus, the stronger the vacuum in the process vessel, the greater the signal output by the transmitter.

This last statement deserves some qualification. It used to be, the way analog pneumatic and electronic transmitters were designed many years ago, that the only way to obtain an increasing signal from a DP instrument was to ensure the "high" port pressure *rose* in relation to the "low" port pressure (or conversely stated, to ensure the "low" port pressure *dropped* in relation to the "high" side pressure). However, with the advent of digital electronic technology, it became rather easy to program a DP instrument with a *negative* range, for example 0 to -10 PSI. This way, a *decreasing* pressure as interpreted by the transmitter would yield an *increasing* output signal.

It is rare to find a pressure transmitter calibrated in such a way, but bear in mind that it is possible. This opens the possibility of using a regular "gauge" pressure transmitter (where the

"high" port connects to the process vessel and the "low" port is always vented to atmosphere by virtue of a special flange on the instrument) as a vacuum instrument. If a gauge pressure transmitter is given a negative calibration span, any decreasing pressure seen at the "high" port will yield an increasing output signal.

18.5.2 Inferential measurement applications

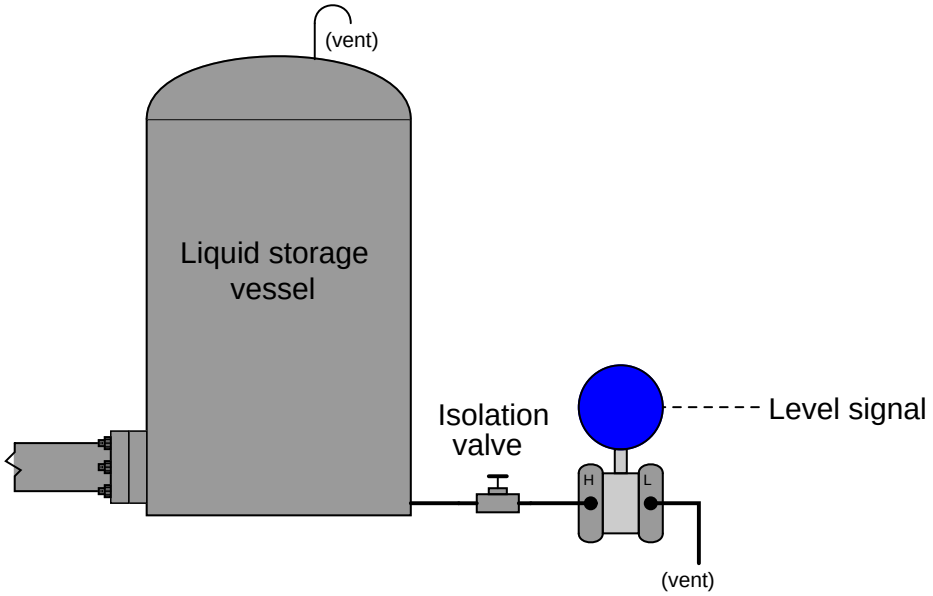
In addition to its usefulness as a direct pressure-measurement device, the DP transmitter may *infer* a great many other process variables known to generate sensible pressures. It is in this capacity that the differential pressure transmitter demonstrates its greatest versatility.

Inferring liquid level

Liquids generate pressure proportional to height (depth) due to their weight. The pressure generated by a vertical column of liquid is proportional to the column height (h), and liquid's mass density (ρ), and the acceleration of gravity (g):

$$P = \rho g h$$

Knowing this, we may use a DP transmitter as a liquid level-sensing device if we know the density of the liquid remains fairly constant[12]:

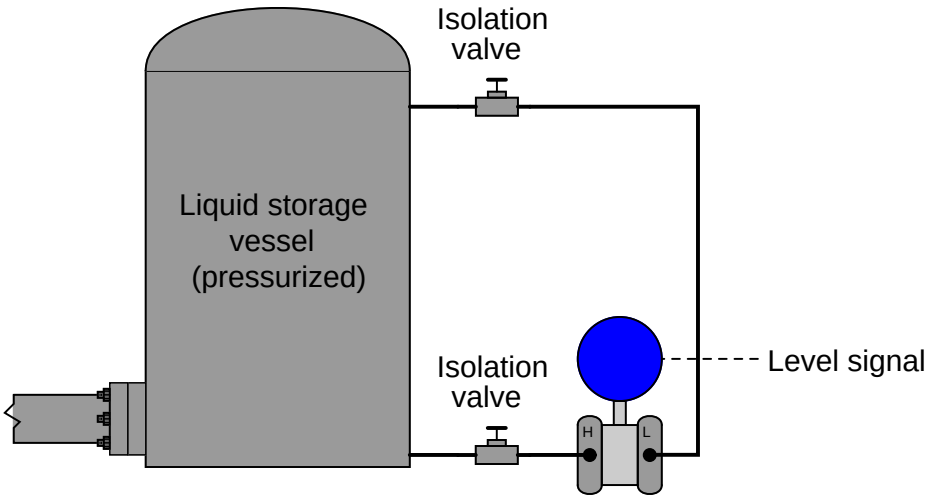


As liquid level in the vessel increases, the amount of hydrostatic pressure applied to the transmitter's "high" port increases in direct proportion. Thus, the transmitter's increasing signal represents the height of liquid inside the vessel:

$$h = \frac{P}{\rho g}$$

[12] We simply assume Earth's gravitational acceleration (g) to be constant as well.

This simple technique works even if the vessel is under pressure from a gas or a vapor (rather than being vented as was the case in the previous example). All we need to do to compensate for this other pressure is to connect the DP transmitter's "low" port to the top of the vessel so it senses nothing but the gas pressure:

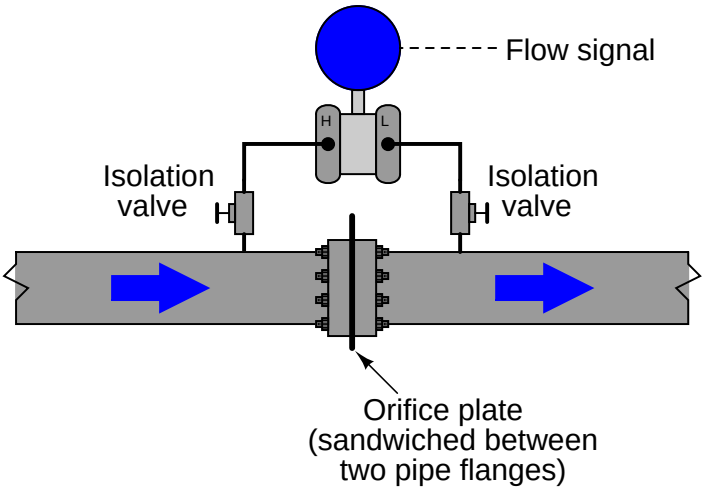


Since the transmitter responds only to differences of pressure between its two sensing ports, and the only cause for a difference of pressure in this application will be pressure generated by the height of a liquid column, the transmitter's signal becomes an exclusive representation of liquid level in the vessel, rejecting potential measurement errors caused by changes in gas pressure within the vessel. Any gas pressure within the vessel will be sensed equally by both ports on the transmitter, thus "canceling" and being of no effect to the measurement. Only changes in liquid level within the vessel will cause the "high" port pressure to change independently of the "low" port pressure, causing the transmitter's output signal to change.

Inferring gas and liquid flow

Another common inferential measurement using DP transmitters is the measurement of fluid flow through a pipe. Pressure dropped across a constriction in the pipe varies in relation to flow rate (Q) and fluid density (ρ). So long as fluid density remains fairly constant, we may measure pressure drop across a piping constriction and use that measurement to infer flow rate.

The most common form of constriction used for this purpose is called an *orifice plate*, being nothing more than a metal plate with a precisely machined hole in the center. As fluid passes through this hole, its velocity changes, causing a pressure drop to form:



Once again, we see the common-mode rejection abilities of the pressure transmitter used for practical advantage. Since both ports of the transmitter connect to the same process line, static fluid pressure within that line has no effect on the measurement. Only *differences* of pressure between the upstream and downstream sides of the constriction (orifice plate) cause the transmitter to register flow.

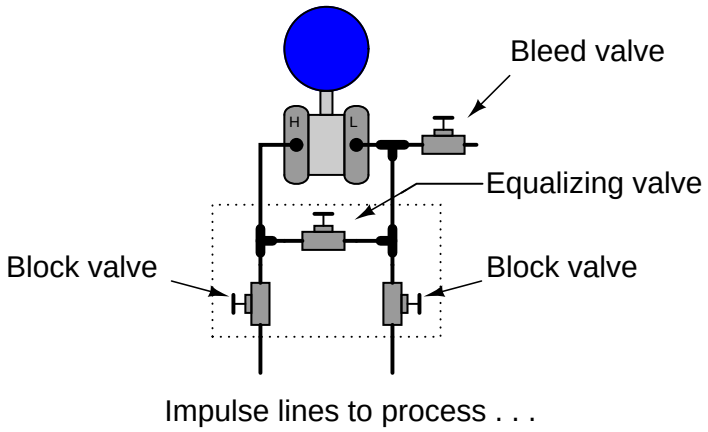
18.6 Pressure sensor accessories

Multiple accessories exist for pressure-sensing devices to function optimally in challenging process environments. Sometimes, we must use special accessories to protect the pressure instrument against hazards of certain process fluids. One such hazard is pressure *pulsation*, for example at the discharge of a piston-type (positive-displacement) high-pressure pump. Pulsating pressure can quickly damage mechanical sensors such as bourdon tubes, either by wear of the mechanism transferring pressure element motion to an indicating needle, and/or fatigue of the metal element itself.

18.6.1 Valve manifolds

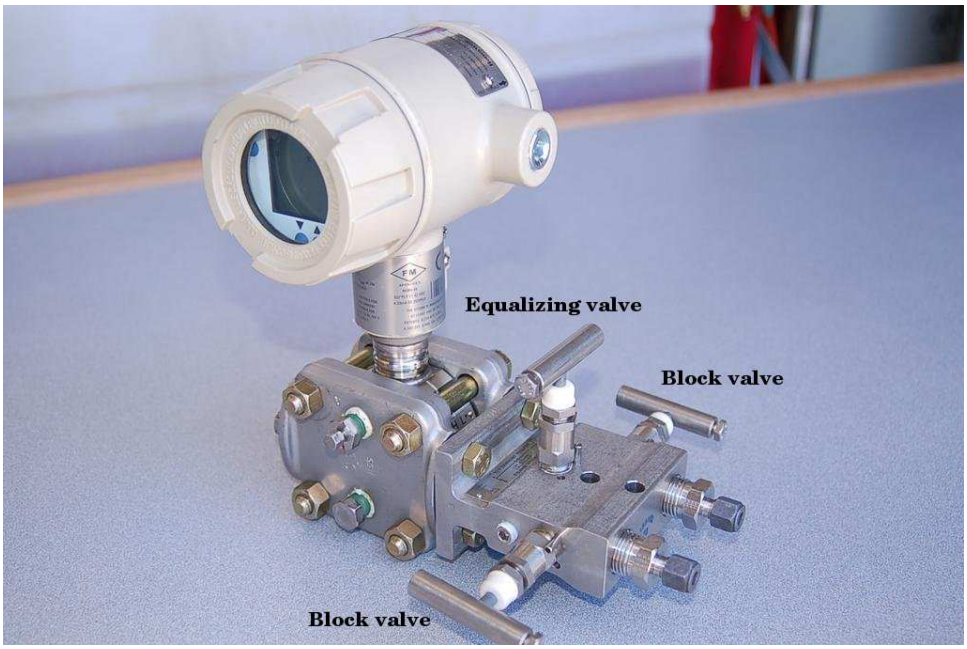
An important accessory to the DP transmitter is the *three-valve manifold*. This device incorporates three manual valves to isolate and equalize pressure from the process to the transmitter, for maintenance and calibration purposes.

The following illustration shows the three valves comprising a three-valve manifold (within the dotted-line box), as well as a fourth valve called a "bleed" valve used to vent trapped fluid pressure to atmosphere:

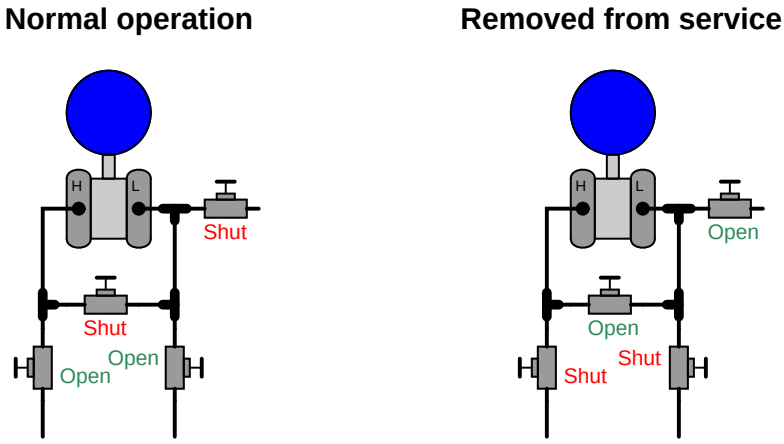


While this illustration shows the three valves as separate devices, connected together and to the transmitter by tubing, three-valve manifolds are more commonly manufactured as monolithic devices: the three valves cast together into one block of metal, attaching to the pressure transmitter by way of a flanged face with O-ring seals. Bleed valves are most commonly found as separate devices threaded into one or more of the ports on the transmitter's diaphragm chambers.

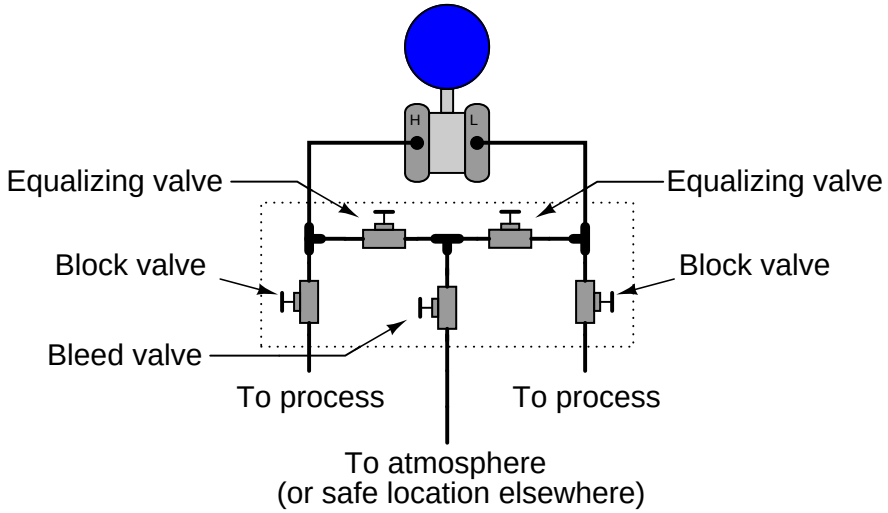
The following photograph shows a three-valve manifold bolted to a Honeywell model ST3000 differential pressure transmitter. A bleed valve fitting may be seen inserted into the upper port on the nearest diaphragm capsule flange:



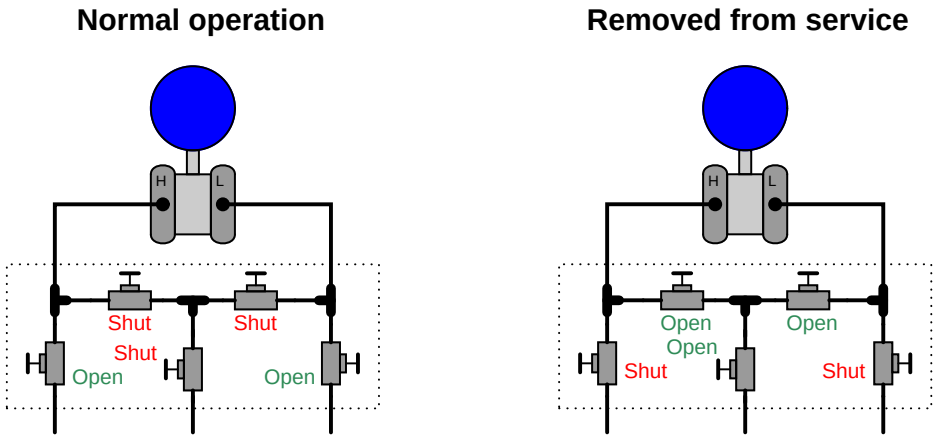
In normal operation, the two block valves are left open to allow process fluid pressure to reach the transmitter. The equalizing valve is left tightly shut so no fluid can pass between the "high" and "low" pressure sides. To isolate the transmitter from the process for maintenance, one must first close the block valves, then open the equalizing valve to ensure the transmitter "sees" no differential pressure. The "bleed" valve is opened at the very last step to relieve pent-up fluid pressure within the manifold and transmitter chambers:



A variation on this theme is the *five-valve manifold*, shown in this illustration:



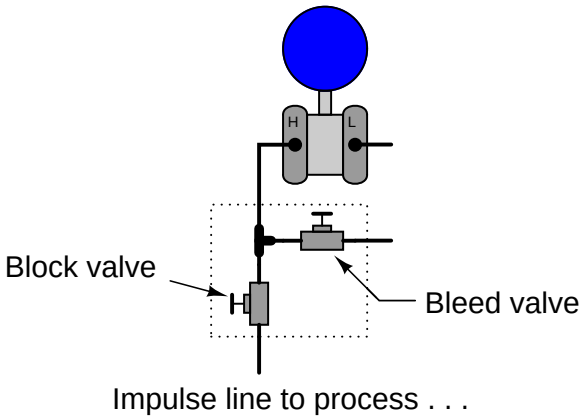
Manifold valve positions for normal operation and maintenance are as follows:



It is critically important that the equalizing valve(s) never be open while both block valves are open! Doing so will allow process fluid to flow through the equalizing valve(s) from the high-pressure side of the process to the low-pressure side of the process. If the impulse tubes connecting the manifold to the process are intentionally filled with a *fill fluid* (such as glycerin, to displace process water from entering the impulse tubes; or water in a steam system), this fill fluid will be lost. Also, if the process fluid is dangerously hot or radioactive, a combination of open equalizing and block valves will let that dangerous fluid reach the transmitter and manifold, possibly causing damage or creating a personal hazard. Speaking from personal experience, I once made this mistake on a DP transmitter connected to a steam system, causing hot steam to flow through the manifold and overheat the equalizing valve so that it seized open and could not be shut again! The only way

I was able to stop the flow of hot steam through the manifold was to locate and shut a sliding-gate hand valve between the impulse tube and the process pipe. Fortunately, this cast iron valve was not damaged by the heat and was still able to shut off the flow.

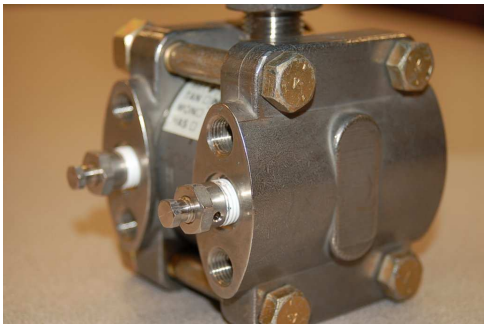
Pressure transmitter valve manifolds also come in single block-and-bleed configurations, for gauge pressure applications. Here, the "low" pressure port of the transmitter is vented to atmosphere, with only the "high" pressure port connected to the impulse line:

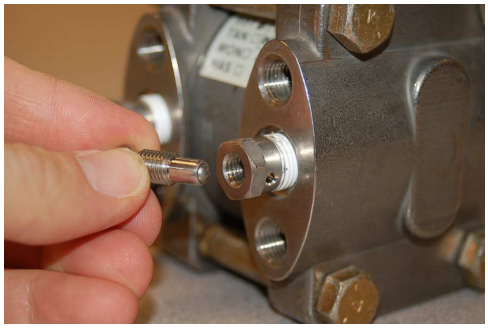


The following photograph shows a bank of eight pressure transmitters, seven out of the eight being equipped with a single block-and-bleed manifold. The eighth transmitter (bottom row, second-from left) sports a 5-valve manifold:

18.6.2 Bleed (vent) fittings

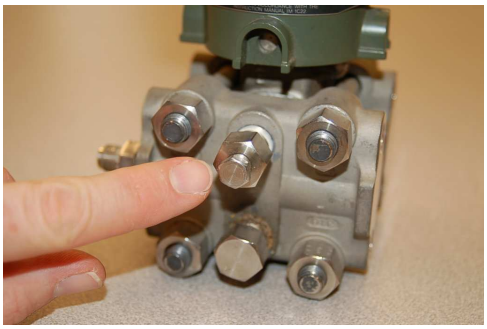
Before removing a pressure transmitter from live service, the technician must "bleed" or "vent" stored fluid pressure to atmosphere in order to achieve a *zero energy state* prior to disconnecting the transmitter from the impulse lines. Some valve manifolds provide a bleed valve for doing just this, but many do not[13]. An inexpensive and common accessory for pressure-sensing instruments (especially transmitters) is the *bleed valve fitting* or *vent valve fitting*, installed on the instrument as a discrete device. The most common bleed fitting is equipped with 1/4 inch male NPT pipe threads, for installation into one of the 1/4 inch NPT threaded pipe holes typically provided on pressure transmitter flanges. The bleed is operated with a small wrench, loosening a ball-tipped plug off its seat to allow process fluid to escape through a small vent hole in the side of the fitting. The following photographs show close-up views of a bleed fitting both assembled (left) and with the plug fully extracted from the fitting (right). The bleed hole may be clearly seen in both photographs:

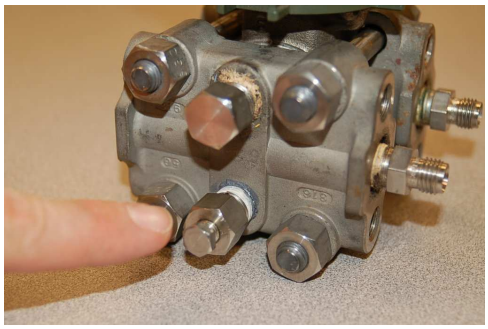




When installed directly on the flanges of a pressure instrument, these bleed valves may be used to bleed unwanted fluids from the pressure chambers, for example bleeding air bubbles from an instrument intended to sense water pressure, or bleeding condensed water out of an instrument intended to sense compressed air pressure.

The following photographs show bleed fittings installed two different ways on the side of a pressure transmitter flange, one way to bleed gas out of a liquid process (located on top) and the other way to bleed liquid out of a gas process (located on bottom):



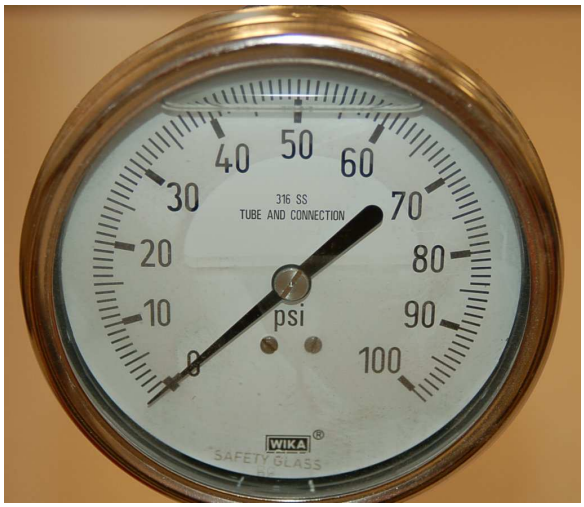


[13]The standard 3-valve manifold, for instance, does not provide a bleed valve – only block and equalizing valves.

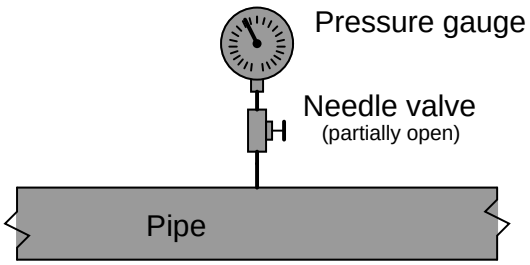
18.6.3 Pressure pulsation damping

A simple way to mitigate the effects of pulsation on a pressure gauge is to fill the inside of the gauge with a viscous liquid such as glycerin or oil. The inherent friction of this fill liquid has a "shock-absorber" quality which damps the gauge mechanism's oscillatory motion and helps protect against damage from pulsations or from external vibration. This method is ineffectual for high-amplitude pulsations, though.

An oil-filled pressure gauge may be seen in the following photograph. Note the air bubble near the top of the gauge face, which is the only visual indication of an oil filling:

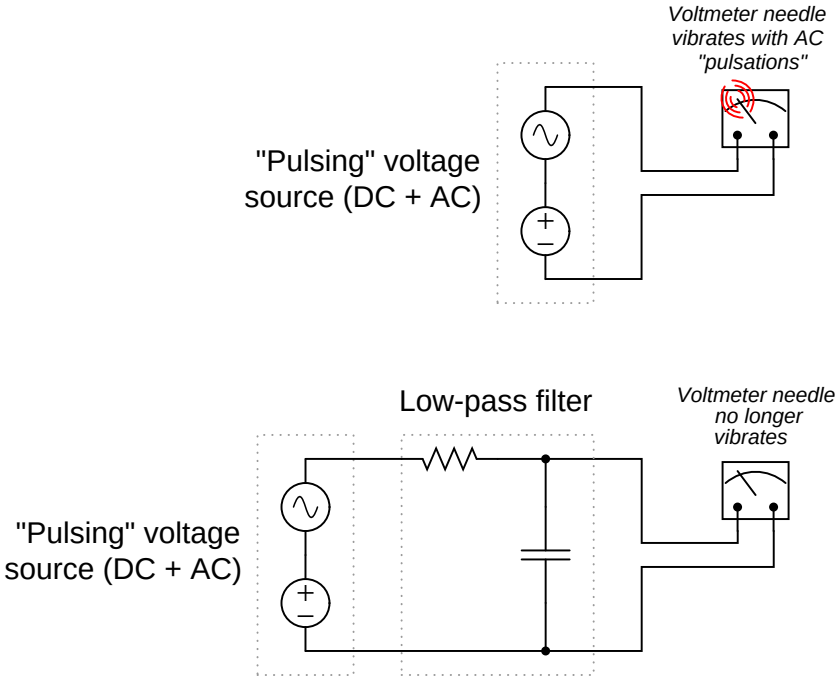


A more sophisticated method for damping pulsations seen by a pressure instrument is called a *snubber*, and it consists of a fluid restriction placed between with the pressure sensor and the process. The simplest example of a snubber is a simple *needle valve* (an adjustable valve designed for low flow rates) placed in a mid-open position, restricting fluid flow in and out of a pressure gauge:



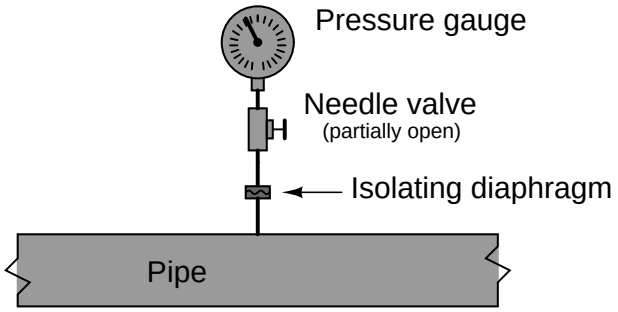
At first, the placement of a throttling valve between the process and a pressure-measuring instrument seems rather strange, because there should not be any continuous flow in or out of the gauge for such a valve to throttle! However, a *pulsing* pressure causes a small amount of *alternating* flow in and out of the pressure instrument, owing to the expansion and contraction of the mechanical pressure-sensing element (bellows, diaphragm, or bourdon tube). The needle valve provides a restriction for this flow which, when combined with the fluid capacitance of the pressure

instrument, combine to form a low-pass filter of sorts. By impeding the flow of fluid in and out of the pressure instrument, that instrument is prevented from "seeing" the high and low peaks of the pulsating pressure. Instead, the instrument registers a much steadier pressure over time. An electrical analogy for a pressure snubber is an RC low-pass filter circuit "damping" voltage pulsations from reaching a voltmeter:



One potential problem with the needle valve solution is that the small orifice inside the valve may plug up over time with debris or deposits from dirty process fluid. This, of course, would be bad because if that valve were to ever completely plug, the pressure instrument would stop responding to any changes in process pressure at all, or perhaps just become too slow in responding to major changes.

A solution to this problem is to fill the pressure sensor mechanism with a clean liquid (called a *fill fluid*), then transfer pressure from the process fluid to the fill fluid (and then to the pressure-sensing element) using a slack diaphragm or some other membrane:

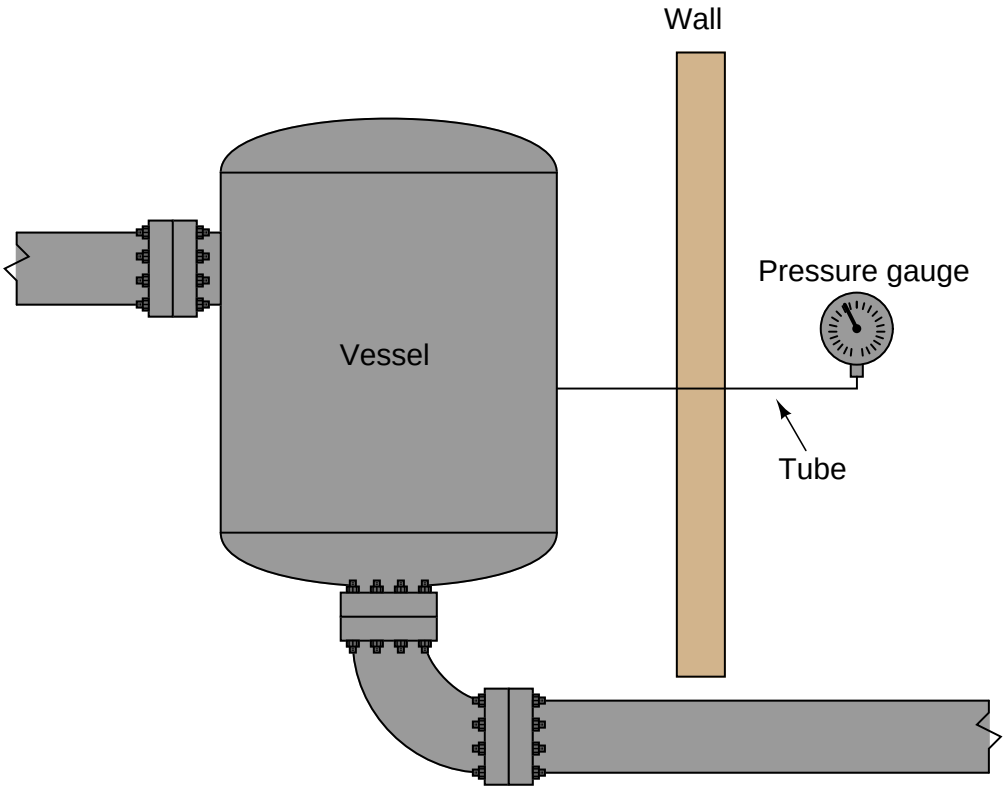


In order for the fill fluid and isolating diaphragm to work effectively, there cannot be any gas bubbles in the fill fluid – it must be a "solid" hydraulic system from the diaphragm to the sensing element. The presence of gas bubbles means that the fill fluid is compressible, which means the isolating diaphragm may have to move more than necessary to transfer pressure to the instrument's sensing element. This will introduce pressure measurement errors if the isolating diaphragm begins to tense from excessive motion (and thereby oppose some process fluid pressure from fully transferring to the fill fluid), or hit a "stop" point where it cannot move any further (thereby preventing any further transfer of pressure from process fluid to fill fluid)[14]. For this reason, isolating diaphragm systems for pressure instruments are usually "packed" with fill fluid at the point and time of manufacture, then sealed in such a way that they cannot be opened for any form of maintenance. Consequently, any fill fluid leak in such a system immediately ruins it.

[14] This concept will be immediately familiar to anyone who has ever had to "bleed" air bubbles out of an automobile brake system. With air bubbles in the system, the brake pedal has a "spongy" feel when depressed, and much pedal motion is required to achieve adequate braking force. After bleeding all air out of the brake fluid tubes, the pedal motion feels much more "solid" than before, with minimal motion required to achieve adequate braking force. Imagine the brake pedal being the isolating diaphragm, and the brake pads being the pressure sensing element inside the instrument. If enough gas bubbles exist in the tubes, the brake pedal might stop against the floor when fully pressed, preventing full force from ever reaching the brake pads! Likewise, if the isolating diaphragm hits a hard motion limit due to gas bubbles in the fill fluid, the sensing element will not experience full process pressure!

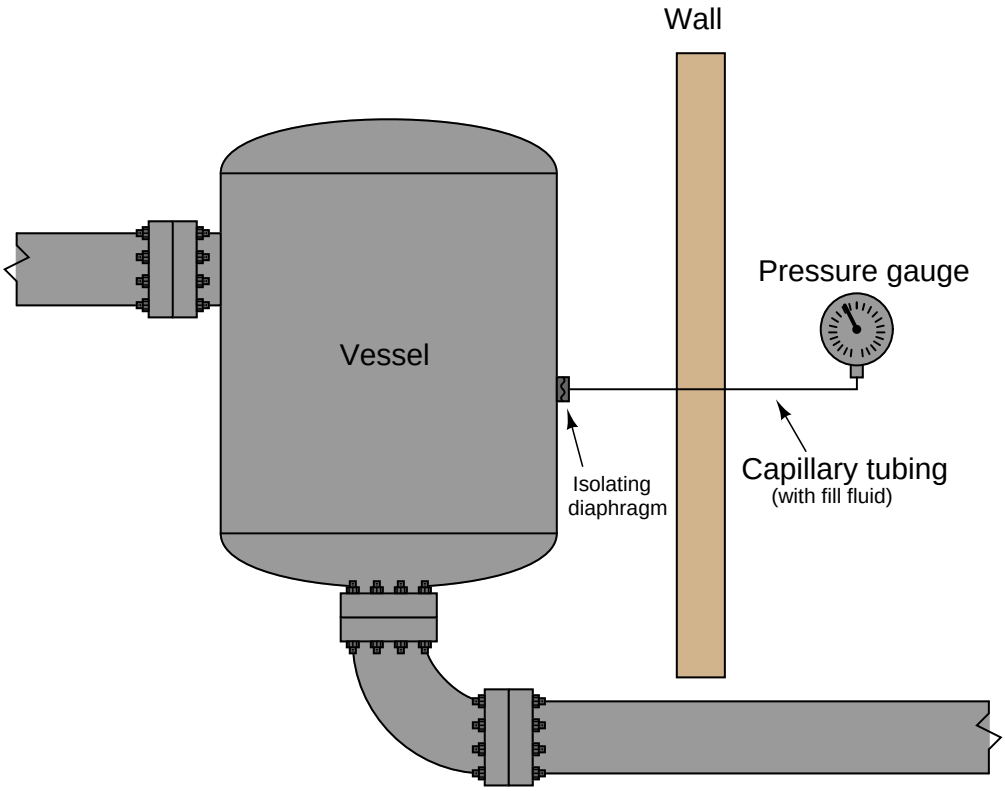
18.6.4 Remote and chemical seals

Isolating diaphragms have merit even in scenarios where pressure pulsations are not a problem. Consider the case of a food-processing system where we must remotely measure pressure inside a mixing vessel:



The presence of the tube connecting the vessel to the pressure gauge poses a hygiene problem. Stagnant process fluid (in this case, some liquid food product) inside the tube can support microbial growth, which will eventually contaminate the vessel no matter how well or how often the vessel is cleaned. Even automated *Clean-In-Place* and *Steam-In-Place* (*CIP* and *SIP*, respectively) protocols where the vessel is chemically purged between batches cannot prevent this problem because the cleaning agents never purge the entire length of the tubing (ultimately, to the bourdon tube or other sensing element inside the gauge).

Here, we see a valid application of an isolating diaphragm and fill fluid. If we mount an isolating diaphragm to the vessel in such a way that the process fluid directly contacts the diaphragm, sealed fill fluid will be the only material inside the tubing carrying that pressure to the instrument. Furthermore, the isolating diaphragm will be directly exposed to the vessel interior, and therefore cleaned with every CIP cycle. Thus, the problem of microbial contamination is completely avoided:



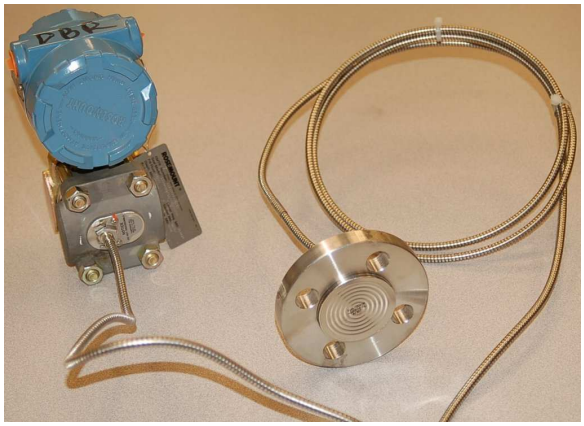
Such systems are often referred to as *remote seals*, and they are available on a number of different pressure instruments including gauges, transmitters, and switches. If the purpose of an isolating diaphragm and fill fluid is to protect the sensitive instrument from corrosive or otherwise harsh chemicals, it is often referred to as a *chemical seal*.

The following photograph shows a pressure gauge equipped with a chemical seal diaphragm. Note that the chemical seal on this particular gauge is close-coupled to the gauge, since the only goal here is protection of the gauge from harsh process fluids, not the ability to remotely mount the gauge:

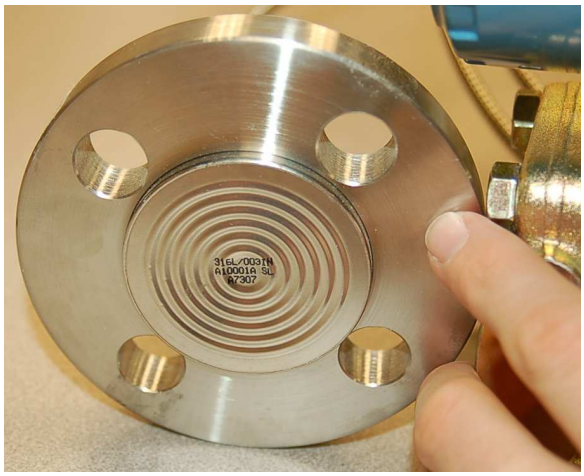
A view facing the bottom of the flange reveals the thin metal isolating diaphragm keeping process fluid from entering the gauge mechanism. Only inert fill fluid occupies the space between this diaphragm and the gauge's bourdon tube:

The only difference between this chemical-seal gauge and a remote-seal gauge is that a remote-seal gauge uses a length of very small-diameter tubing called *capillary tubing* to transfer fill fluid pressure from the sealing diaphragm to the gauge mechanism.

Direct-reading gauges are not the only type of pressure instrument that may benefit from having remote seals. Electronic pressure *transmitters* are also manufactured with remote seals for the same reasons: protection of the transmitter sensor from harsh process fluid, or prevention of "dead-end" tube lengths where organic process fluid would stagnate and harbor microbial growths. The following photograph shows a pressure transmitter equipped with a remote sealing diaphragm. The capillary tube is protected by a coiled metal ("armor") sheath:



A close-up view of the sealing diaphragm shows its corrugated design, allowing the metal to easily flex and transfer pressure to the fill fluid within the capillary tubing[15]:



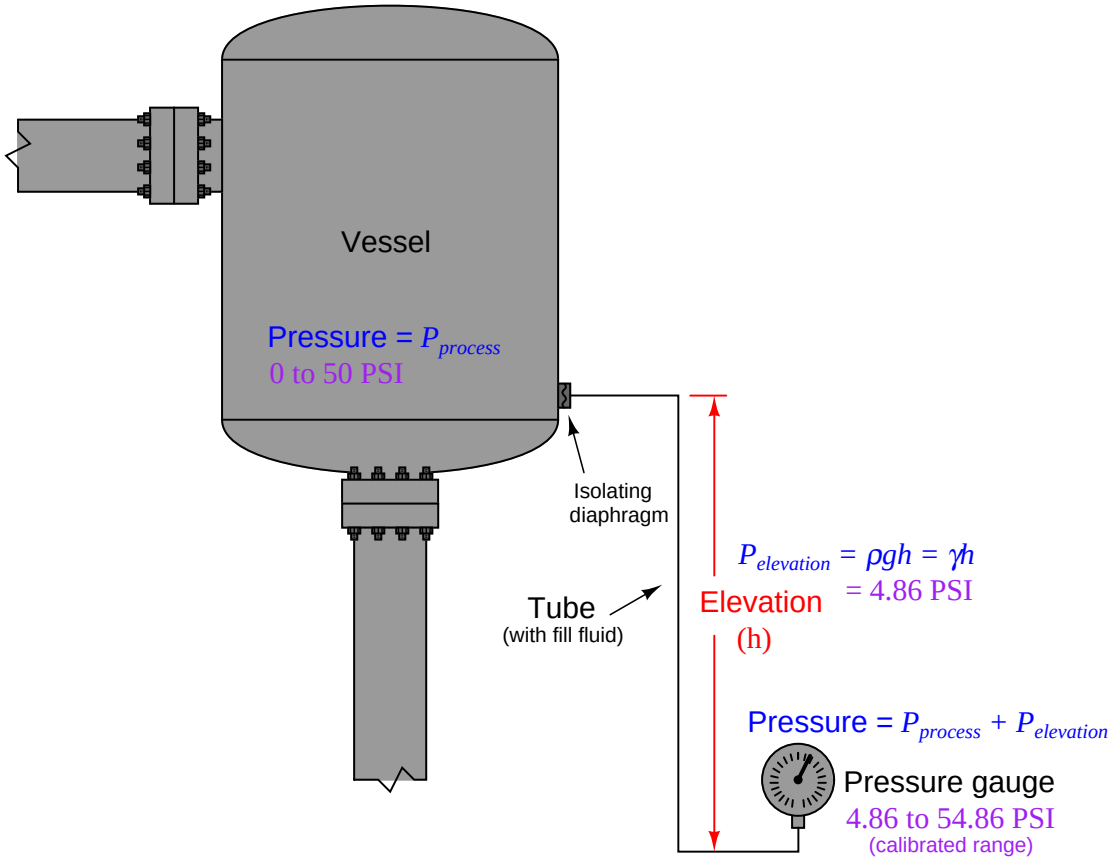
Just like the isolating diaphragms of the pressure-sensing capsule, these remote diaphragms need only transfer process fluid pressure to the fill fluid and (ultimately) to the taut sensing diaphragm inside the instrument. Therefore, these diaphragms perform their function best if they are designed to easily flex. This allows the taut sensing diaphragm to provide the vast majority of the opposing force to the fluid pressure, as though it were the only spring element in the fluid system.

[15] Like all instrument diaphragms, this one is sensitive to damage from contact with sharp objects. If the diaphragm ever becomes nicked, dented, or creased, it will tend to exhibit hysteresis in its motion, causing calibration errors for the instrument. For this reason, isolating diaphragms are often protected from contact by a plastic plug when the instrument is shipped from the manufacturer. This plug must be removed from the instrument before placing it into service.

The connection point between the capillary tube and the transmitter's sensor capsule is labeled with a warning never to disassemble, since doing so would allow air to enter the filled system (or fill fluid to escape from the system) and thereby ruin its accuracy:

In order for a remote seal system to work, the hydraulic "connection" between the sealing diaphragm and the pressure-sensing element must be completely gas-free so there will be a "solid" transfer of motion from one end to the other.

A potential problem with using remote diaphragms is the hydrostatic pressure generated by the fill fluid if the pressure instrument is located far away (vertically) from the process connection point. For example, a pressure gauge located far below the vessel it connects to will register a *greater* pressure than what is actually inside the vessel, because the vessel's pressure adds to the hydrostatic pressure caused by the liquid in the tubing:



This pressure may be calculated by the formula ρgh or γh where ρ is the mass density of the fill liquid or γ is the weight density of the fill liquid. For example, a 12 foot capillary tube height filled with a fill liquid having a weight density of 58.3 lb/ft^3 will generate an elevation pressure of almost 700 lb/ft^2, or 4.86 PSI. If the pressure instrument is located below the process connection point, this 4.86 PSI offset must be incorporated into the instrument's calibration range. If we desire this pressure instrument to accurately measure a process pressure range of 0 to 50 PSI, we would have to calibrate it for an actual range of 4.86 to 54.86 PSI.

The reverse problem exists where the pressure instrument is located *higher* than the process connection: here the instrument will register a *lower* pressure than what is actually inside the vessel, offset by the amount predicted by the hydrostatic pressure formulae ρgh or γh.

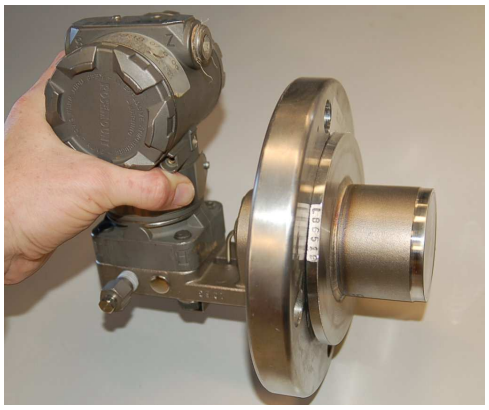
In all fairness, this problem is not limited to remote seal systems – even non-isolated systems

where the tubing is filled with process liquid will exhibit this offset error. However, in filled-capillary systems a vertical offset is *guaranteed* to produce a pressure offset because fill fluids are always liquid, and liquids generate pressure in direct proportion to the vertical height of the liquid column (and to the density of that liquid).

A similar problem unique to isolated-fill pressure instruments is measurement error caused by temperature extremes. Suppose the liquid-filled capillary tube of a remote seal pressure instrument comes too near a hot steam pipe, furnace, or some other source of high temperature. The expansion of the fill fluid may cause the isolation diaphragm to extend to the point where it begins to tense and add a pressure to the fill fluid above and beyond that of the process fluid. Cold temperatures may wreak havoc with filled capillary tubes as well, if the fill fluid congeals or even freezes such that it no longer flows as it should.

Proper mounting of the instrument and proper selection of the fill fluid[16] will help to avoid such problems. All in all, the potential for trouble with remote- and chemical-seal pressure instruments is greatly offset by their benefits in the right applications.

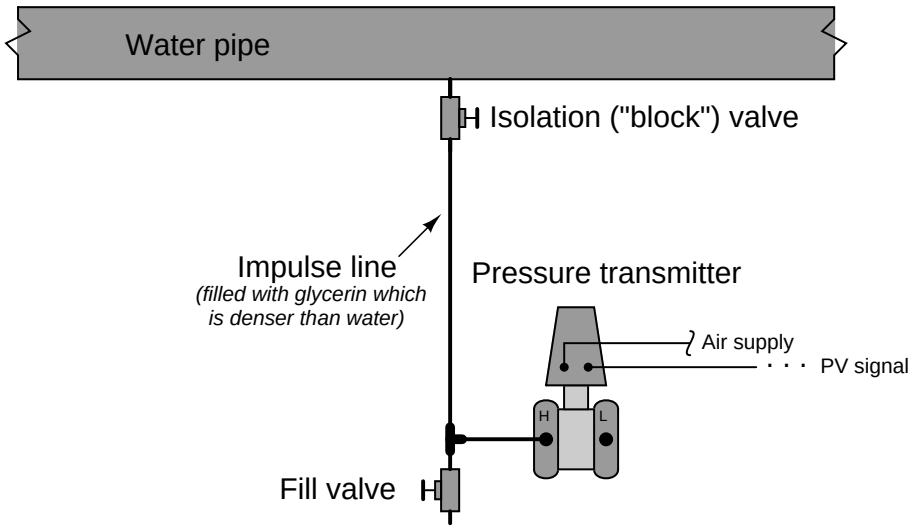
Some pressure transmitters are equipped with close-coupled seals rather than remote seals, for applications where it is best to avoid impulse tube connections to the process (e.g. liquids that tend to coagulate or in hygienic processes). A Rosemount extended-diaphragm pressure transmitter appears in the left-hand photograph, while a Yokogawa transmitter of the same basic design is shown installed in a working process in the right-hand photograph:



[16] Most pressure instrument manufacturers offer a range of fill fluids for different applications. Not only is temperature a consideration in the selection of the right fill fluid, but also potential contamination of or reaction with the process if the isolating diaphragm ever suffers a leak!

18.6.5 Filled impulse lines

An alternate method for isolating a pressure-sensing instrument from direct contact with process fluid is to either *fill* or *purge* the impulse lines with a harmless fluid. Filling impulse tubes with a static fluid works when gravity is able to keep the fill fluid in place, such as in this example of a pressure transmitter connected to a water pipe by a glycerin-filled impulse line:

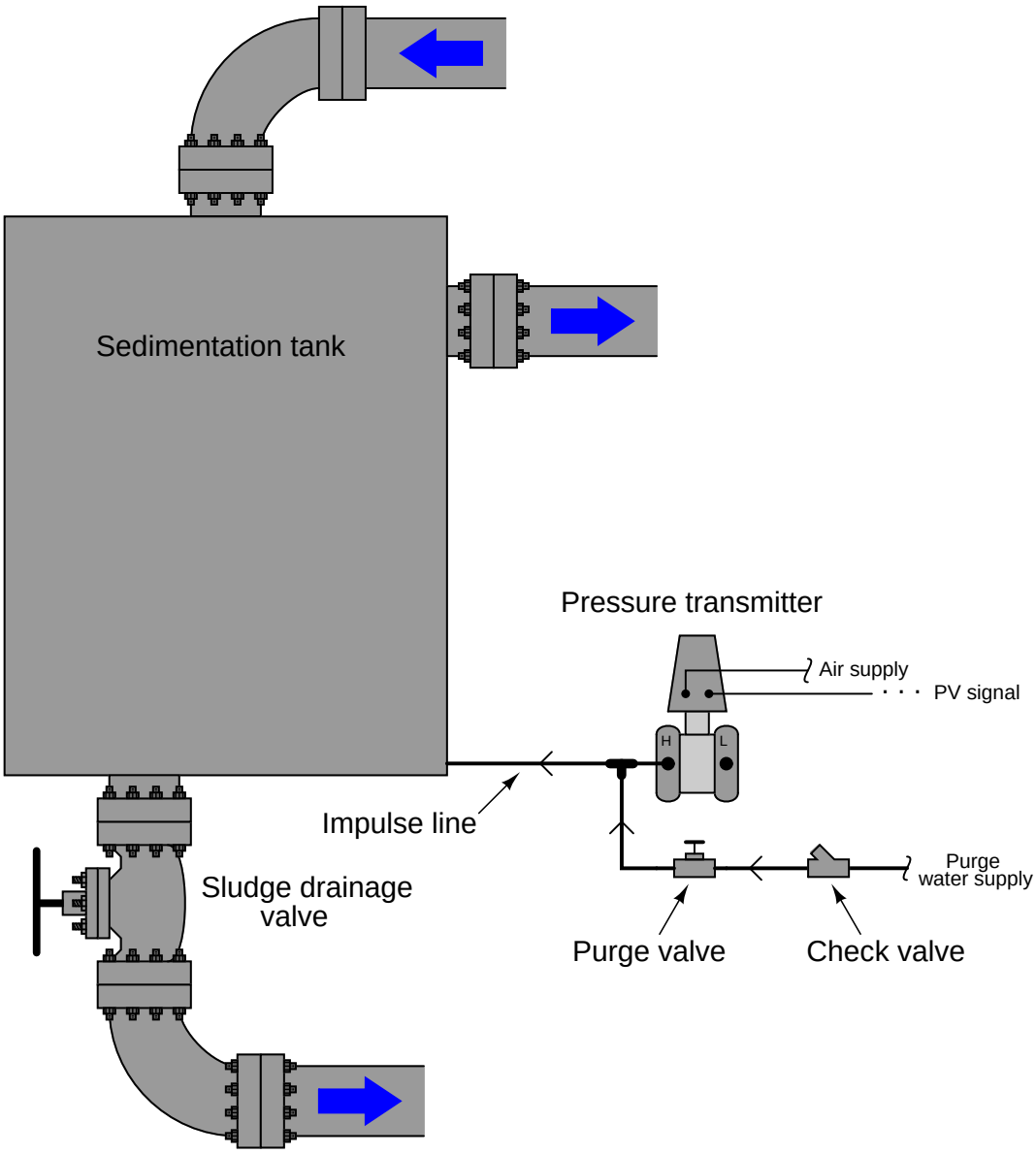


A reason someone might do this is for freeze protection, since glycerin freezes at a lower temperature than water. If the impulse line were filled with process water, it might freeze solid in cold weather conditions (the water in the pipe cannot freeze so long as it is forced to flow). The greater density of glycerin keeps it placed in the impulse line, below the process water line. A fill valve is provided near the transmitter so a technician may re-fill the impulse line with glycerin (using a hand pump) if ever needed.

As with a remote diaphragm, a filled impulse line will generate its own pressure proportional to the height difference between the point of process connection and the pressure-sensing element. If the height difference is substantial, the pressure offset resulting from this difference in elevation will require compensation by means of an intentional "zero shift" of the pressure instrument when it is calibrated.

18.6.6 Purged impulse lines

Continuous purge of an impulse line is an option when the line is prone to plugging. Consider this example, where pressure is measured at the bottom of a sedimentation vessel:



A continuous flow of clean water enters through a "purge valve" and flows through the impulse line, keeping it clear of sediment while still allowing the pressure instrument to sense pressure at the bottom of the vessel. A *check valve* guards against reverse flow through the purge line, in case process fluid pressure ever exceeds purge supply pressure. Purged systems are very useful, but a few details are necessary to consider before deciding to implement such a strategy:

- How reliable is the supply of purge fluid? If this stops for any reason, the impulse line may plug!
- Is the purge fluid supply pressure guaranteed to exceed the process pressure at all times, for proper direction of purge flow?
- What options exist for purge fluids that will not adversely react with the process?
- What options exist for purge fluids that will not contaminate the process?
- How expensive will it be to maintain this constant flow of purge fluid into the process?

Also, it is important to limit the flow of purge fluid to a rate that will not create a falsely high pressure measurement due to restrictive pressure drop across the length of the impulse line, yet flow freely enough to achieve the goal of plug prevention. In many installations, a visual flow indicator is installed in the purge line to facilitate optimum purge flow adjustment. Such flow indicators are also helpful for troubleshooting, as they will indicate if anything happens to stop the purge flow.

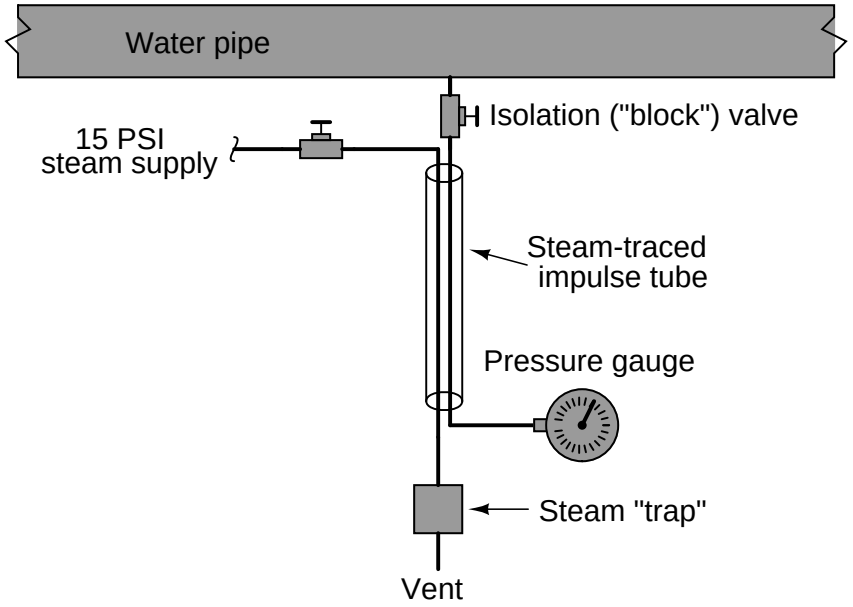
In the previous example, the purge fluid was clean water. Many options exist for purge fluids other than water, though. Gases such as air, nitrogen, or carbon dioxide are often used in purged systems, for both gas and liquid process applications.

Purged impulse lines, just like filled lines and diaphragm-isolated lines, will generate hydrostatic pressure with vertical height. If the purge fluid is a liquid, this elevation-dependent pressure may be an offset to include in the instrument's calibration. If the purge fluid is a gas (such as air), however, any height difference may be ignored because the density of the gas is negligible.

18.6.7 Heat-traced impulse lines

If impulse lines are filled with liquid, there may exist a possibility for that liquid to freeze in cold-weather conditions. This possibility depends, of course, on the type of liquid filling the impulse lines and how cold the weather gets in that geographic location.

One safeguard against impulse line freezing is to *trace* the impulse lines with some form of active heating medium, steam and electrical being the most common. "Steam tracing" consists of a copper tube carrying low-pressure steam, bundled alongside one or more impulse tubes, enclosed in a thermally insulating jacket.



Steam flows through the shut-off valve, through the tube in the insulated bundle, transferring heat to the impulse tube as it flows past. Cooled steam condenses into water and collects in the *steam trap* device located at the lowest elevation on the steam trace line. When the water level builds up to a certain level inside the trap, a float-operated valve opens to vent the water. This allows more steam to flow into the tracing tube, keeping the impulse line continually heated.

The steam trap naturally acts as a sort of thermostat as well, even though it only senses condensed water level and not temperature. The rate at which steam condenses into water depends on how cold the impulse tube is. The colder the impulse tube (caused by colder ambient conditions), the more heat energy drawn from the steam, and consequently the faster condensation rate of steam into water. This means water will accumulate faster in the steam trap, which means it will "blow down" more often. More frequent blow-down events means a greater flow rate of steam into the tracing tube, which adds more heat to the tubing bundle and raises its temperature. Thus, the system is naturally regulating, with its own negative feedback loop to maintain bundle temperature at a relatively stable point[17].

[17] In fact, after you become accustomed to the regular "popping" and "hissing" sounds of steam traps blowing down, you can interpret the blow-down frequency as a crude ambient temperature thermometer! Steam traps seldom

The following photograph shows a picture of a steam trap:

Steam traps are not infallible, being susceptible to freezing (in *very* cold weather) and sticking open (wasting steam by venting it directly to atmosphere). However, they are generally reliable devices, capable of adding tremendous amounts of heat to impulse tubing for protection against freezing.

blow down during warm weather, but their "popping" is much more regular (one every minute or less) when ambient temperatures drop well below the freezing point of water.

Electrically traced impulse lines are an alternative solution for cold-weather problems. The "tracing" used is a twin-wire cable (sometimes called *heat tape*) that acts as a resistive heater. When power is applied, the cable heats up, thus imparting thermal energy to the impulse tubing it is bundled with. This next photograph shows the end of a section of electrical heat tape, rated at 33 watts per meter (10 watts per foot) at 10 degrees Celsius (50 degrees Fahrenheit):

This particular heat tape also has a maximum current rating of 20 amps (at 120 volts). Since heat tape is really just a continuous parallel circuit, longer lengths of it draw greater current. This maximum total current rating therefore places a limit on the usable length of the tape.

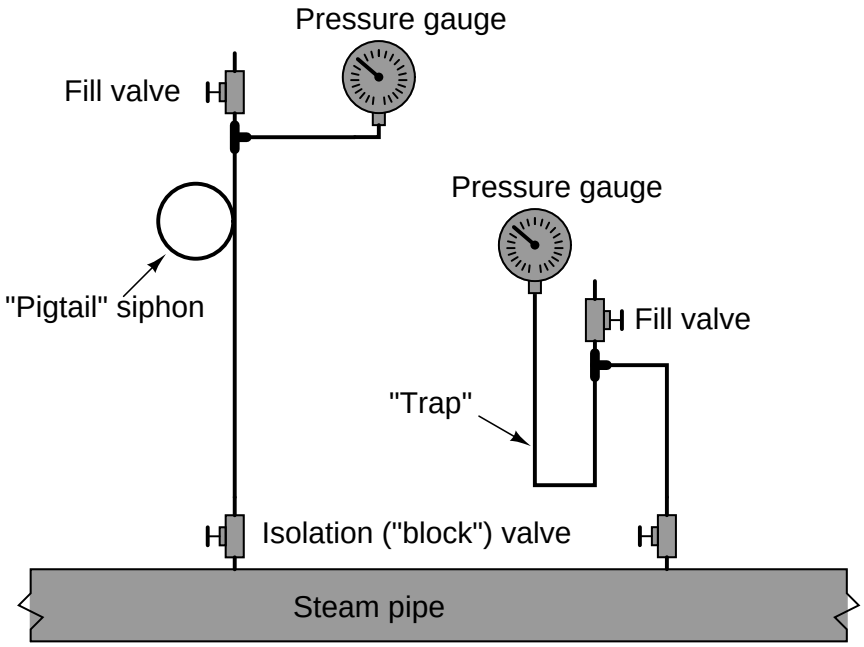
Heat tape may be self-regulating, or controlled with an external thermostat. Self-regulating heat tape exhibits an electrical resistance that varies with temperature, automatically self-regulating its own temperature without the need for external controls.

Both steam and electrical heat tracing are used to protect instruments themselves from cold weather freezing, not just the impulse lines. In these applications it is important to remember that only the liquid-filled portions of the instrument need freeze protection, not the electronics portions!

18.6.8 Water traps and pigtail siphons

Many industrial processes utilize high-pressure steam for direct heating, performing mechanical work, combustion control, and as a chemical reactant. Measuring the pressure of steam is important both for its end-point use and its generation (in a boiler). One problem with doing this is the relatively high temperature of steam at the pressures common in industry, which can cause damage to the sensing element of a pressure instrument if directly connected.

A simple yet effective solution to this problem is to intentionally create a "low" spot in the impulse line where condensed steam (water) will accumulate and act as a liquid barrier to prevent hot steam from reaching the pressure instrument. The principle is much the same as a plumber's trap used underneath sinks, creating a liquid seal to prevent noxious gases from entering a home from the sewer system. A loop of tube or pipe called a *pigtail siphon* achieves the same purpose:

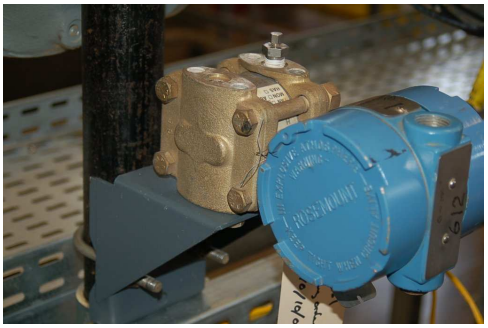


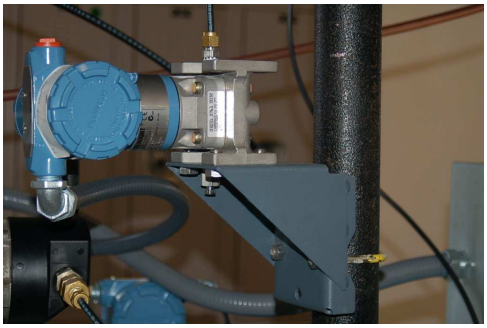
The following photograph shows a pigtail siphon connected to a pressure gauge sensing pressure on a steam line:

18.6.9 Mounting brackets

An accessory specifically designed for a variety of field-mounted instruments including DP transmitters is the *2 inch pipe mounting bracket*. Such a bracket is manufactured from heavy-gauge sheet metal and equipped with a U-bolt designed to clamp around any 2 inch black iron pipe. Holes stamped in the bracket match mounting bolts on the capsule flanges of most common DP transmitters, providing a mechanically stable means of attaching a DP transmitter to a framework in a process area.

The following photographs show several different instruments mounted to pipe sections using these brackets:





18.6.10 Heated enclosures

In installations where the ambient temperature may become very cold, a protective measure against fluid freezing inside a pressure transmitter is to house the transmitter in an insulated, heated enclosure. The next photograph shows just such an enclosure with the cover removed:

Not surprisingly, this installation works well to protect all kinds of temperature-sensitive instruments from extreme cold. Here, we see an explosive gas sensor mounted inside a slightly different style of insulated enclosure, with the lid opened up for inspection:

18.7 Process/instrument suitability

On a fundamental level, pressure is universal. Regardless of the fluid in question; liquid or gas, hot or cold, corrosive or inert, pressure is nothing more than the amount of force exerted by that fluid over a unit area:

$$P = \frac{F}{A}$$

It should come as no surprise, then, that the common mechanical sensing elements for measuring pressure (bellows, diaphragm, bourdon tube, etc.) are equally applicable to all fluid pressure measurement applications, at least in principle. It is normally a matter of proper material selection and element strength (material thickness) to make a pressure instrument suitable for any range of process fluids.

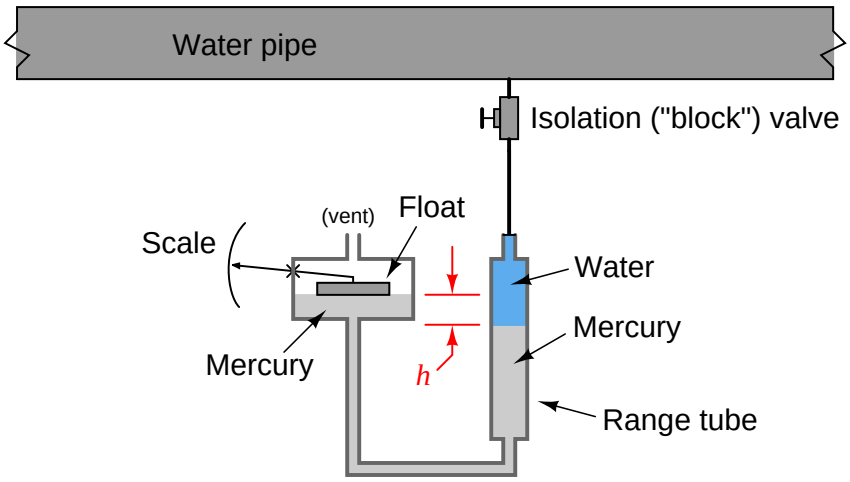
Fill fluids used in pressure instruments – whether it be the dielectric liquid inside a differential capacitance sensor, the fill liquid of a remote or chemical seal system, or liquid used to fill a vertical section of impulse tubing – must be chosen so as to not adversely react with or contaminate the process.

Pure oxygen processes require that no system component have traces of hydrocarbon fluids present. While oxygen itself is not explosive, it greatly accelerates the combustion and explosive potential of any flammable substance. Therefore, a pressure gauge calibrated using oil as the working fluid in a deadweight tester would definitely *not* be suitable for pure oxygen service! The same may be said for a DP transmitter with a hydrocarbon-based fill inside its pressure-sensing capsule[18].

Pharmaceutical, medical, and food manufacturing processes require strict purity and the ability to disinfect all elements in the process system at will. Stagnant lines are not allowed in such processes, as microbe cultures may flourish in such "dead end" piping. Remote seals are very helpful in overcoming this problem, but the fill fluids used in remote systems must be chosen such that a leak in the isolating diaphragm will not contaminate the process.

[18] Although this fluid would not *normally* contact pure oxygen in the process, it could if the isolating diaphragm inside the transmitter were to ever leak.

Manometers, of course, are rather limited in their application, as their operation depends on direct contact between process fluid and manometer liquid. In the early days of industrial instrumentation, liquid mercury was a very common medium for process manometers, and it was not unusual to see a mercury manometer used in direct contact with a process fluid such as oil or water to provide pressure indication:



Thankfully, those days are gone. Mercury (chemical symbol "Hg") is a toxic metal and therefore hazardous to work with. Calibration of these manometers was also challenging due to the column height of the process liquid in the impulse line and the range tube. When the process fluid is a gas, the difference in mercury column height directly translates to sensed pressure by the hydrostatic pressure formula $P = \rho g h$ or $P = \gamma h$. When the process fluid is a liquid, though, the shifting of mercury columns also creates a change in height of the process liquid column, which means the indicated pressure is a function of the height difference (h) *and* the difference in density between the process liquid and mercury. Consequently, the indications provided by mercury manometers in liquid pressure applications were subject to correction according to process liquid density.

References

Beckerath, Alexander von; Eberlein, Anselm; Julien, Hermann; Kersten, Peter; and Kreutzer, Jochem, *WIKA-Handbook, Pressure and Temperature Measurement*, WIKA Alexander Wiegand GmbH & Co., Klingenberg, Germany, 1995.

"Digital Sensor Technology" (PowerPoint slideshow presentation), Yokogawa Corporation of America.

Fribance, Austin E., *Industrial Instrumentation Fundamentals*, McGraw-Hill Book Company, New York, NY, 1962.

Kallen, Howard P., *Handbook of Instrumentation and Controls*, McGraw-Hill Book Company, Inc., New York, NY, 1961.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Measurement and Analysis Volume I*, Fourth Edition, CRC Press, New York, NY, 2003.

Patrick, Dale R. and Patrick, Steven R., *Pneumatic Instrumentation*, Delmar Publishers, Inc., Albany, NY, 1993.

Technical Note: "Rosemount 1199 Fill Fluid Specifications", Rosemount, Emerson Process Management, 2005.

Chapter 19

Continuous level measurement

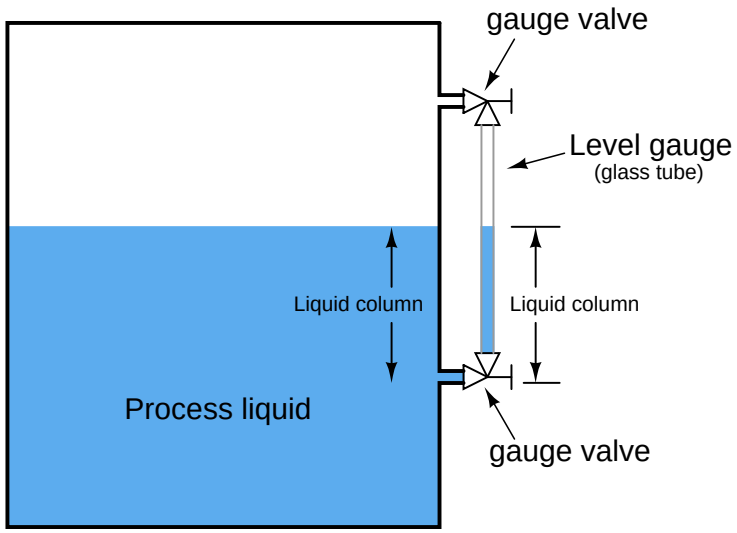
Many industrial processes require the accurate measurement of fluid or solid (powder, granule, etc.) height within a vessel. Some process vessels hold a stratified combination of fluids, naturally separated into different layers by virtue of differing densities, where the height of the *interface* point between liquid layers is of interest.

A wide variety of technologies exist to measure the level of substances in a vessel, each exploiting a different principle of physics. This chapter explores the major level-measurement technologies in current use.

19.1 Level gauges (sightglasses)

The *level gauge*, or *sightglass* is to liquid level measurement as manometers are to pressure measurement: a very simple and effective technology for direct visual indication of process level. In its simplest form, a level gauge is nothing more than a clear tube through which process liquid may be seen. The following photograph shows a simple example of a sightglass:

A functional diagram of a sightglass shows how it visually represents the level of liquid inside a vessel such as a storage tank:

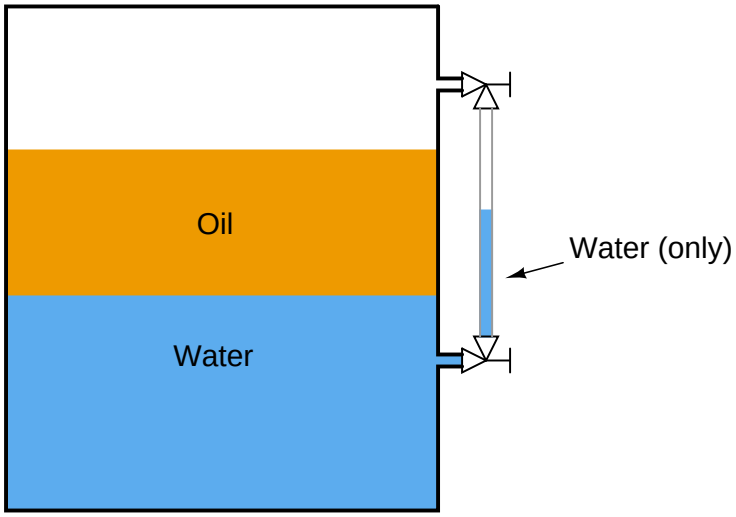


A level gauge is not unlike a U-tube manometer, with equal pressures applied to both liquid columns (one column being the liquid in the gauge sightglass, the other column being the liquid in the vessel).

Level gauge valves exist to allow replacement of the glass tube without emptying or depressurizing the process vessel. These valves are usually equipped with flow-limiting devices in the event of a tube rupture, so too much process fluid does not escape even when the valves are fully open.

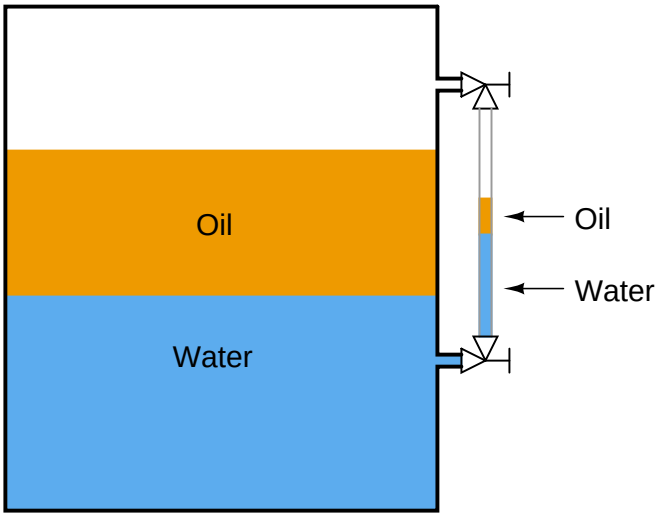
Some level gauges called *reflex gauges* are equipped with special optics to facilitate the viewing of clear liquids, which is problematic for simple glass-tube sightglasses.

As simple and apparently trouble-free as level gauges may seem, there are special circumstances where they will register incorrectly. One such circumstance is in the presence of a lighter liquid layer existing between the connection ports of the gauge. If a lighter (less dense) liquid exists above a heavier (denser) liquid in the process vessel, the level gauge may not show the proper interface, if at all:

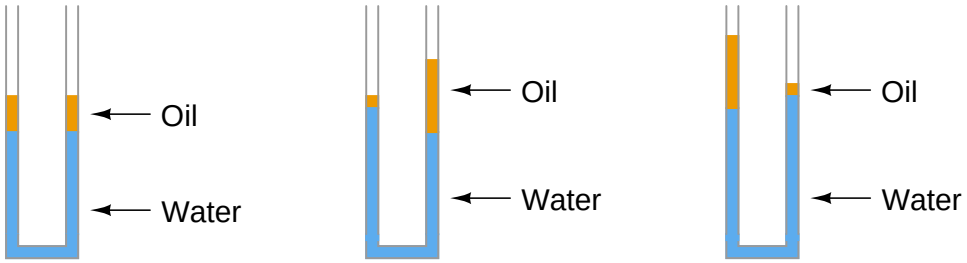


Here we see how a column of water in the sightglass shows less (total) level than the combination of water and oil inside the process vessel. Since the oil lies between the two level gauge ports into the vessel (sometimes called *nozzles*), it cannot enter the sightglass tube, and therefore the level gauge will continue to show just water.

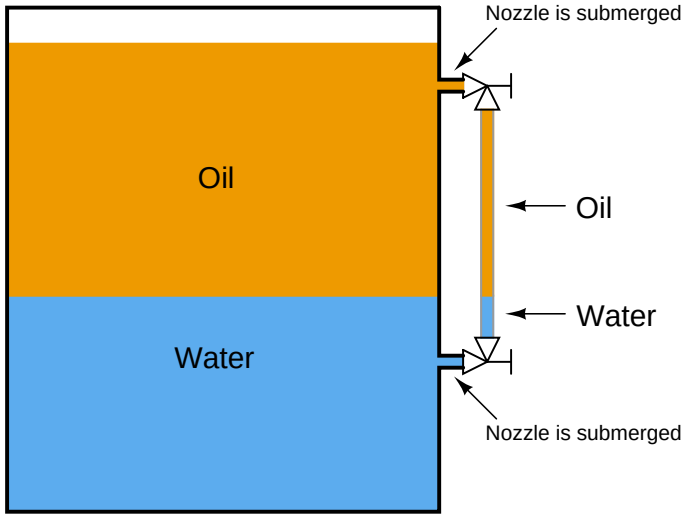
If by chance some oil does find its way into the sightglass tube – either by the interface level dropping below the lower nozzle or the total level rising above the upper nozzle – the oil/water interface shown inside the level gauge may not continue to reflect the true interface inside the vessel once the interface and total levels return to their previous positions:



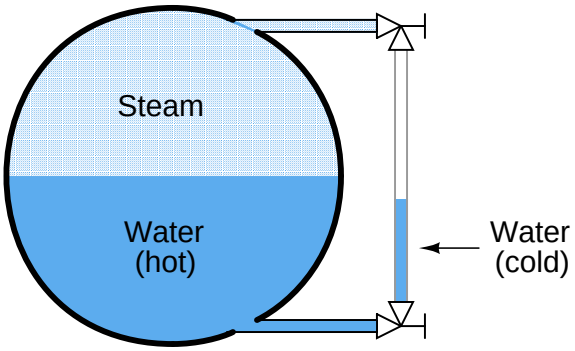
Recall that the level gauge and vessel together form a U-tube manometer. So long as the pressures from each liquid column are the same, the columns balance each other. The problem is, many different liquid-liquid interface columns can have the same hydrostatic pressure without being identical to one another:



The only way to ensure proper two-part liquid interface level indication in a sightglass is to keep both ports (nozzles) submerged:



Another troublesome scenario for level gauges is when the liquid inside the vessel is substantially hotter than the liquid in the gauge, causing the densities to be different. This is commonly seen on boiler level gauges, where the water inside the sightglass cools off substantially from its former temperature inside the boiler drum:

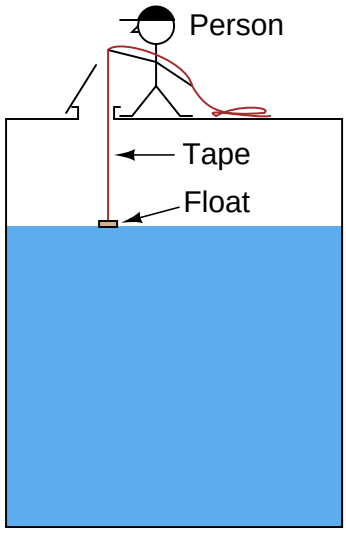


Looking at the sightglass as a U-tube manometer again, we see that unequal-height liquid columns may indeed balance each other's hydrostatic pressures if the two columns are comprised of liquids with different densities. The weight density of water is 62.4 lb/ft^3 at standard temperature, but may be as low as only 36 lb/ft^3 at temperatures common for power generation boilers.

19.2 Float

Perhaps the simplest form of solid or liquid level measurement is with a *float*: a device that rides on the surface of the fluid or solid within the storage vessel. The float itself must be of substantially lesser density than the substance of interest, and it must not corrode or otherwise react with the substance.

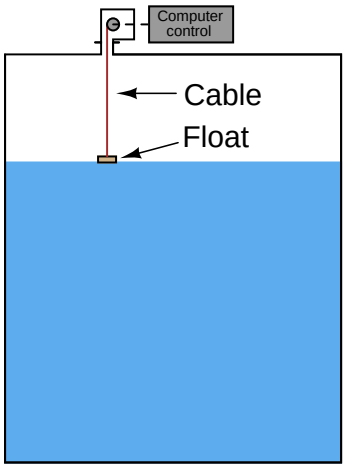
Floats may be used for manual "gauging" of level, as illustrated here:



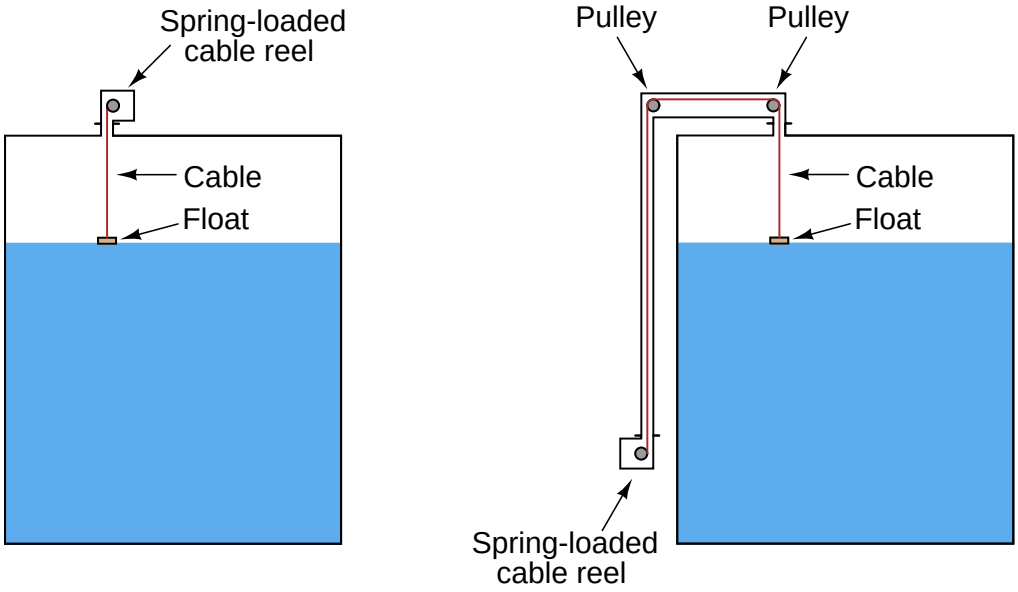
A person lowers a float down into a storage vessel using a flexible measuring tape, until the tape goes slack due to the float coming to rest on the material surface. At that point, the person notes the length indicated on the tape (reading off the lip of the vessel access hole). This distance is called the *ullage*, being the distance from the top of the vessel to the surface of the process material. *Fillage* of the vessel may be determined by subtracting this "ullage" measurement from the known height of the vessel.

Obviously, this method of level measurement is tedious and may pose risk to the person conducting the measurement. If the vessel is pressurized, this method is simply not applicable.

If we automate the person's function using a small winch controlled by a computer – having the computer automatically lower the float down to the material surface and measure the amount of cable played out at each measurement cycle – we may achieve better results without human intervention. Such a level gauge may be enclosed in such a way to allow pressurization of the vessel, too:



A simpler version of this technique uses a spring-reel to constantly tension the cable holding the float, such that the float continuously rides on the surface of the liquid in the vessel[1]:

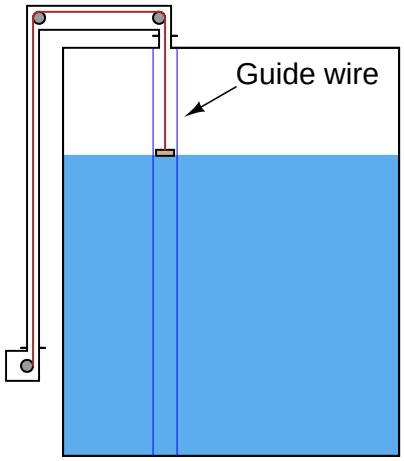


[1] A spring-loaded cable float only works with liquid level measurement, while a retracting float will measure liquids and solids with equal ease. The reason for this limitation is simple: a float that always contacts the material surface is likely to become buried if the material in question is a solid (powder or granules), which must be fed into the vessel from above.

The following photograph shows the "measurement head" of a spring-reel tape-and-float liquid level transmitter, with the vertical pipe housing the tape on its way to the top of the storage tank where it will turn 180 degrees via two pulleys and attach to the float inside the tank:

The spring reel's angular position may be measured by a multi-turn potentiometer or a rotary encoder (located inside the "head" unit), then converted to an electronic signal for transmission to a remote display, control, and/or recording system. Such systems are used extensively for measurement of water and fuel in storage tanks.

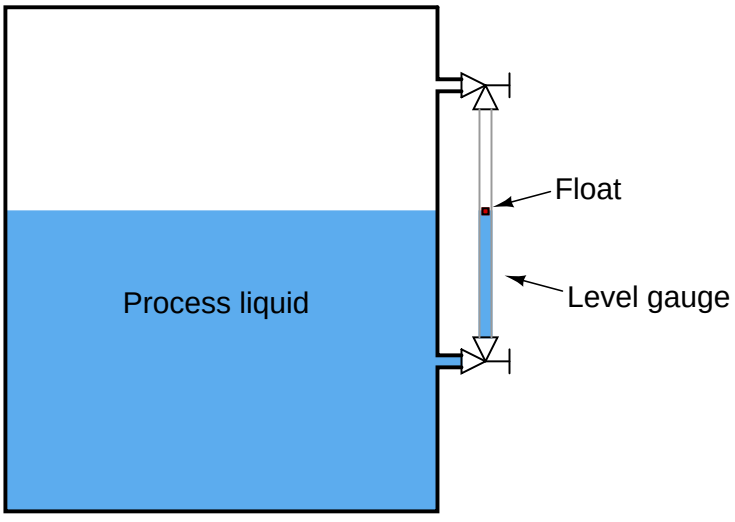
If the liquid inside the vessel is subject to turbulence, *guide wires* may be necessary to keep the float cable in a vertical orientation:



The guide wires are anchored to the floor and roof of the vessel, passing through ring lugs on the float to keep it from straying laterally.

One of the potential disadvantages of tape-and-float level measurement systems is fouling of the tape (and guide wires) if the substance is sticky or unclean.

A variation on the theme of float level measurement is to place a small float inside the tube of a sightglass-style level gauge:



The float's position inside the tube may be readily detected by ultrasonic waves, magnetic sensors or any other applicable means. Locating the float inside a tube eliminates the need for guide wires or a sophisticated tape retraction or tensioning system. If no visual indication is necessary, the

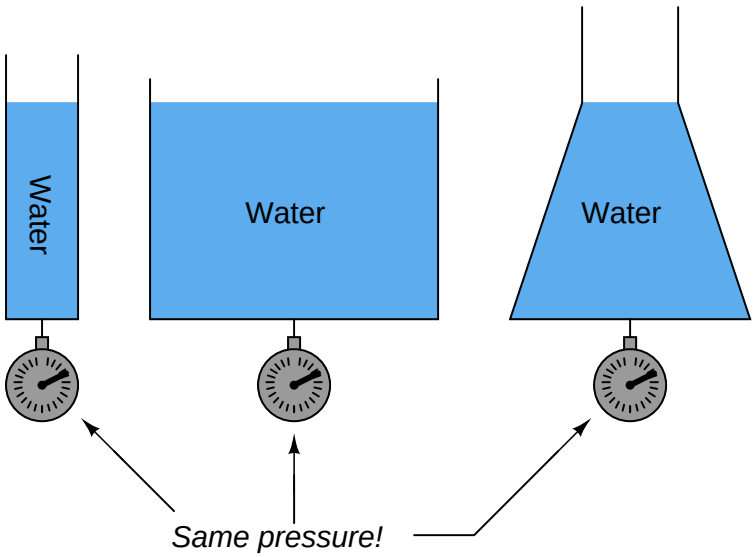
level gauge tube may be constructed out of metal instead of glass, greatly reducing the risk of tube breakage. All the problems inherent to sightglasses, however, still apply to this form of float instrument.

Another variation on the theme of float level measurement is to use a principle called *magnetostriction* to detect the position of the float along a metal guide rod called a *waveguide*. This instrument design is discussed in more detail in a later section of this chapter (see 19.5.4, beginning on page 915).

19.3 Hydrostatic pressure

A vertical column of fluid exerts a pressure due to the column's weight. The relationship between column height and fluid pressure at the bottom of the column is constant for any particular fluid (density) regardless of vessel width or shape.

This principle makes it possible to infer the height of liquid in a vessel by measuring the pressure generated at the bottom:



The mathematical relationship between liquid column height and pressure is as follows:

$$P = \rho g h \qquad\qquad P = \gamma h$$

Where,

P = Hydrostatic pressure
ρ = Mass density of fluid in kilograms per cubic meter (metric) or slugs per cubic foot (British)
g = Acceleration of gravity
γ = Weight density of fluid in newtons per cubic meter (metric) or pounds per cubic foot (British)
h = Height of vertical fluid column above point of pressure measurement

For example, the pressure generated by a column of oil 12 feet high having a weight density (γ) of 40 pounds per cubic foot is:

$$P = \gamma h$$

$$P = \left(\frac{12 \text{ ft}}{1}\right)\left(\frac{40 \text{ lb}}{\text{ft}^3}\right)$$

$$P = \frac{480 \text{ lb}}{\text{ft}^2}$$

Note the cancellation of units, resulting in a pressure value of 480 pounds per square foot (PSF). To convert into the more common pressure unit of pounds per square inch, we may multiply by the proportion of square feet to square inches, eliminating the unit of square feet by cancellation and leaving square inches in the denominator:

$$P = \left(\frac{480 \text{ lb}}{\text{ft}^2}\right)\left(\frac{1^2 \text{ ft}^2}{12^2 \text{ in}^2}\right)$$

$$P = \left(\frac{480 \text{ lb}}{\text{ft}^2}\right)\left(\frac{1 \text{ ft}^2}{144 \text{ in}^2}\right)$$

$$P = \frac{3.\overline{33} \text{ lb}}{\text{in}^2} = 3.\overline{33} \text{ PSI}$$

Thus, a pressure gauge attached to the bottom of the vessel holding a 12 foot column of this oil would register $3.\overline{33}$ PSI. It is possible to customize the scale on the gauge to read directly in feet of oil (height) instead of PSI, for convenience of the operator who must periodically read the gauge. Since the mathematical relationship between oil height and pressure is both linear and direct, the gauge's indication will always be proportional to height.

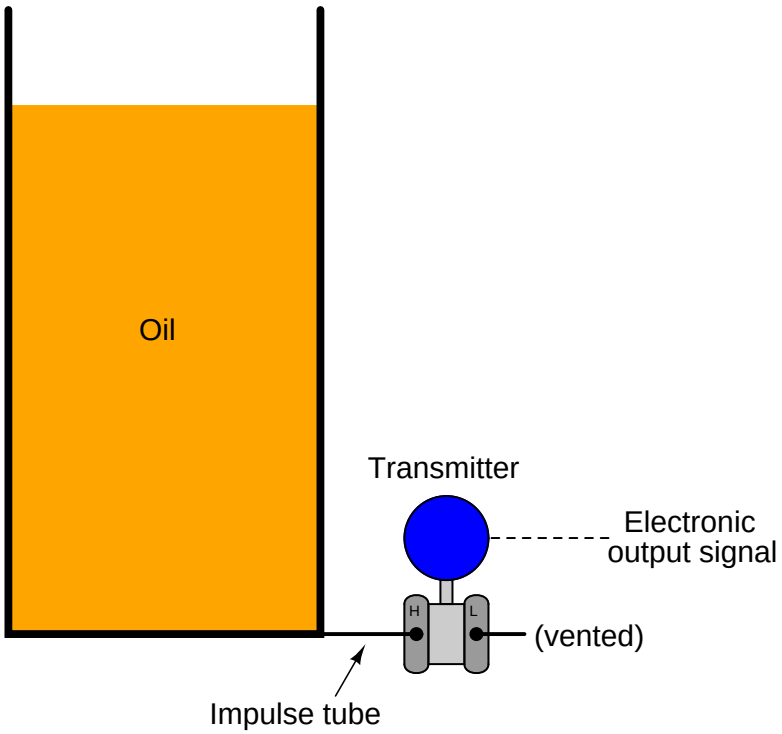
Any type of pressure-sensing instrument may be used as a liquid level transmitter by means of this principle. In the following photograph, you see a Rosemount model 1151 pressure transmitter being used to measure the height of colored water inside a clear plastic tube:

The critically important factor in liquid level measurement using hydrostatic pressure is liquid density. One must accurately know the liquid's density in order to have any hope of measuring that liquid's level using hydrostatic pressure, since density is an integral part of the height/pressure relationship ($P = \rho gh$ and $P = \gamma h$). Having an accurate assessment of liquid density also implies that density must remain relatively constant despite other changes in the process. If the liquid density is subject to random variation, the accuracy of any hydrostatic pressure-based level instrument will correspondingly vary.

It should be noted, though, that changes in liquid density will have absolutely no effect on hydrostatic measurement of liquid *mass*, so long as the vessel has a constant cross-sectional area throughout its entire height. A simple thought experiment proves this: imagine a vessel partially full of liquid, with a pressure transmitter attached to the bottom to measure hydrostatic pressure.

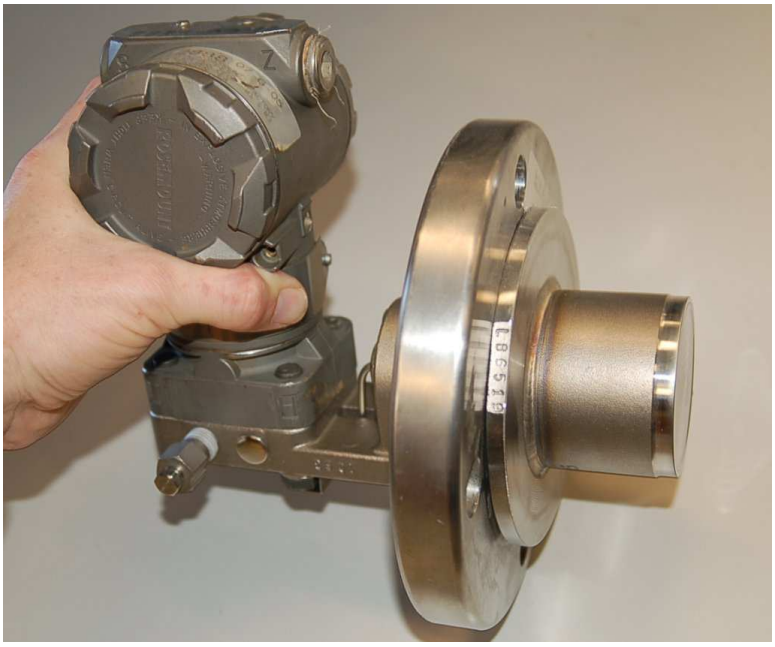
Now imagine the temperature of that liquid increasing, such that its volume expands and has a lower density than before. Assuming no addition or loss of liquid to or from the vessel, any increase in liquid level will be strictly due to volume expansion (density decrease). Liquid level inside this vessel will rise, but the transmitter will sense the exact same hydrostatic pressure as before, since the rise in level is precisely countered by the decrease in density (if h increases by the same factor that γ decreases, then $P = \gamma h$ must remain the same!). In other words, hydrostatic pressure is seen to be a direct indication of the liquid *mass* contained within the vessel, regardless of changes in liquid density.

Differential pressure transmitters are the most common pressure-sensing device used in this capacity to infer liquid level within a vessel. In the hypothetical case of the oil vessel just considered, the transmitter would connect to the vessel in this manner (with the high side toward the process and the low side vented to atmosphere):



Connected as such, the differential pressure transmitter functions as a gauge pressure transmitter, responding to hydrostatic pressure exceeding ambient (atmospheric) pressure. As liquid level increases, the hydrostatic pressure applied to the "high" side of the differential pressure transmitter also increases, driving the transmitter's output signal higher.

Some pressure-sensing instruments are built specifically for hydrostatic measurement of liquid level in vessels, doing away with impulse tubing altogether in favor of a special kind of sealing diaphragm extending slightly into the vessel through a flanged pipe entry (commonly called a *nozzle*). A Rosemount hydrostatic level transmitter with an extended diaphragm is shown here:



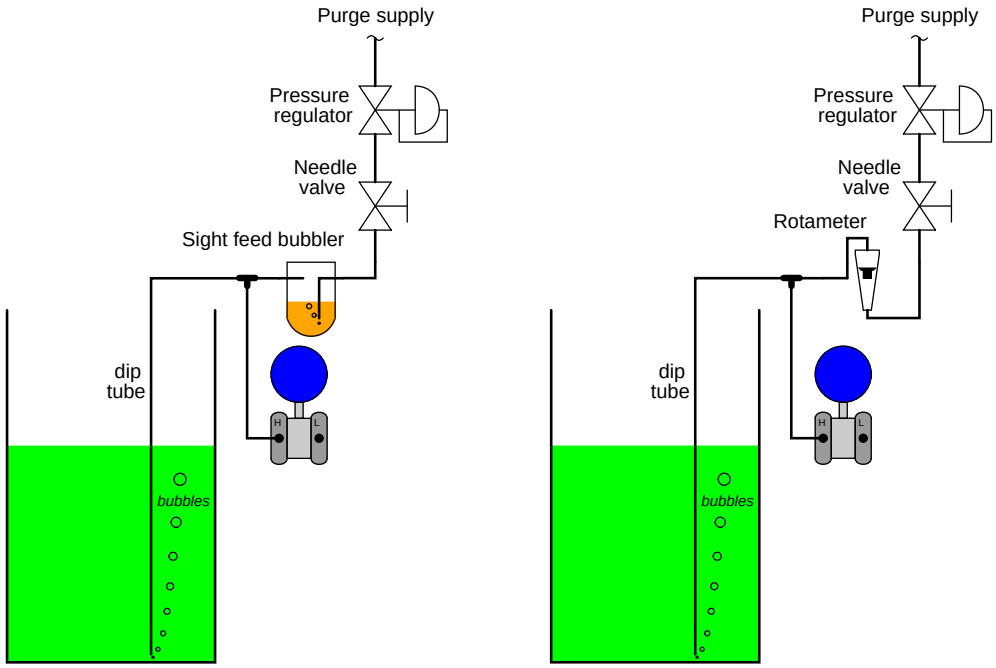
The calibration table for a transmitter close-coupled to the bottom of an oil storage tank would be as follows, assuming a zero to twelve foot measurement range for oil height, an oil density of 40 pounds per cubic foot, and a 4-20 mA transmitter output signal range:

Oil level	Percent of range	Hydrostatic pressure	Transmitter output
0 ft	0 %	0 PSI	4 mA
3 ft	25 %	0.833 PSI	8 mA
6 ft	50 %	1.67 PSI	12 mA
9 ft	75 %	2.50 PSI	16 mA
12 ft	100 %	3.33 PSI	20 mA

19.3.1 Bubbler systems

An interesting variation on this theme of direct hydrostatic pressure measurement is the use of a purge gas to measure hydrostatic pressure in a liquid-containing vessel. This eliminates the need for direct contact of the process liquid against the pressure-sensing element, which can be advantageous if the process liquid is corrosive.

Such systems are often called *bubble tube* or *dip tube* systems, the former name being appropriately descriptive for the way purge gas bubbles out the end of the tube as it is submerged in process liquid. A key detail of a bubble tube system is to provide a means of limiting gas flow through the tube, so the purge gas backpressure properly reflects hydrostatic pressure at the end of the tube with no additional pressure due to frictional losses of purge flow through the length of the tube. Most bubble tube systems, therefore, are provided with some means of monitoring purge gas flow, typically with a *rotameter* or with a *sightfeed bubbler*:



If the purge gas flow is not too great, gas pressure measured anywhere in the tube system downstream of the needle valve will be equal to the hydrostatic pressure of the process liquid at the bottom of the tube where the gas escapes. In other words, the purge gas acts to transmit the liquid's hydrostatic pressure to some remote point where a pressure-sensing instrument is located. A general rule-of-thumb is to limit purge gas flow to the point where you can easily count individual bubbles exiting the bubble tube (or inside the sightfeed bubbler if one is provided on the system).

As with all purged systems, certain criteria must be met for successful operation. Listed here are a few pertinent questions to consider for a bubble tube system:

- How reliable is the supply of purge fluid? If this stops for any reason, the level measurement may be in error!

- Is the purge fluid supply pressure guaranteed to exceed the hydrostatic pressure at all times, to ensure continuous purging (bubbling)?
- What options exist for purge gases that will not adversely react with the process?
- What options exist for purge gases that will not contaminate the process?
- How expensive will it be to maintain this constant flow of purge gas into the process?

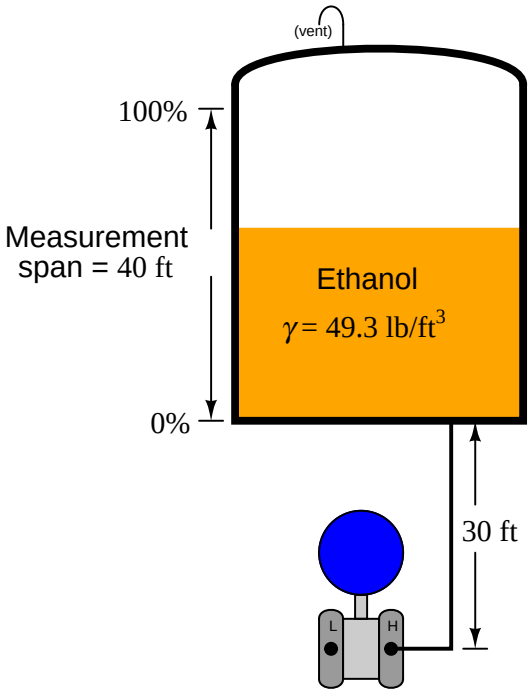
One measurement artifact of a bubble tube system is a slight variation in pressure each time a new bubble breaks away from the end of the tube. The amount of pressure variation is approximately equal to the hydrostatic pressure of process fluid at a height equal to the diameter of the bubble, which in turn will be approximately equal to the diameter of the bubble tube. For example, a 1/4 inch diameter dip tube will experience pressure oscillations with a peak-to-peak amplitude of approximately 1/4 inch elevation of process liquid. The frequency of this pressure oscillation, of course, will be equal to the rate at which individual bubbles escape out the end of the dip tube.

Usually, this is a small variation when considered in the context of the measured liquid height in the vessel. A pressure oscillation of approximately 1/4 inch compared to a measurement range of 0 to 10 feet, for example, is only about 0.2% of span. Modern pressure transmitters have the ability to "filter" or "damp" pressure variations over time, which is a useful feature for minimizing the effect such a pressure variation will have on system performance.

19.3.2 Transmitter suppression and elevation

A very common scenario for liquid level measurement is where the pressure-sensing instrument is not located at the same level as the 0% measurement point. The following photograph shows an example of this, where a Rosemount model 3051 differential pressure transmitter is being used to sense hydrostatic pressure of colored water inside a (clear) vertical plastic tube:

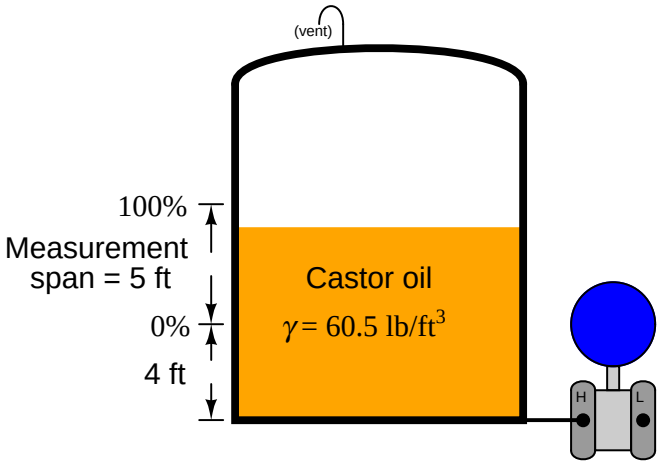
Consider the example of a pressure sensor measuring the level of liquid ethanol in a storage tank. The measurement range for liquid height in this ethanol storage tank is 0 to 40 feet, but the transmitter is located 30 feet below the tank:



This means the transmitter's impulse line contains a 30-foot elevation head of ethanol, so the transmitter "sees" 30 feet of ethanol when the tank is empty and 70 feet of ethanol when the tank is full. A 3-point calibration table for this instrument would look like this, assuming a 4 to 20 mA DC output signal range:

Ethanol level in tank	Percent of range	Pressure (inches of water)	Pressure (PSI)	Output (mA)
0 ft	0 %	284 "W.C.	10.3 PSI	4 mA
20 ft	50 %	474 "W.C.	17.1 PSI	12 mA
40 ft	100 %	663 "W.C.	24.0 PSI	20 mA

Another common scenario is where the transmitter is mounted at or near the vessel's bottom, but the desired level measurement range does not extend to the vessel bottom:



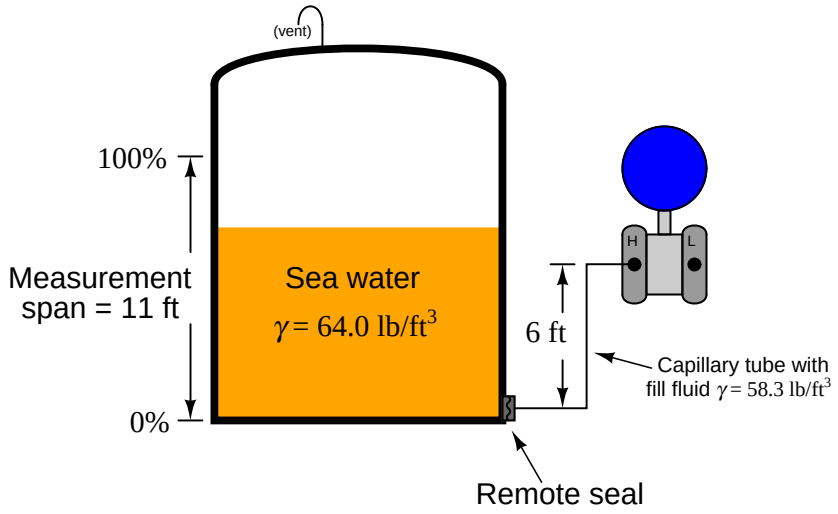
In this example, the transmitter is mounted exactly at the same level as the vessel bottom, but the level measurement range goes from 4 feet to 9 feet (a 5 foot span). At the level of castor oil deemed 0%, the transmitter "sees" a hydrostatic pressure of 1.68 PSI (46.5 inches of water column) and at the 100% castor oil level the transmitter "sees" a pressure of 3.78 PSI (105 inches water column). Thus, these two pressure values would define the transmitter's lower and upper range values (LRV and URV), respectively.

The term for describing either of the previous scenarios, where the lower range value (LRV) of the transmitter's calibration is a positive number, is called *zero suppression*[2]. If the zero offset is reversed (e.g. the transmitter mounted at a location *higher* than the 0% process level), it is referred to as *zero elevation*[3].

[2]Or alternatively, zero *depression*.

[3]There is some disagreement among instrumentation professionals as to the definitions of these two terms. According to Béla G. Lipták's *Instrument Engineers' Handbook, Process Measurement and Analysis* (Fourth Edition, page 67), "suppressed zero range" refers to the transmitter being located below the 0% level (the LRV being a positive pressure value), while "suppression," "suppressed range," and "suppressed span" mean exactly the opposite (LRV is a negative value). The Yokogawa Corporation defines "suppression" as a condition where the LRV is a positive pressure ("Autolevel" Application Note), as does the Michael MacBeth in his CANDU Instrumentation & Control course (lesson 1, module 4, page 12), Foxboro's technical notes on bubble tube installations (pages 4 through 7), and Rosemount's product manual for their 1151 Alphaline pressure transmitter (page 3-7). Interestingly, the Rosemount document defines "zero range suppression" as synonymous with "suppression," which disagrees with Lipták's distinction. My advice: draw a picture if you want the other person to clearly understand what you mean!

If the transmitter is elevated above the process connection point, it will most likely "see" a negative pressure (vacuum) with an empty vessel owing to the pull of liquid in the line leading down from the instrument to the vessel. It is vitally important in elevated transmitter installations to use a *remote seal* rather than an open impulse line, so liquid cannot dribble out of this line and into the vessel[4]:



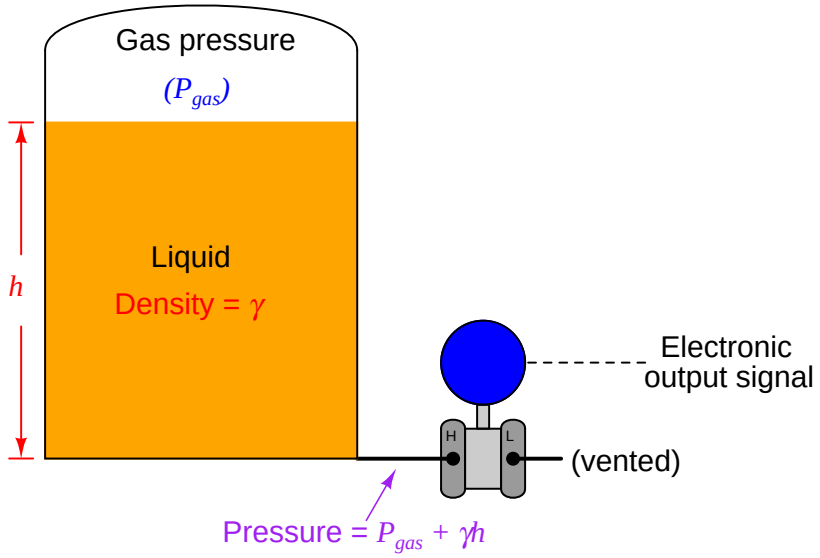
In this example, we see a remote seal system with a fill fluid having a density of 58.3 lb/ft^3, and a process level measurement range of 0 to 11 feet of sea water (density = 64 lb/ft^3). The transmitter elevation is 6 feet, which means it will "see" a vacuum of -2.43 PSI (-67.2 inches of water column) when the vessel is completely empty. This, of course, will be the transmitter's calibrated lower range value (LRV). The upper range value (URV) will be the pressure "seen" with 11 feet of sea water in the vessel. This much sea water will contribute an additional 4.89 PSI of hydrostatic pressure at the level of the remote seal diaphragm, causing the transmitter to experience a pressure of +2.46 PSI[5].

[4] As you are about to see, the calibration of an elevated transmitter depends on us knowing how much hydrostatic pressure (or vacuum, in this case) is generated within the tube connecting the transmitter to the process vessel. If liquid were to ever escape from this tube, the hydrostatic pressure would be unpredictable, and so would be the accuracy of our transmitter as a level-measuring instrument. A remote seal diaphragm guarantees no fill fluid will be lost if and when the process vessel goes empty.

[5] The sea water's positive pressure at the remote seal diaphragm adds to the negative pressure already generated by the downward length of the capillary tube's fill fluid (-2.43 PSI), which explains why the transmitter only "sees" 2.46 PSI of pressure at the 100% full mark.

19.3.3 Compensated leg systems

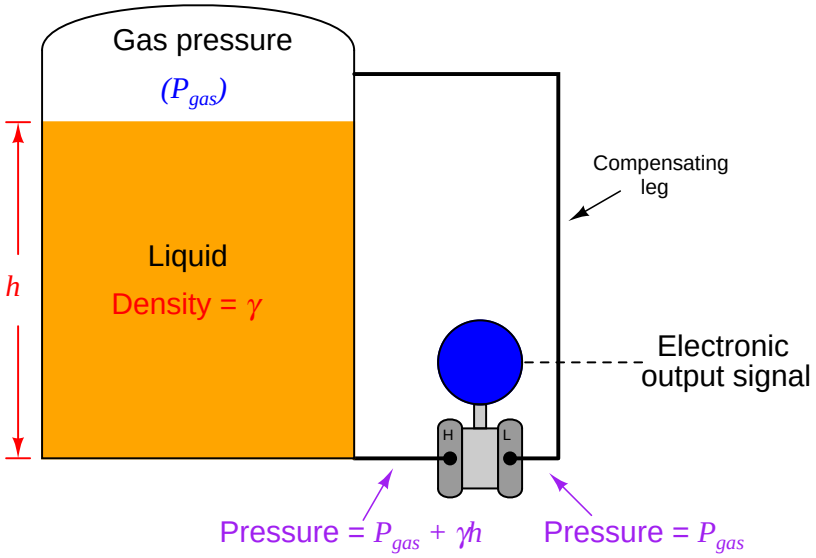
The simple and direct relationship between liquid height in a vessel and pressure at the bottom of that vessel is ruined if another source of pressure exists inside the vessel other than hydrostatic (elevation head). This is virtually guaranteed to be the case if the vessel in question is unvented. Any gas or vapor pressure accumulation in an enclosed vessel will add to the hydrostatic pressure at the bottom, causing any pressure-sensing instrument to falsely register a high level:



A pressure transmitter has no way of "knowing" how much of the sensed pressure is due to liquid elevation and how much of it is due to pressure existing in the vapor space above the liquid. Unless a way can be found to compensate for any non-hydrostatic pressure in the vessel, this extra pressure will be interpreted by the transmitter as additional liquid level.

Moreover, this error will change as gas pressure inside the vessel changes, so it cannot simply be "calibrated away" by a static zero shift within the instrument. The only way to hydrostatically measure liquid level inside an enclosed (non-vented) vessel is to continuously compensate for gas pressure.

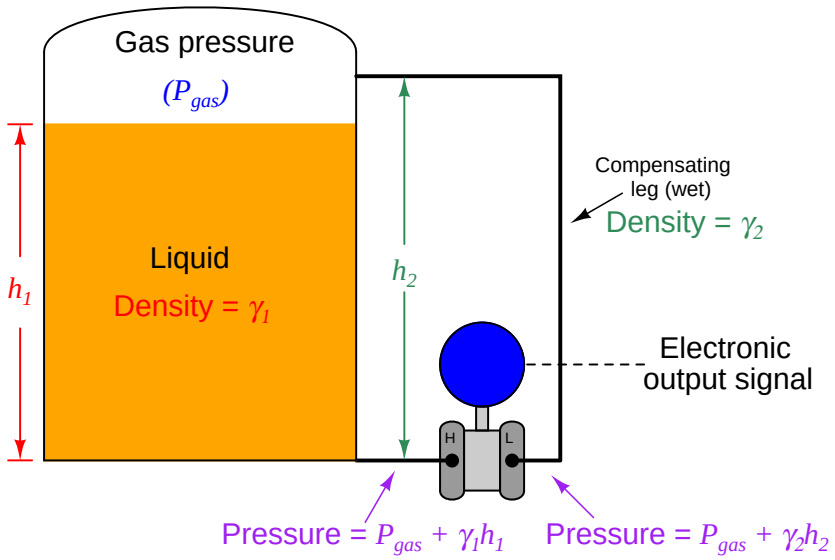
Fortunately, the capabilities of a *differential* pressure transmitter make this a simple task. All we need to do is connect a second impulse line (called a *compensating leg*), from the "Low" port of the transmitter to the top of the vessel, so the "Low" side of the transmitter experiences nothing but the gas pressure enclosed by the vessel, while the "High" side experiences the *sum* of gas and hydrostatic pressures. Since a differential pressure transmitter responds only to *differences* in pressure between "High" and "Low" sides, it will naturally subtract the gas pressure (P_{gas}) to yield a measurement based solely on hydrostatic pressure (γh):



$$(P_{gas} + \gamma h) - P_{gas} = \gamma h$$

The amount of gas pressure inside the vessel now becomes completely irrelevant to the transmitter's indication, because its effect is canceled at the differential pressure instrument's sensing element. If gas pressure inside the vessel were to increase while liquid level remained constant, the pressure sensed at *both* ports of the differential pressure transmitter would increase by the exact same amount, with the pressure *difference* between the "high" and "low" ports remaining absolutely constant with the constant liquid level. This means the instrument's output signal is a representation of hydrostatic pressure only, which represents liquid height (assuming a known liquid density γ).

Unfortunately, it is common for enclosed vessels to hold condensible vapors, which may over time fill a compensating leg full of liquid. If the tube connecting the "Low" side of a differential pressure transmitter fills completely with a liquid, this will add a hydrostatic pressure to that side of the transmitter, causing another calibration shift. This *wet leg* condition makes level measurement more complicated than a *dry leg* condition where the only pressure sensed by the transmitter's "Low" side is gas pressure (P_{gas}):

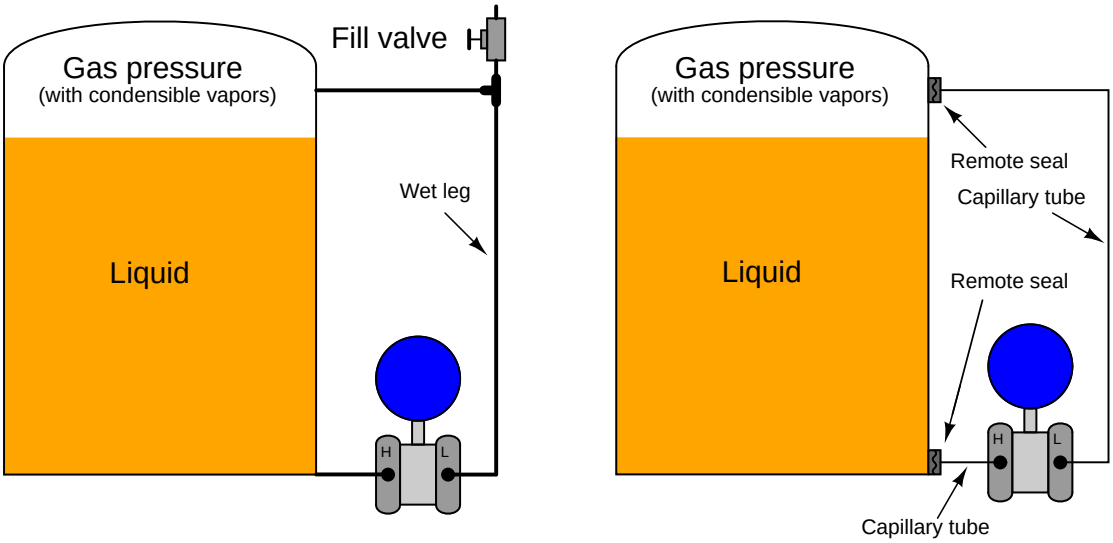


$$(P_{gas} + \gamma_1 h_1) - (P_{gas} + \gamma_2 h_2) = \gamma_1 h_1 - \gamma_2 h_2$$

Gas pressure still cancels due to the differential nature of the pressure transmitter, but now the transmitter's output indicates a difference of hydrostatic pressures between the vessel and the wet leg, rather than just the hydrostatic pressure of the vessel's liquid level. Fortunately, the hydrostatic pressure generated by the wet leg will be constant, so long as the density of the condensed vapors filling that leg (γ_2) is constant. If the wet leg's hydrostatic pressure is constant, we can compensate for it by calibrating the transmitter with an intentional zero shift, so it indicates as though it were measuring hydrostatic pressure on a vented vessel.

$$\text{Differential pressure } = \gamma_1 h_1 - \text{Constant}$$

We may ensure a constant density of wet leg liquid by intentionally filling that leg with a liquid known to be denser than the densest condensed vapor inside the vessel. We could also use a differential pressure transmitter with remote seals and capillary tubes filled with liquid of known density:



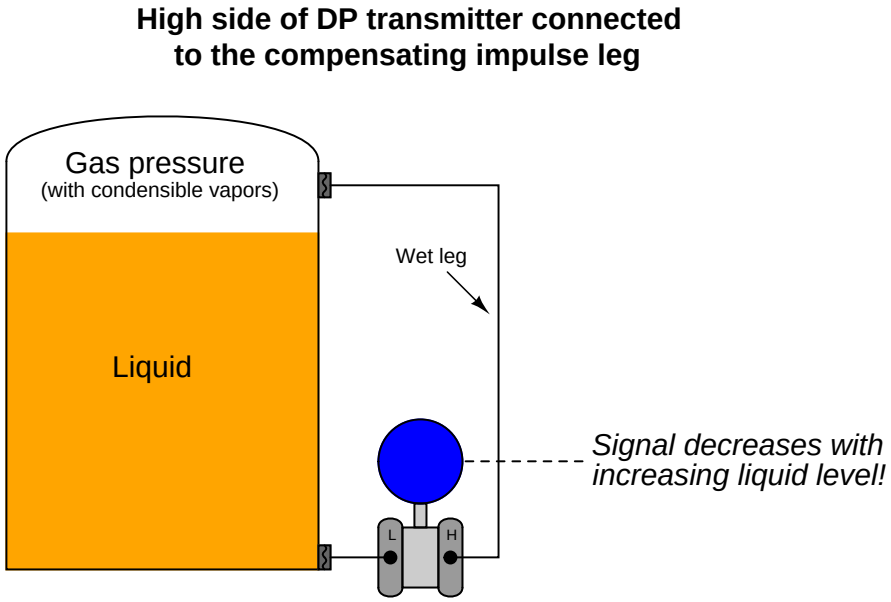
The following example shows the calibration table for a compensated-leg (wet) hydrostatic level measurement system, for a gasoline storage vessel and water as the wet leg fill fluid. Here, I am assuming a density of 41.0 lb/ft^3 for gasoline and 62.4 lb/ft^3 for water, with a 0 to 10 foot measurement range and an 11 foot wet leg height:

Gasoline level	Percent of range	Pressure at transmitter	Transmitter output
0 ft	0 %	-4.77 PSI	4 mA
2.5 ft	25 %	-4.05 PSI	8 mA
5 ft	50 %	-3.34 PSI	12 mA
7.5 ft	75 %	-2.63 PSI	16 mA
10 ft	100 %	-1.92 PSI	20 mA

Note that due to the superior density and height of the wet (water) leg, the transmitter *always* sees a negative pressure (pressure on the "Low" side exceeds pressure on the "High" side). With some older differential pressure transmitter designs, this was a problem. Consequently, it is common to see "wet leg" hydrostatic transmitters installed with the "Low" port connected to the bottom of the vessel and the "High" port connected to the compensating leg. In fact, it is *still* common to see modern differential pressure transmitters installed in this manner[6], although modern transmitters may be calibrated for negative pressures just as easily as for positive pressures. It is vitally important

[6]Sometimes this is done out of habit, other times because instrument technicians do not know the capabilities of new technology.

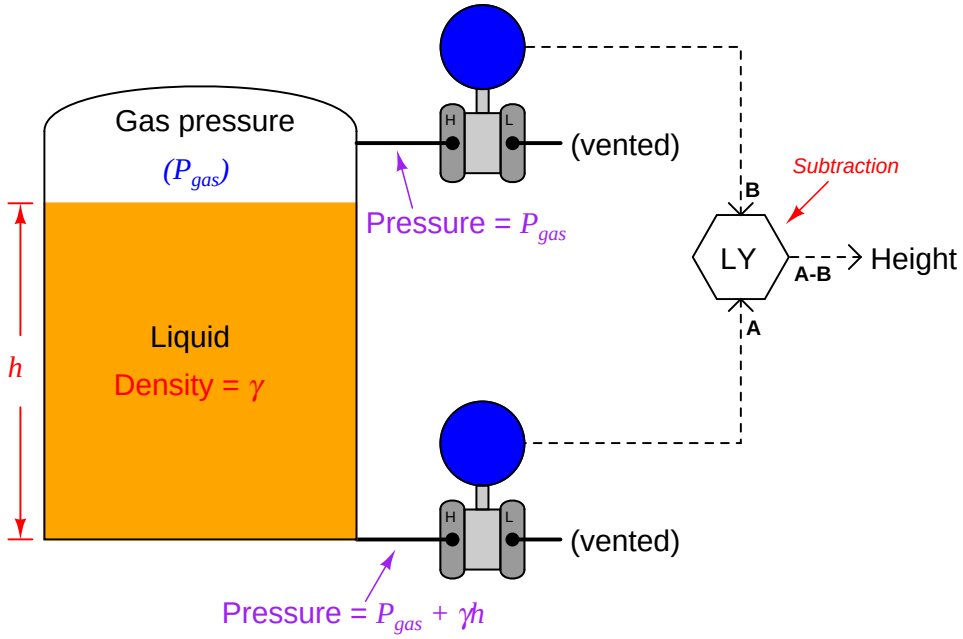
to recognize that any differential pressure transmitter connected as such (for any reason) will respond in reverse fashion to increases in liquid level. That is to say, as the liquid level in the vessel rises, the transmitter's output signal will *decrease* instead of increase:



Either way of connecting the transmitter to the vessel will suffice for measuring liquid level, so long as the instrumentation receiving the transmitter's signal is properly configured to interpret the signal. The choice of which way to connect the transmitter to the vessel should be driven by fail-safe system design, which means to design the measurement system such that the most probably system failures – including broken signal wires – result in the control system "seeing" the most dangerous process condition and therefore taking the safest action.

19.3.4 Tank expert systems

An alternative to using a compensating leg to subtract gas pressure inside an enclosed vessel is to simply use a second pressure transmitter and electronically subtract the two pressures in a computing device:

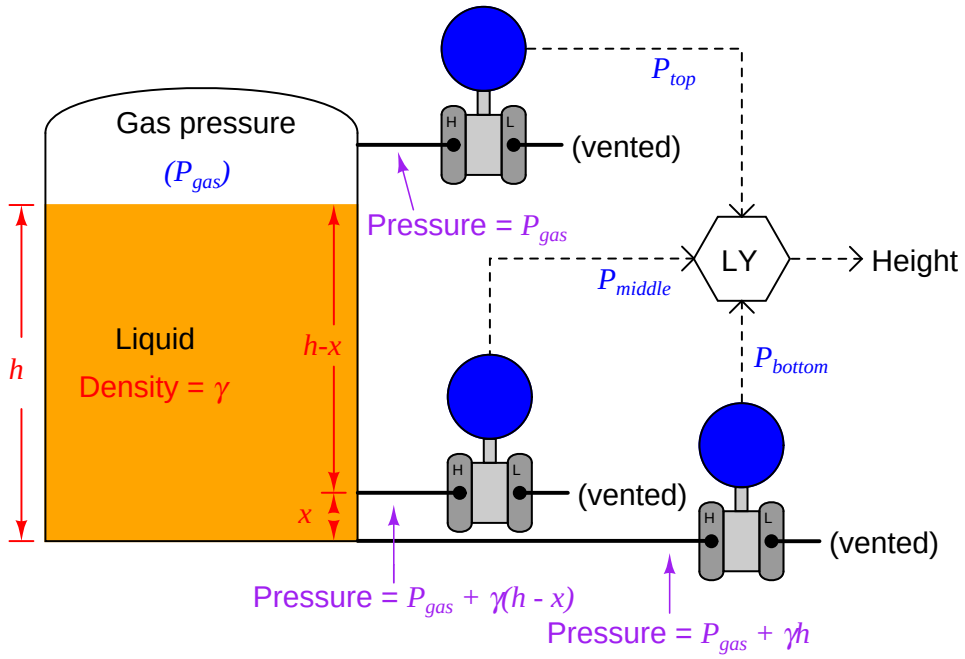


This approach enjoys the distinct advantage of avoiding a potentially wet compensating leg, but suffers the disadvantages of extra cost and greater error due to the potential calibration drift of *two* transmitters rather than just one. Such a system is also impractical in applications where the gas pressure is substantial compared to the hydrostatic (elevation head) pressure[7].

[7] This is due to limited transmitter resolution. Imagine an application where the elevation head was 10 PSI (maximum) yet the vapor space pressure was 200 PSI. The majority of each transmitter's working range would be "consumed" measuring gas pressure, with hydrostatic head being a mere 5% of the measurement range. This would make precise measurement of liquid level very difficult, akin to trying to measure the sound intensity of a whisper in a noisy room.

If we add a third pressure transmitter to this system, located a known distance (x) above the bottom transmitter, we have all the pieces necessary for what is called a *tank expert system*. These systems are used on large storage tanks operating at or near atmospheric pressure, and have the ability to measure infer liquid height, liquid density, total liquid volume, and total liquid mass stored in the tank:

A "tank expert" system



The pressure difference between the bottom and middle transmitters will change only if the liquid density changes[8], since the two transmitters are separated by a known and fixed height difference.

[8] Assuming the liquid level is equal to or greater than x. Otherwise, the pressure difference between P_{bottom} and P_{middle} will depend on liquid density *and* liquid height. However, this condition is easy to check: the level computer simply checks to see if P_{middle} and P_{top} are unequal. If so, then the computer knows the liquid level exceeds x and it is safe to calculate density. If not, and P_{middle} registers the same as P_{top}, the computer knows those two transmitters are both registering gas pressure only, and it knows to stop calculating density.

Algebraic manipulation shows us how the measured pressures may be used by the level computer (LY) to continuously calculate liquid density (γ):

$$P_{bottom} - P_{middle} = (P_{gas} + \gamma h) - [P_{gas} + \gamma(h - x)]$$

$$P_{bottom} - P_{middle} = P_{gas} + \gamma h - P_{gas} - \gamma(h - x)$$

$$P_{bottom} - P_{middle} = P_{gas} + \gamma h - P_{gas} - \gamma h + \gamma x$$

$$P_{bottom} - P_{middle} = \gamma x$$

$$\frac{P_{bottom} - P_{middle}}{x} = \gamma$$

Once the computer knows the value of γ, it may calculate the height of liquid in the tank with great accuracy based on the pressure measurements taken by the bottom and top transmitters:

$$P_{bottom} - P_{top} = (P_{gas} + \gamma h) - P_{gas}$$

$$P_{bottom} - P_{top} = \gamma h$$

$$\frac{P_{bottom} - P_{top}}{\gamma} = h$$

With all the computing power available in the LY, it is possible to characterize the tank such that this height measurement converts to a precise volume measurement[9] (V), which may then be converted into a total mass (m) measurement based on the mass density of the liquid (ρ) and the acceleration of gravity (g). First, the computer calculates mass density based on the proportionality between mass and weight (shown here starting with the equivalence between the two forms of the hydrostatic pressure formula):

$$\rho g h = \gamma h$$

$$\rho g = \gamma$$

$$\rho = \frac{\gamma}{g}$$

[9] The details of this math depend entirely on the shape of the tank. For vertical cylinders – the most common shape for vented storage tanks – volume and height are related by the simple formula $V = \pi r^2 h$ where r is the radius of the tank's circular base. Other tank shapes and orientations may require much more sophisticated formulae to calculate stored volume from height. See section 24.2, beginning on page 1251, for more details on this subject.

Armed with the mass density of the liquid inside the tank, the computer may now calculate total liquid mass stored inside the tank:

$$m = \rho V$$

Dimensional analysis shows how units of mass density and volume cancel to yield only units of mass in this last equation:

$$[\text{kg}] = \left[\frac{\text{kg}}{\text{m}^3}\right] \left[\text{m}^3\right]$$

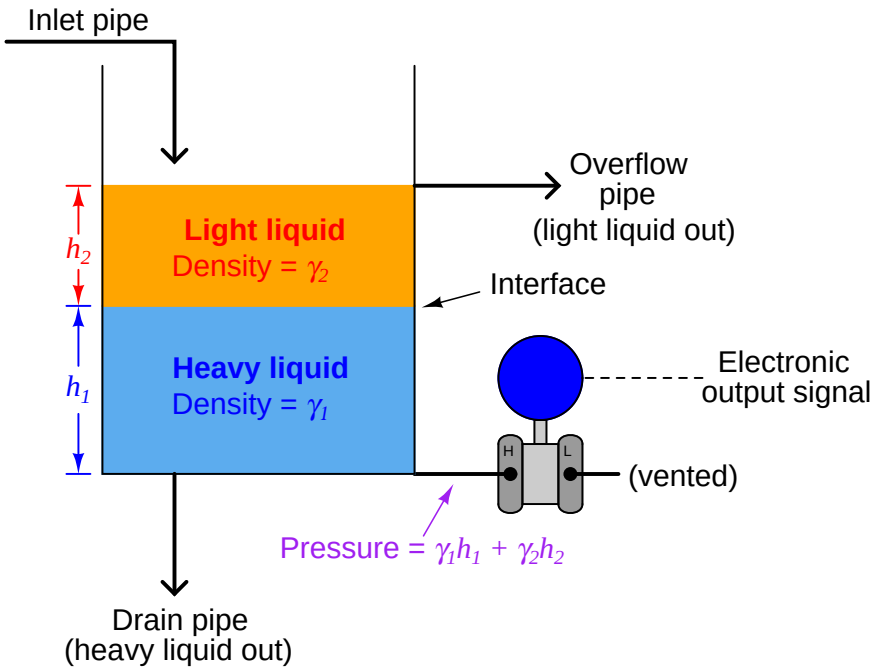
Here we see a vivid example of how several measurements may be inferred from just a few actual process (in this case, pressure) measurements. Three pressure measurements on this tank allow us to compute four inferred variables: liquid density, liquid height, liquid volume, and liquid mass.

The accurate measurement of liquids in storage tanks is not just useful for process operations, but also for conducting business affairs. Whether the liquid represents raw material purchased from a supplier, or a processed product ready to be pumped out to a customer, both parties have a vested interest in knowing the exact quantity of liquid bought or sold. Measurement applications such as this are known as *custody transfer*, because they represent the transfer of custody (ownership) of a substance exchanged in a business agreement. In some instances, both buyer and seller operate and maintain their own custody transfer instrumentation, while in other instances there is but one instrument, the calibration of which validated by a neutral party.

19.3.5 Hydrostatic interface level measurement

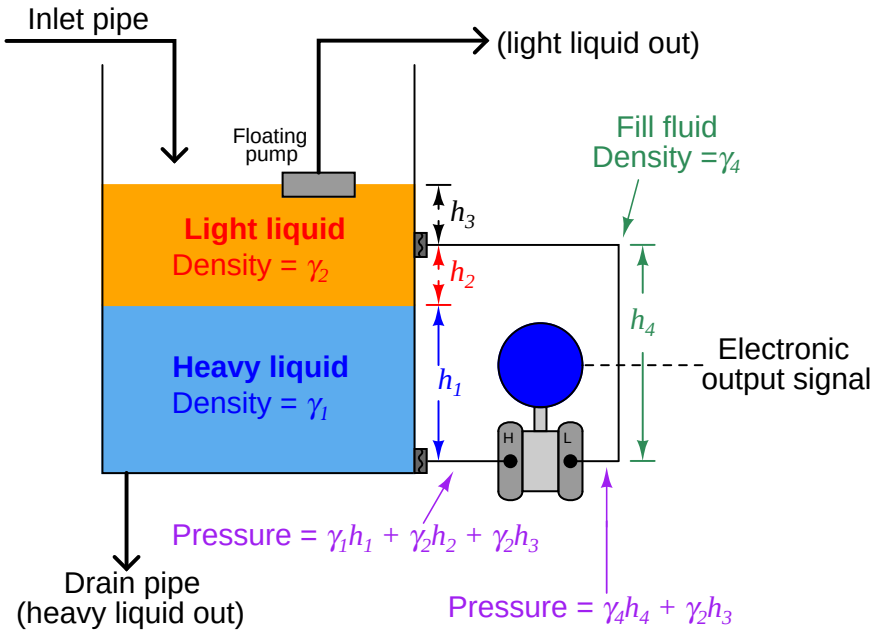
Hydrostatic pressure sensors may be used to detect the level of a liquid-liquid interface, if and only if the total height of liquid sensed by the instrument is fixed. A single hydrostatic-based level instrument cannot discern between a changing interface level and a changing total level, so the latter must be fixed in order to measure the former.

One way of fixing total liquid height is to equip the vessel with an overflow pipe, and ensure that drain flow is always less than incoming flow (forcing some flow to always go through the overflow pipe). This strategy naturally lends itself to separation processes, where a mixture of light and heavy liquids are separated by their differing densities:



Here we see a practical application for liquid-liquid interface level measurement. If the goal is to separate two liquids of differing densities from one another, we need only the light liquid to exit out the overflow pipe and only the heavy liquid to exit out the drain pipe. This means we must control the interface level to stay between those two piping points on the vessel. If the interface drifts too far up, heavy liquid will be carried out the overflow pipe; and if we let the interface drift too far down, light liquid will flow out of the drain pipe. The first step in controlling any process variable is to measure that variable, and so here we are faced with the necessity of measuring the interface point between the light and heavy liquids.

Another way of fixing the total height seen by the transmitter is to use a compensating leg located at a point on the vessel always lower than the total liquid height. In this example, a transmitter with remote seals is used:



Since both sides of the differential pressure transmitter "see" the hydrostatic pressure generated by the liquid column above the top connection point ($\gamma_2 h_3$), this term naturally cancels:

$$(\gamma_1 h_1 + \gamma_2 h_2 + \gamma_2 h_3) - (\gamma_4 h_4 + \gamma_2 h_3)$$

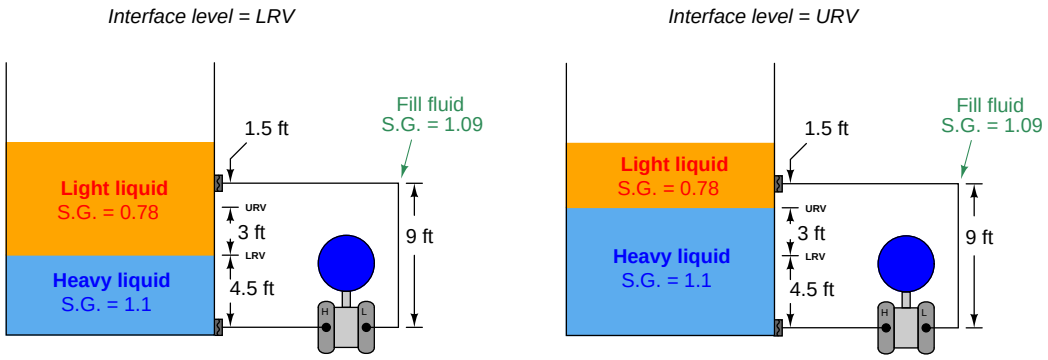
$$\gamma_1 h_1 + \gamma_2 h_2 + \gamma_2 h_3 - \gamma_4 h_4 - \gamma_2 h_3$$

$$\gamma_1 h_1 + \gamma_2 h_2 - \gamma_4 h_4$$

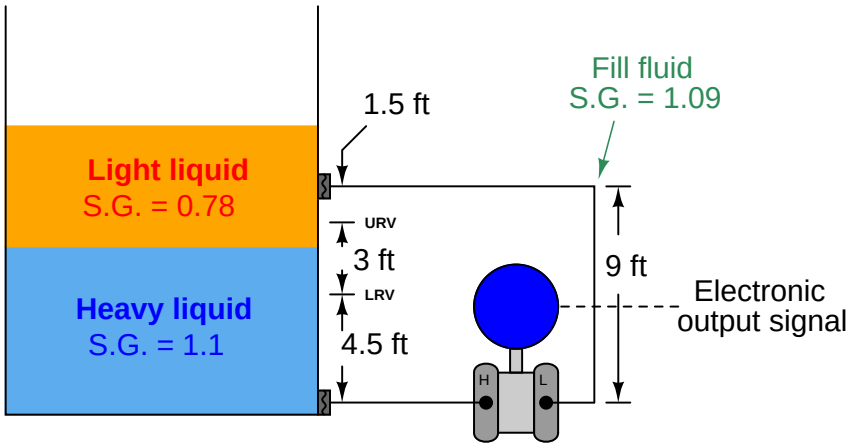
The hydrostatic pressure in the compensating leg is constant ($\gamma_4 h_4 = \text{Constant}$), since the fill fluid never changes density and the height never changes. This means the transmitter's sensed pressure will differ from that of an uncompensated transmitter merely by a constant offset, which may be "calibrated out" so as to have no impact on the measurement:

$$\gamma_1 h_1 + \gamma_2 h_2 - \text{Constant}$$

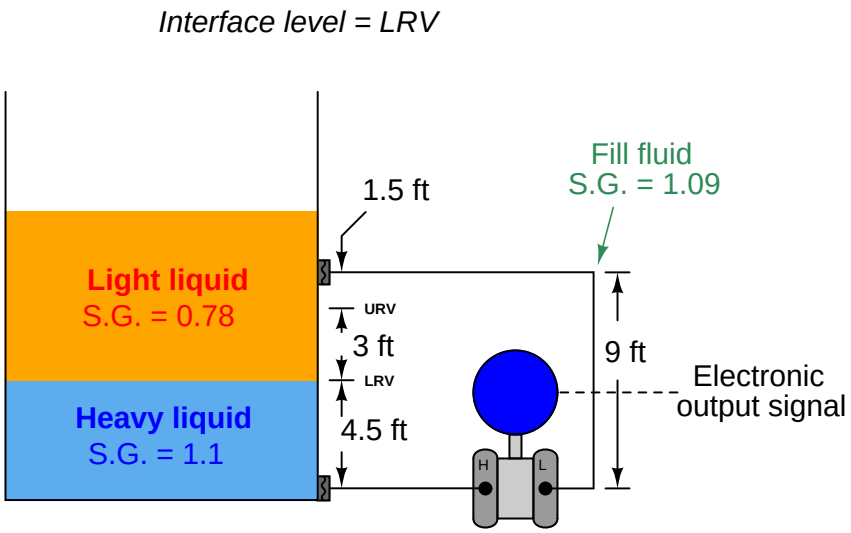
At first, it may seem as though determining the calibration points (lower- and upper-range values: LRV and URV) for a hydrostatic interface level transmitter is impossibly daunting given all the different pressures involved. A recommended problem-solving technique to apply here is that of a *thought experiment*, where we imagine what the process will "look like" at lower-range value condition and at the upper-range value condition, drawing two separate illustrations:



For example, suppose we must calibrate a differential pressure transmitter to measure the interface level between two liquids having specific gravities of 1.1 and 0.78, respectively, over a span of 3 feet. The transmitter is equipped with remote seals, each containing a halocarbon fill fluid with a specific gravity of 1.09. The physical layout of the system is as follows:



As the first step in our "thought experiment," we imagine what the process would look like with the interface at the LRV level, calculating hydrostatic pressures seen at each side of the transmitter:



We know from our previous exploration of this setup that any hydrostatic pressure resulting from liquid level *above* the top remote seal location is irrelevant to the transmitter, since it is "seen" on both sides of the transmitter and thus cancels out. All we must do, then, is calculate hydrostatic pressures as though the total liquid level stopped at that upper diaphragm connection point.

First, calculating the hydrostatic pressure "seen" at the high port of the transmitter[10]:

$$P_{high} = 4.5 \text{ feet of heavy liquid } + 4.5 \text{ feet of light liquid}$$

$$P_{high} = 54 \text{ inches of heavy liquid } + 54 \text{ inches of light liquid}$$

$$P_{high} \text{ "W.C. } = (54 \text{ inches of heavy liquid })(1.1) + (54 \text{ inches of light liquid })(0.78)$$

$$P_{high} \text{ "W.C. } = 59.4 \text{ "W.C. } + 42.12 \text{ "W.C.}$$

$$P_{high} = 101.52 \text{ "W.C.}$$

Next, calculating the hydrostatic pressure "seen" at the low port of the transmitter:

$$P_{low} = 9 \text{ feet of fill fluid}$$

$$P_{low} = 108 \text{ inches of fill fluid}$$

$$P_{low} \text{ "W.C. } = (108 \text{ inches of fill fluid })(1.09)$$

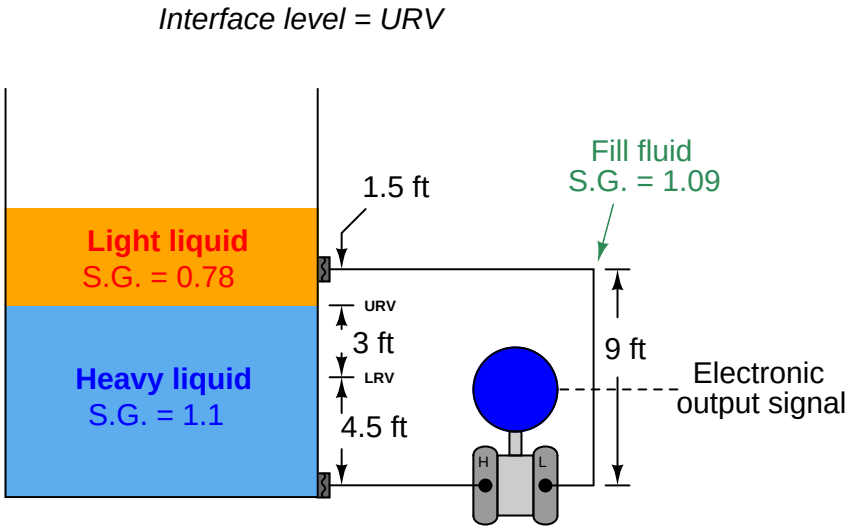
$$P_{low} = 117.72 \text{ "W.C.}$$

The differential pressure applied to the transmitter in this condition is the difference between the high and low port pressures, which becomes the lower range value (LRV) for calibration:

$$P_{LRV} = 101.52 \text{ "W.C. } - 117.72 \text{ "W.C. } = -16.2 \text{ "W.C.}$$

[10] Here I will calculate all hydrostatic pressures in units of inches water column. This is relatively easy because we have been given the specific gravities of each liquid, which make it easy to translate actual liquid column height into column heights of pure water.

As the second step in our "thought experiment," we imagine what the process would look like with the interface at the URV level, calculating hydrostatic pressures seen at each side of the transmitter:



$$P_{high} = 7.5 \text{ feet of heavy liquid } + 1.5 \text{ feet of light liquid}$$

$$P_{high} = 90 \text{ inches of heavy liquid } + 18 \text{ inches of light liquid}$$

$$P_{high} \text{ "W.C. } = (90 \text{ inches of heavy liquid })(1.1) + (18 \text{ inches of light liquid })(0.78)$$

$$P_{high} \text{ "W.C. } = 99 \text{ "W.C. } + 14.04 \text{ "W.C.}$$

$$P_{high} = 113.04 \text{ "W.C.}$$

The hydrostatic pressure of the compensating leg is exactly the same as it was before: 9 feet of fill fluid having a specific gravity of 1.09, which means there is no need to calculate it again. It will still be 117.72 inches of water column. Thus, the differential pressure at the URV point is:

$$P_{URV} = 113.04 \text{ "W.C. } - 117.72 \text{ "W.C. } = -4.68 \text{ "W.C.}$$

Using these two pressure values and some interpolation, we may generate a 5-point calibration table (assuming a 4-20 mA transmitter output signal range) for this interface level measurement system:

Interface level	Percent of range	Pressure at transmitter	Transmitter output
4.5 ft	0 %	-16.2 "W.C.	4 mA
5.25 ft	25 %	-13.32 "W.C.	8 mA
6 ft	50 %	-10.44 "W.C.	12 mA
6.75 ft	75 %	-7.56 "W.C.	16 mA
7.5 ft	100 %	-4.68 "W.C.	20 mA

When the time comes to bench-calibrate this instrument in the shop, the easiest way to do so will be to set the two remote diaphragms on the workbench (at the same level), then apply 16.2 to 4.68 inches of water column pressure to the *low* remote seal diaphragm with the other diaphragm at atmospheric pressure to simulate the desired range of negative differential pressures[11].

The more mathematically inclined reader will notice that the span of this instrument (URV − LRV) is equal to the span of the interface level (3 feet, or 36 inches) multiplied by the difference in specific gravities (1.1 − 0.78):

$$\text{Span in "W.C.} = (36 \text{ inches})(1.1 - 0.78)$$

$$\text{Span} = 11.52 \text{ "W.C.}$$

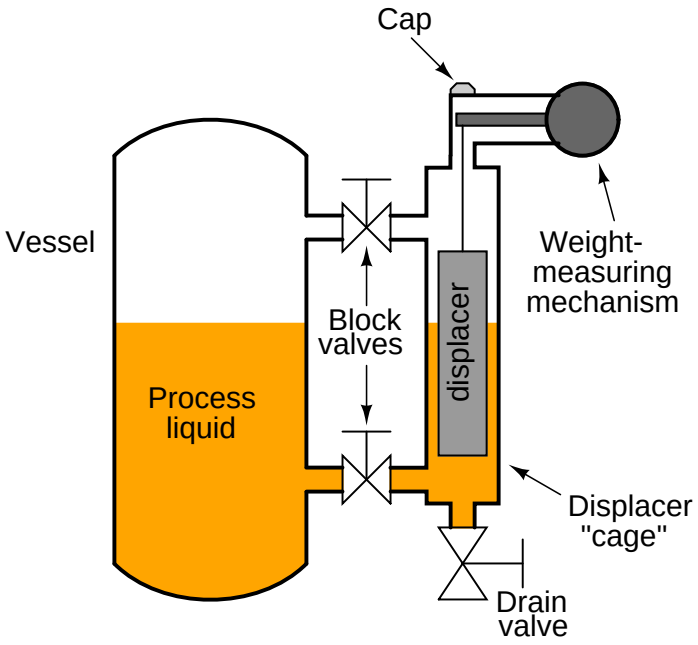
Looking at our two "thought experiment" illustrations, we see that the only difference between the two scenarios is the type of liquid filling that 3-foot region between the LRV and URV marks. Therefore, the only difference between the transmitter's pressures in those two conditions will be the difference in height multiplied by the difference in density. Not only is this an easy way for us to quickly calculate the necessary transmitter span, but it also is a way for us to check our previous work: we see that the difference between the LRV and URV pressures is indeed a difference of 11.52 inches water column just as this method predicts.

[11] Remember that a differential pressure instrument cannot "tell the difference" between a positive pressure applied to the low side, an equal vacuum applied to the high side, or an equivalent difference of two positive pressures with the low side's pressure exceeding the high side's pressure. Simulating the exact process pressures experienced in the field to a transmitter on a workbench would be exceedingly complicated, so we "cheat" by simplifying the calibration setup and applying the equivalent difference of pressure only to the "low" side.

19.4 Displacement

Displacer level instruments exploit *Archimedes' Principle* to detect liquid level by continuously measuring the weight of a rod immersed in the process liquid. As liquid level increases, the displacer rod experiences a greater buoyant force, making it appear lighter to the sensing instrument, which interprets the loss of weight as an increase in level and transmits a proportional output signal.

In practice a displacer level instrument usually takes the following form:



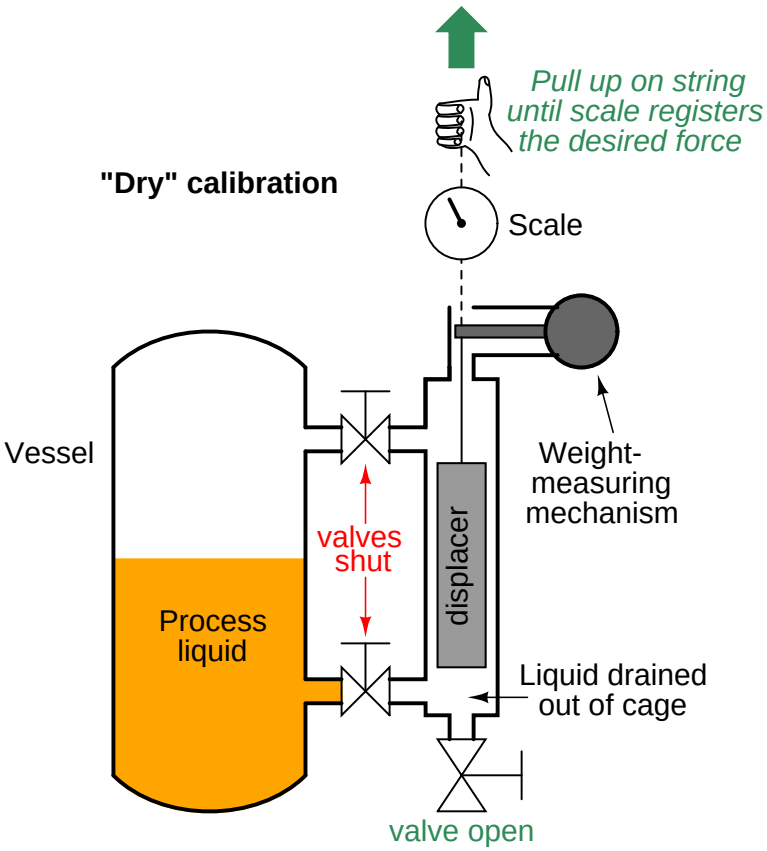
The following photograph shows a Fisher "Level-Trol" model pneumatic transmitter measuring condensate level in a *knockout drum*[12] for natural gas service. The instrument itself appears on the right-hand side of the photo, topped by a grey-colored "head" with two pneumatic pressure gauges visible. The displacer "cage" is the vertical pipe immediately behind and below the head unit. Note that a sightglass level gauge appears on the left-hand side of the knockout chamber (or *condensate boot*) for visual indication of condensate level inside the process vessel:

[12]So-called for its ability to "knock out" (separate and collect) condensible vapors from the gas stream. This particular photograph was taken at a natural gas compression facility, where it is very important the gas to be compressed is dry (since liquids are essentially incompressible). Sending even relatively small amounts of liquid into a compressor may cause the compressor to catastrophically fail!

Two photos of a disassembled Level-Trol displacer instrument appear here, showing how the displacer fits inside the cage pipe:

The cage pipe is coupled to the process vessel through two block valves, allowing isolation from the process. A drain valve allows the cage to be emptied of process liquid for instrument service and zero calibration.

Full-range calibration may be performed by flooding the cage with process liquid (a *wet* calibration), or by suspending the displacer with a string and precise scale (a *dry* calibration), pulling upward on the displacer at just the right amount to simulate buoyancy at 100% liquid level:



Calculation of this buoyant force is a simple matter. According to Archimedes' Principle, buoyant force is always equal to the weight of the fluid volume displaced. In the case of a displacer-based level instrument at full range, this usually means the entire volume of the displacer element is submerged in the liquid. Simply calculate the volume of the displacer (if it is a cylinder, $V = \pi r^2 l$, where r is the cylinder radius and l is the cylinder length) and multiply that volume by the weight density (γ):

$$F_{buoyant} = \gamma V$$

$$F_{buoyant} = \gamma \pi r^2 l$$

For example, if the weight density of the process fluid is 57.3 pounds per cubic foot and the displacer is a cylinder measuring 3 inches in diameter and 24 inches in length, the necessary force to simulate a condition of buoyancy at full level may be calculated as follows:

$$\gamma = \left(\frac{57.3 \text{ lb}}{\text{ft}^3}\right)\left(\frac{1 \text{ ft}^3}{12^3 \text{ in}^3}\right) = 0.0332\frac{\text{lb}}{\text{in}^3}$$

$$V = \pi r^2 l = \pi(1.5 \text{ in})^2(24 \text{ in}) = 169.6 \text{ in}^3$$

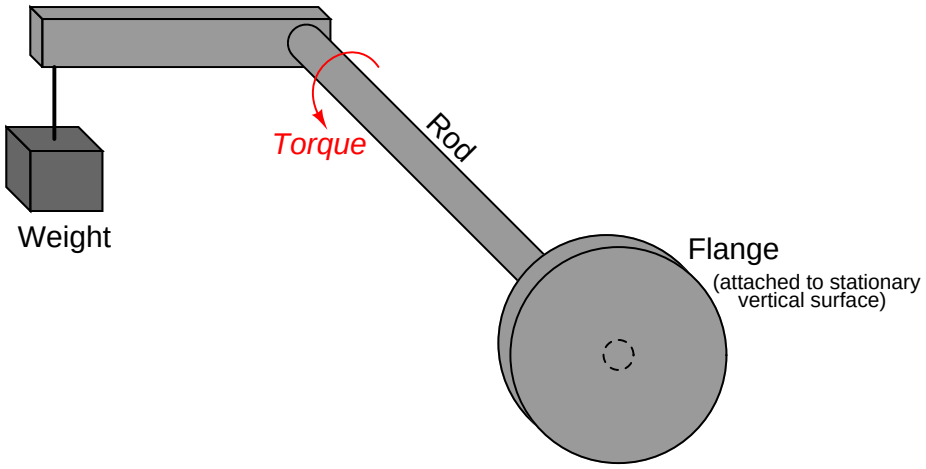
$$F_{buoyant} = \gamma V = \left(0.0332\frac{\text{lb}}{\text{in}^3}\right)\left(169.6 \text{ in}^3\right) = 5.63 \text{ lb}$$

Note how important it is to maintain consistency of units! The liquid density was given in units of pounds per cubic *foot* and the displacer dimensions in *inches*, which would have caused serious problems without a conversion between feet and inches. In my example work, I opted to convert density into units of pounds per cubic inch, but I could have just as easily converted the displacer dimensions into feet to arrive at a displacer volume in units of cubic feet.

19.4.1 Torque tubes

An interesting engineering problem for displacement-type level transmitters is how to transfer the sensed weight of the displacer to the transmitter mechanism while positively sealing process vapor pressure from that same mechanism. The most common solution to this problem is an ingenious mechanism called a *torque tube*. Unfortunately, torque tubes can be rather difficult to understand unless you have direct hands-on access to one, and so this section will explore the concept in more detail than is customarily available in reference manuals.

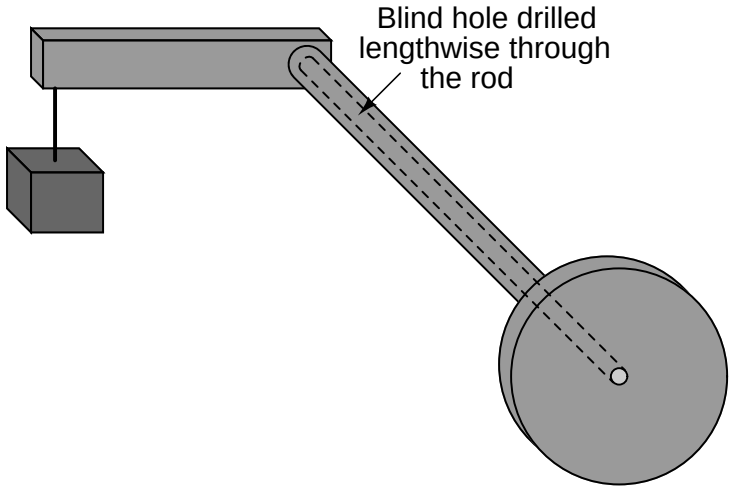
Imagine a solid, horizontal, metal rod with a flange at one end and a perpendicular lever at the other end. The flange is mounted to a stationary surface, and a weight suspended from the end of the lever. A dashed-line circle shows where the rod is welded to the center of the flange:



The downward force of the weight acting on the lever imparts a twisting force (torque) to the rod, causing it to slightly twist along its length. The more weight hung at the end of the lever, the more the rod will twist[13]. So long as the torque applied by the weight and lever never exceeds the elastic limit of the rod, the rod will continue to act as a spring. If we knew the "spring constant" of the rod, and measured its torsional deflection, we could in fact use this slight motion to measure the magnitude of the weight hung at the end of the lever.

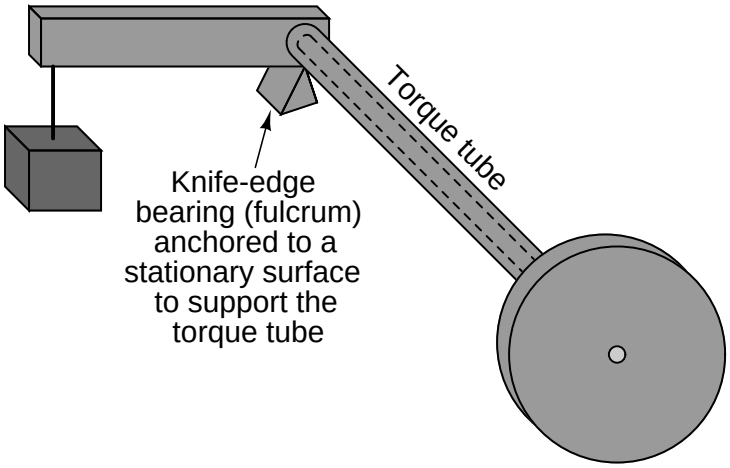
[13] To anyone familiar with the front suspension of a 1960's vintage Chevrolet truck, or the suspension of the original Volkswagen "Beetle" car, the concept of a *torsion bar* should be familiar. These vehicles used straight, spring-steel rods to provide suspension force instead of the more customary coil springs used in modern vehicles. However, even the familiar coil spring is an example of torsional forces at work: a coil spring is nothing more than a torsion bar bent in a coil shape. As a coil spring is stretched or compressed, torsional forces develop along the circumferential length of the spring coil, which is what makes the spring "try" to maintain a fixed height.

Now imagine drilling a long hole through the rod, lengthwise, that almost reaches the end where the lever attaches. In other words, imagine a *blind hole* through the center of the rod, starting at the flange and ending just shy of the lever:

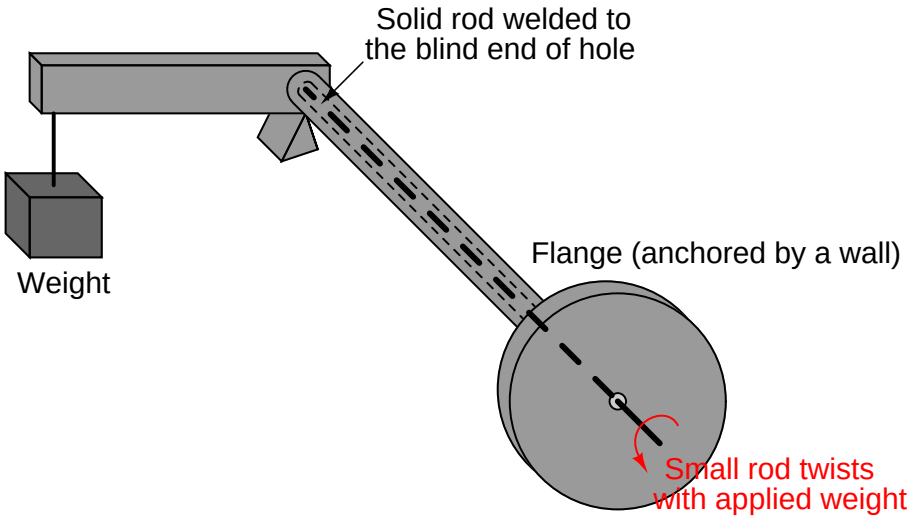


The presence of this long hole does not change much about the behavior of the assembly, except perhaps to alter the rod's spring constant. With less girth, the rod will be a weaker spring, and will twist to a greater degree with applied weight at the end of the lever. More importantly for the purpose of this discussion, though, the long hole transforms the rod into a *tube* with a sealed end. Instead of a being a "torsion bar," the rod is now more properly called a *torque tube*, twisting ever so slightly with applied weight at the end of the lever.

In order to give the torque tube vertical support so it does not sag downward with the applied weight, a supporting *knife-edge bearing* is often placed underneath the end of the lever where it attaches to the torque tube. The purpose of this fulcrum is to provide vertical support for the weight while forming a virtually frictionless pivot point, ensuring the only stress applied to the torque tube is *torque* from the lever:



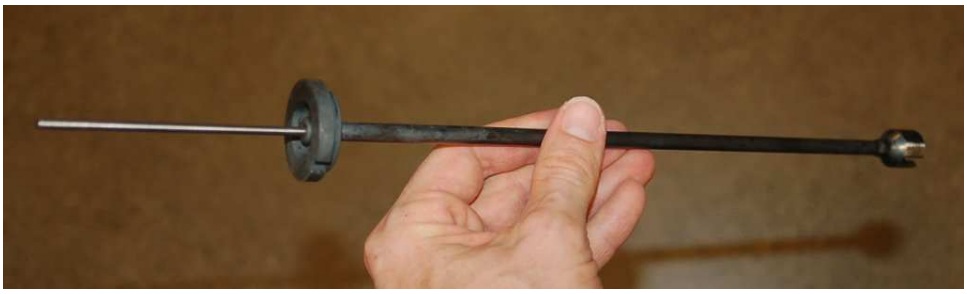
Finally, imagine another solid metal rod (slightly smaller diameter than the hole) spot-welded to the far end of the blind hole, extending out beyond the end of the flange:



The purpose of this smaller-diameter rod is to transfer the twisting motion of the torque tube to a point past the flange where it may be sensed. Imagine the flange anchored to a vertical wall, while a variable weight tugs downward at the end of the lever. The torque tube will flex in a twisting motion with the variable force, but now we are able to see just how much it twists by watching the rotation of the smaller rod on the near side of the wall. The weight and lever may be completely hidden from our view by this wall, but the small rod's twisting motion nevertheless reveals how much the torque tube yields to the force of the weight.

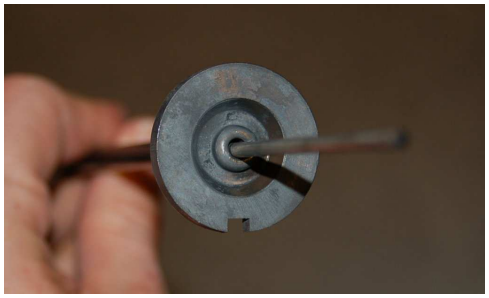
We may apply this torque tube mechanism to the task of measuring liquid level in a pressurized vessel by replacing the weight with a displacer, attaching the flange to a nozzle welded to the vessel, and aligning a motion-sensing device with the small rod end to measure its rotation. As liquid level rises and falls, the apparent weight of the displacer varies, causing the torque tube to slightly twist. This slight twisting motion is then sensed at the end of the small rod, in an environment isolated from the process fluid pressure.

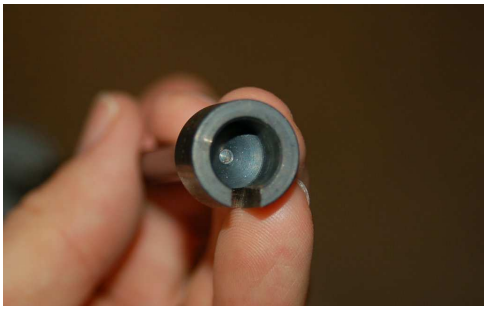
A photograph taken of a real torque tube from a Fisher "Level-Trol" level transmitter shows its external appearance:



The dark-colored metal is the elastic steel used to suspend the weight by acting as a torsional spring, while the shiny portion is the inner rod used to transfer motion. As you can see, the torque tube itself is not very wide in diameter. If it were, it would be far too stiff of a spring to be of practical use in a displacer-type level instrument, since the displacer is not typically very heavy, and the lever is not long.

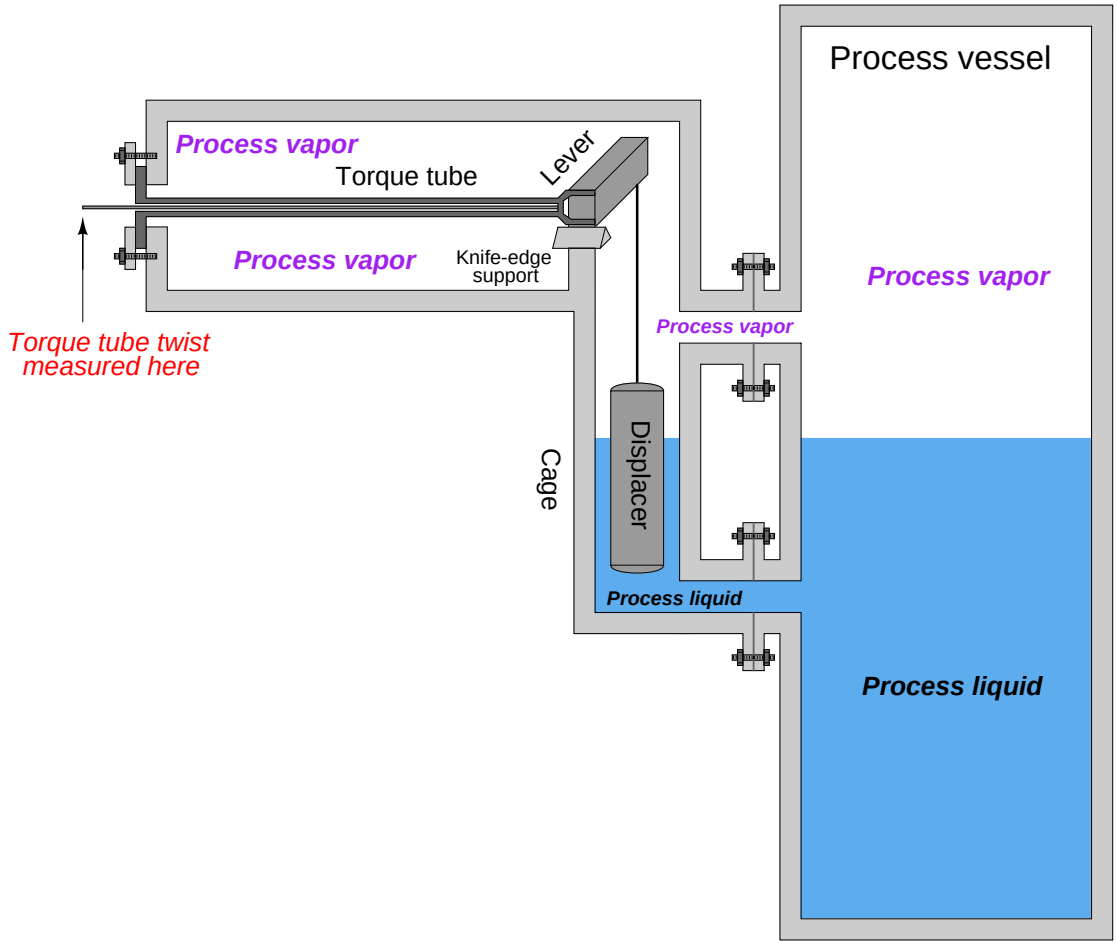
Looking closer at each end of the torque tube reveals the open end where the small-diameter rod protrudes (left) and the "blind" end of the tube where it attaches to the lever (right):





If we were to slice the torque tube assembly in half, lengthwise, its cross-section would look something like this:

This next illustration shows the torque tube as part of a whole displacement-style level transmitter[14]:



As you can see from this illustration, the torque tube serves three distinct purposes when applied to a displacer-type level measurement application: (1) to serve as a torsional spring suspending the weight of the displacer, (2) to seal off process fluid pressure from the position-sensing mechanism, and (3) to transfer motion from the far end of the torque tube into the sensing mechanism.

[14]This illustration is simplified, omitting such details as access holes into the cage, block valves between the cage and process vessel, and any other pipes or instruments attached to the process vessel. Also, the position-sensing mechanism normally located at the far left of the assembly is absent from this drawing.

In pneumatic level transmitters, the sensing mechanism used to convert the torque tube's twisting motion into a pneumatic (air pressure) signal is typically of the *motion-balance* design. The Fisher Level-Trol mechanism, for example, uses a C-shaped bourdon tube with a nozzle at the end to follow a baffle attached to the small rod. The center of the bourdon tube is aligned with the center of the torque tube. As the rod rotates, the baffle advances toward the nozzle at the bourdon tube tip, causing backpressure to rise, which in turn causes the bourdon tube to flex. This flexing draws the nozzle away from the advancing baffle until a balanced condition exists. Rod motion is therefore balanced by bourdon tube motion, making this a motion-balance pneumatic system:

19.4.2 Displacement interface level measurement

Displacer level instruments may be used to measure liquid-liquid interfaces just the same as hydrostatic pressure instruments. One important requirement is that the displacer always be fully submerged. If this rule is violated, the instrument will not be able to "tell" the difference between a low (total) liquid level and a low interface level.

If the displacer instrument has its own "cage," it is important that both pipes connecting the cage to the process vessel (sometimes called "nozzles") be submerged. This ensures the liquid interface inside the cage matches the interface inside the vessel. If the upper nozzle ever goes dry, the same problem can happen with a caged displacer instrument as with a "sightglass" level gauge (see page 849 for a detailed explanation of this problem.).

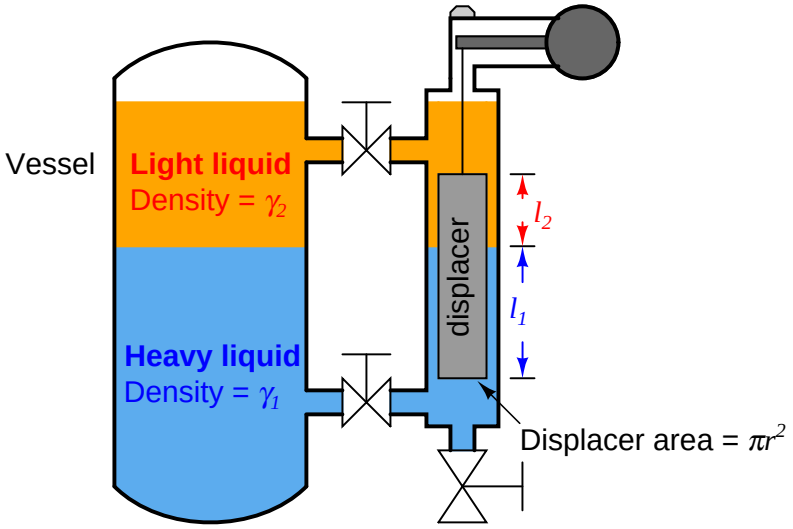
Calculating buoyant force on a displacer element due to a combination of two liquids is not as difficult as it may sound. Archimedes' Principle still holds: that buoyant force is equal to the weight of the fluid(s) displaced. All we need to do is calculate the combined weights and volumes of the displaced liquids to calculate buoyant force. For a single liquid, buoyant force is equal to the weight density of that liquid (γ) multiplied by the volume displaced (V):

$$F_{buoyant} = \gamma V$$

For a two-liquid interface, the buoyant force is equal to the sum of the two liquid weights displaced, each liquid weight term being equal to the weight density of that liquid multiplied by the displaced volume of that liquid:

$$F_{buoyant} = \gamma_1 V_1 + \gamma_2 V_2$$

Assuming a displacer of constant cross-sectional area throughout its length, the volume for each liquid's displacement is simply equal to the same area (πr^2) multiplied by the length of the displacer submerged in that liquid:



$$F_{buoyant} = \gamma_1 \pi r^2 l_1 + \gamma_2 \pi r^2 l_2$$

Since the area (πr^2) is common to both buoyancy terms in this equation, we may factor it out for simplicity's sake:

$$F_{buoyant} = \pi r^2 (\gamma_1 l_1 + \gamma_2 l_2)$$

Determining the calibration points of a displacer-type level instrument for interface applications is relatively easy if the LRV and URV conditions are examined as a pair of "thought experiments" just as we did with hydrostatic interface level measurement. First, we imagine what the displacer's condition would "look like" with the interface at the lower range value, then we imagine a different scenario with the interface at the upper range value.

Suppose we have a displacer instrument measuring the interface level between two liquids having specific gravities of 0.850 and 1.10, with a displacer length of 30 inches and a displacer diameter of 2.75 inches (radius = 1.375 inches). Let us further suppose that the LRV in this case is where the interface is at the displacer's bottom and the URV is where the interface is at the displacer's top. The placement of the LRV and URV interface levels at the extreme ends of the displacer's length simplifies our LRV and URV calculations, as the LRV "thought experiment" will simply be the displacer completely submerged in light liquid and the URV "thought experiment" will simply be the displacer completely submerged in heavy liquid.

Calculating the LRV buoyant force:

$$F_{buoyant}\ (\text{LRV}) = \pi r^2 \gamma_2 L$$

Calculating the URV buoyant force:

$$F_{buoyant}\ (\text{URV})\ = \pi r^2 \gamma_1 L$$

The buoyancy for any measurement percentage between the LRV (0%) and URV (100%) may be calculated by interpolation:

$$\gamma_1 = \left(62.4\ \frac{\text{lb}}{\text{ft}^3}\right)(1.10) = 68.6\ \frac{\text{lb}}{\text{ft}^3} = 0.0397\ \frac{\text{lb}}{\text{in}^3}$$

$$\gamma_2 = \left(62.4\ \frac{\text{lb}}{\text{ft}^3}\right)(0.85) = 53.0\ \frac{\text{lb}}{\text{ft}^3} = 0.0307\ \frac{\text{lb}}{\text{in}^3}$$

$$F_{buoyant}\ (\text{LRV})\ = \pi(1.375\ \text{in})^2 \left(0.0307\ \frac{\text{lb}}{\text{in}^3}\right)(30\ \text{in}) = 5.47\ \text{lb}$$

$$F_{buoyant}\ (\text{URV})\ = \pi(1.375\ \text{in})^2 \left(0.0397\ \frac{\text{lb}}{\text{in}^3}\right)(30\ \text{in}) = 7.08\ \text{lb}$$

Interface level (inches)	Buoyant force (pounds)
0	5.47
7.5	5.87
15	6.27
22.5	6.68
30	7.08

19.5 Echo

A completely different way of measuring liquid level in vessels is to bounce a traveling wave off the surface of the liquid – typically from a location at the top of the vessel – using the time-of-flight for the waves as an indicator of distance[15], and therefore an indicator of liquid height inside the vessel. Echo-based level instruments enjoy the distinct advantage of immunity to changes in liquid density, a factor crucial to the accurate calibration of hydrostatic and displacement level instruments. In this regard, they are quite comparable with float-based level measurement systems.

From a historical perspective, hydrostatic and displacement level instruments have a richer pedigree. These instruments are simpler in nature than echo-based instruments, and were practical long before the advent of modern electronic technology. Echo-based instruments require precision timing and wave-shaping circuitry, plus sensitive (and rugged!) transceiver elements, demanding a much higher level of technology. However, modern electronic design and instrument manufacturing practices are making echo-based level instruments more and more practical for industrial applications. At the time of this writing (2008), it is common practice in some industries to replace old displacer level instruments with guided-wave radar instruments, even in demanding applications operating at high pressures[16].

Liquid-liquid interfaces may also be measured with some types of echo-based level instruments, most commonly guided-wave radar.

The single most important factor to the accuracy of an echo-based level instrument is the speed at which the wave travels en route to the liquid surface and back. This wave propagation speed is as fundamental to the accuracy of an echo instrument as liquid density is to the accuracy of a hydrostatic or displacer instrument. So long as this velocity is known and stable, good level measurement accuracy is possible. Although it is true that the calibration of an echo-based level instrument does not depend on process fluid density for the reason it does in hydrostatic- or displacement-based level instruments, this does not necessarily mean the calibration of an echo-based level instrument remains fixed as process fluid density changes. The propagation velocity of the wave used in an echo-based level instrument may indeed be subject to change as the process fluids change temperature or composition. For ultrasonic (sound) echo instruments, the speed of sound is a strong function of medium density. Thus, an ultrasonic level transmitter measuring time-of-flight through a vapor above the liquid may drift out of calibration if the density (i.e. speed of sound) in that vapor changes substantially, which may happen if the vapor's temperature or pressure happens to change. If the sound wave time-of-flight is measured while the waves pass through liquid, the calibration may drift if the speed of sound in that liquid changes substantially, which may happen if the liquid's temperature changes. For radar (radio wave) echo instruments, the speed of radio wave propagation varies according to the dielectric permittivity of the medium. Permittivity is also affected by changes in density for the fluid medium, and so even radar level instruments may suffer calibration drift with process fluid density changes.

Echo-based level instruments may be "fooled" by layers of foam resting on top of the liquid, and the liquid-to-liquid interface detection models may have difficulty detecting non-distinct interfaces (such as emulsions). Irregular structures residing within the vapor space of a vessel (such as access

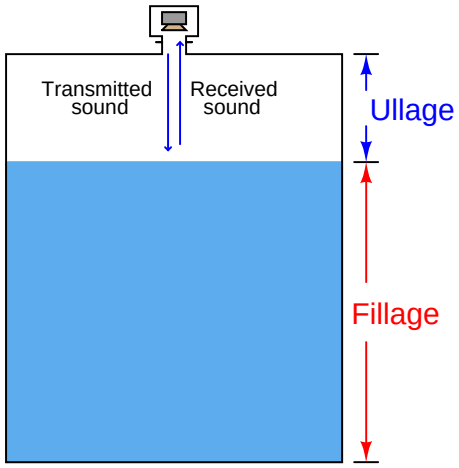
[15] The general term for this form of measurement is *time domain reflectometry*.

[16] My own experience with this trend is within the oil refining industry, where legacy displacer instruments (typically Fisher brand "Level-Trol" units) are being replaced with new guided-wave radar transmitters, both for single-liquid and liquid-liquid interface applications.

portals, mixer paddles and shafts, ladders, etc.) may wreak havoc with echo-based level instruments by casting false echoes back to the instrument, although this problem may be mitigated by installing guide tubes for the waves to travel in, or using wave probes as in the cases of guided-wave radar instruments. Liquid streams pouring in to the vessel through the vapor space may similarly cause problems for an echo instrument. Additionally, all echo-based instruments have *dead zones* where liquid level is too close to the transceiver to be accurately measured or even detected (the echo time-of-flight being too short for the receiving electronics to distinguish from the incident pulse).

19.5.1 Ultrasonic level measurement

Ultrasonic level instruments measure the distance from the transmitter (located at some high point) to the surface of a process material located further below. The time-of-flight for a sound pulse indicates this distance, and is interpreted by the transmitter electronics as process level. These transmitters may output a signal corresponding either to the fullness of the vessel (*fillage*) or the amount of empty space remaining at the top of a vessel (*ullage*).



Ullage is the "natural" mode of measurement for this sort of level instrument, because the sound wave's time-of-flight is a direct function of how much empty space exists between the liquid surface and the top of the vessel. Total tank height will always be the sum of fillage and ullage, though. If the ultrasonic level transmitter is programmed with the vessel's total height, it may calculate fillage via simple subtraction:

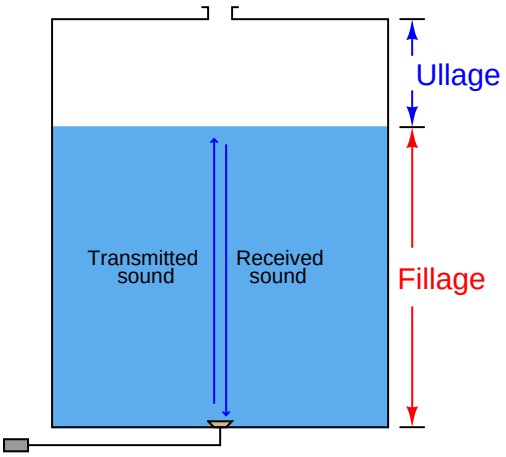
$$\text{Fillage} = \text{Total height} - \text{Ullage}$$

The instrument itself consists of an electronics module containing all the power, computation, and signal processing circuits; plus an ultrasonic transducer[17] to send and receive the sound waves. This transducer is typically piezoelectric in nature, being the equivalent of a very high-frequency audio speaker. The following photographs show a typical electronics module (left) and sonic transducer (right):

The ISA-standard designations for each component would be "LT" (level transmitter) for the electronics module and "LE" (level element) for the transducer, respectively. Even though we call the device responsible for transmitting and receiving the sound waves a *transducer* (in the scientific sense of the word), its function as a process instrument is to be the *primary sensing element* for the level measurement system, and therefore it is more properly designated a "level element" (LE).

[17] In the industrial instrumentation world, the word "transducer" usually has a very specific meaning: a device used to process or convert standardized instrumentation signals, such as 4-20 mA converted into 3-15 PSI, etc. In the general scientific world, however, the word "transducer" describes any device converting one form of energy into another. It is this latter definition of the word that I am using when I describe an ultrasonic "transducer" – a device used to convert electrical energy into ultrasonic sound waves, and visa-versa.

If the ultrasonic transducer is rugged enough, and the process vessel sufficiently free of sludge and other sound-damping materials accumulating at the vessel bottom, the transducer may be mounted at the bottom of the vessel, bouncing sound waves off the liquid surface through the liquid itself rather than through the vapor space:



This arrangement makes fillage the natural measurement, and ullage a derived measurement (calculated by subtraction from total vessel height).

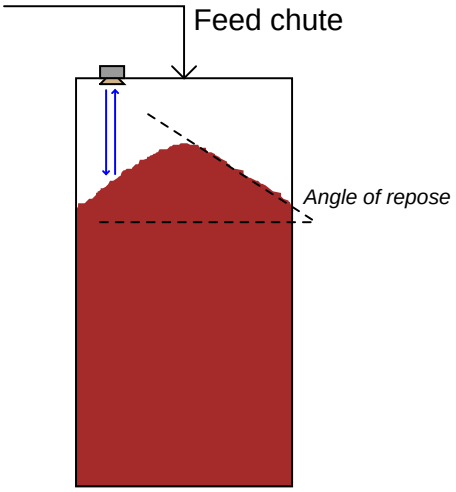
$$\text{Ullage} = \text{Total height} - \text{Fillage}$$

Whether the ultrasonic transducer is mounted above or below the liquid level, the principle of detection is any significant difference in material *density*. If the detection interface is between a gas and a liquid, the abrupt change in density is enough to create a strong reflected signal. However, it is possible for foam and floating solids to also cause echos when the transducer is above-mounted, which may or may not be desirable depending on the application[18].

As mentioned previously, the calibration of an ultrasonic level transmitter depends on the speed of sound through the medium between the transducer and the interface. For top-mounted transducers, this is the speed of sound through the air (or vapor) over the liquid, since this is the medium through which the incident and reflected wave travel time is measured. For bottom-mounted transducers, this is the speed of sound through the liquid. In either case, to ensure good accuracy, one must make sure the speed of sound through the "timed" travel path remains reasonably constant (or else compensate for changes in the speed of sound through that medium by use of temperature or pressure measurements and a compensating algorithm).

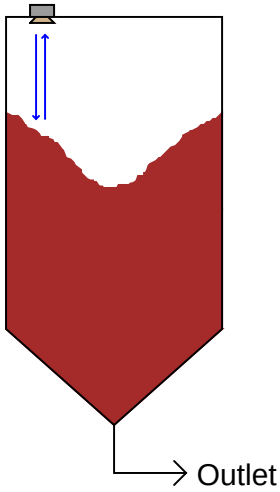
[18] If the goal is to only detect the liquid, then reflections from foam or solids would be bad. However, if the goal of measuring level is to prevent a vessel from overflowing, it is good to measure *anything* floating on the liquid surface!

Ultrasonic level instruments enjoy the advantage of being able to measure the height of solid materials such as powders and grains stored in vessels, not just liquids. Certain challenges unique to these level measurement applications include low material density (not causing strong reflections) and uneven profiles (causing reflections to be scattered laterally instead of straight back to the ultrasonic instrument. A classic problem encountered when measuring the level of a powdered or granular material in a vessel is the *angle of repose* formed by the material as a result of being fed into the vessel at one point:



This angled surface is difficult for an ultrasonic device to detect because it tends to scatter the sound waves laterally instead of reflecting them strongly back toward the instrument. However, even if the scattering problem is not significant, there still remains the problem of interpretation: what is the instrument actually measuring? The detected level near the vessel wall will certainly register less than at the center, but the level detected mid-way between the vessel wall and vessel center may not be an accurate average of those two heights. Moreover, this angle may decrease over time if mechanical vibrations cause the material to "flow" and tumble from center to edge.

In fact, the angle will probably reverse itself if the vessel empties from a center-located chute:

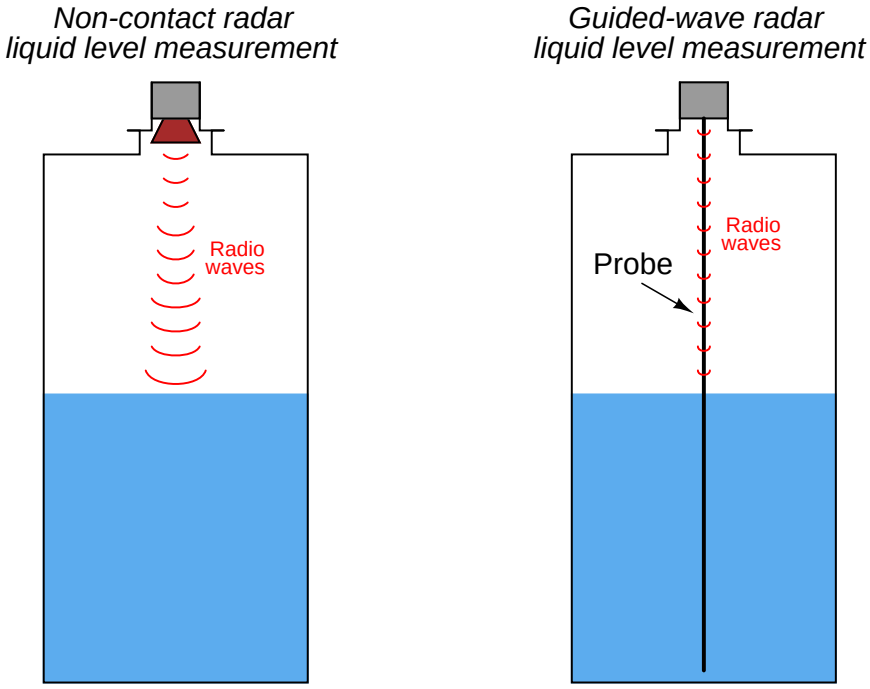


For this reason, solids storage measurement applications demanding high accuracy generally use other techniques, such as weight-based measurement (see section 19.6 for more information).

19.5.2 Radar level measurement

Radar[19] level instruments measure the distance from the transmitter (located at some high point) to the surface of a process material located further below in much the same way as ultrasonic transmitters – by measuring the time-of-flight of a traveling wave. The fundamental difference between a radar instrument and an ultrasonic instrument is the type of wave used: radio waves instead of sound waves. Radio waves are electromagnetic in nature (comprised of alternating electric and magnetic fields), and very high frequency (in the microwave frequency range – GHz). Sound waves are *mechanical* vibrations (transmitted from molecule to molecule in a fluid or solid substance) and of much lower frequency (tens or hundreds of kilohertz – still too high for a human being to detect as a tone) than radio waves.

Some radar level instruments use waveguide "probes" to guide the radio waves into the process liquid while others send radio waves out through open space to reflect off the process material. The instruments using waveguides are called *guided-wave radar* instruments, whereas the radar instruments relying on open space for signal propagation are called *non-contact radar*. The differences between these two varieties of radar instruments is shown in the following illustration:

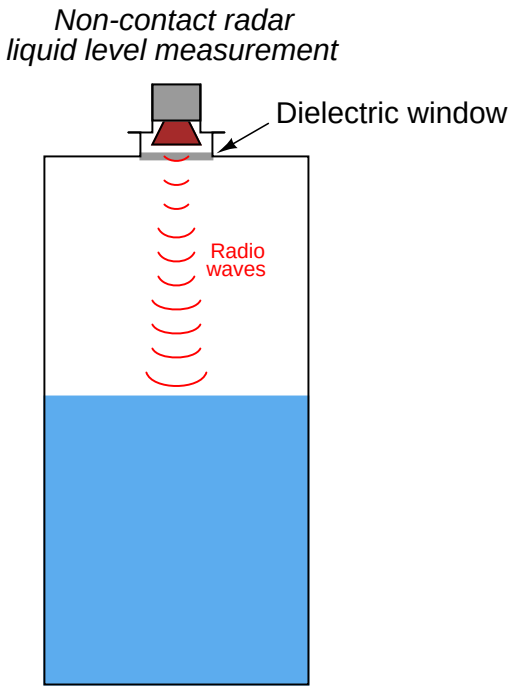


[19] "Radar" is an acronym: RAdio Detection And Ranging. First used as a method for detecting enemy ships and aircraft at long distances over the ocean in World War II, this technology is used for detecting the presence, distance, and/or speed of objects in a wide variety of applications.

Non-contact radar transmitters are always mounted on the top side of a storage vessel. Modern radar transmitters are quite compact, as this photograph shows:

Probes used in guided-wave radar instruments may be single metal rods, parallel pairs of metal rods, or a coaxial metal rod-and-tube structure. Single-rod probes radiate the most energy, whereas coaxial probes do the best job guiding the microwave energy to the liquid interface and back. However, single-rod probes are much more tolerant of process fouling than two-rod or (especially) coaxial probes, where sticky masses of viscous liquid and/or solid matter cling to the probe. Such fouling deposits, if severe enough, will cause radio energy reflections that "look" to the transmitter like the reflection from an actual liquid level or interface.

Non-contact radar instruments rely on antennae to direct microwave energy into the vessel, and to receive the echo (return) energy. These antennae must be kept clean and dry, which may be a problem if the liquid being measured emits condensible vapors. For this reason, non-contact radar instruments are often separated from the vessel interior by means of a *dielectric window* (made of some substance that is relatively "transparent" to radio waves yet acts as an effective vapor barrier):



Radio waves travel at the velocity of light[20], 2.9979×10^8 meters per second in a perfect vacuum. The velocity of a radio wave through space depends on the dielectric permittivity (symbolized by the Greek letter "epsilon," ϵ) of that space. A formula relating wave velocity to relative permittivity (the ratio of a substance's permittivity to that of a perfect vacuum, symbolized as ϵ_r and sometimes called the *dielectric constant* of the substance) and the velocity of light in a perfect vacuum (c) is shown here[21]:

$$v = \frac{c}{\sqrt{\epsilon_r}}$$

[20] In actuality, both radio waves and light waves are electromagnetic in nature. The only difference between the two is frequency: while the radio waves used in radar systems are classified as "microwaves" with frequencies in the gigahertz (GHz) region, visible light waves range in the hundred of terahertz (THz)!

[21] This formula assumes lossless conditions: that none of the wave's energy is converted to heat while traveling through the dielectric. For many situations, this is true enough to assume.

As mentioned previously, the calibration of an echo-based level transmitter depends on knowing the speed of wave propagation through the medium separating the instrument from the process fluid interface. For radar transmitters sensing a single liquid below a gas or vapor, this speed is the speed of light through that gas or vapor space, which we know to be a function of electrical permittivity.

The relative permittivity of air at standard pressure and temperature is very nearly unity (1). This means the speed of light in air under atmospheric pressure and ambient temperature will very nearly be the same as it is for a perfect vacuum (2.9979×10^8 meters per second). If, however, the vapor space above the liquid is not ambient air, and is subject to large changes in temperature and/or pressure[22], the permittivity of that vapor may substantially change and consequently skew the speed of light, and therefore the calibration of the level instrument.

The permittivity of any gas is related to both pressure and temperature by the following formula:

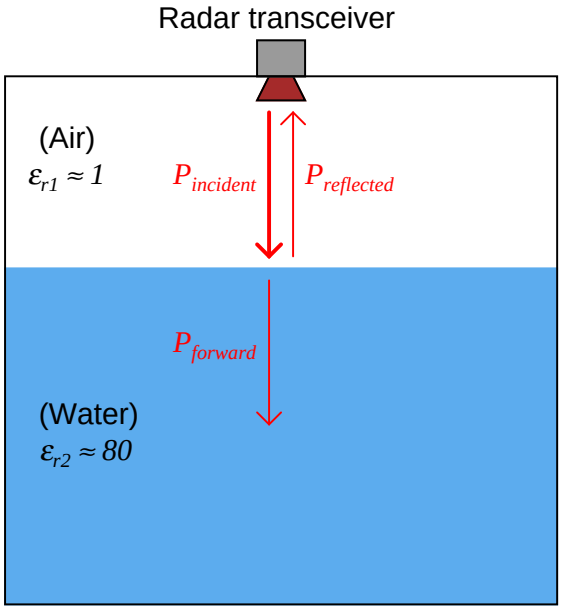
$$\epsilon_r = 1 + (\epsilon_{ref} - 1)\frac{PT_{ref}}{P_{ref}T}$$

Where,

ϵ_r = Relative permittivity of a gas at a given pressure (P) and temperature (T)
ϵ_{ref} = Relative permittivity of the same gas at standard pressure (P_{ref}) and temperature (T_{ref})
P = Absolute pressure of gas (bars)
P_{ref} = Absolute pressure of gas under standard conditions ($\approx$ 1 bar)
T = Absolute temperature of gas (Kelvin)
T_{ref} = Absolute temperature of gas under standard conditions ($\approx$ 273 K)

[22] Or if the chemical composition of the gas or vapor changes dramatically.

If a radio wave encounters a sudden change in dielectric permittivity, some of that wave's energy will be reflected in the form of another wave traveling the opposite direction. In other words, the wave will "echo" when it reaches a discontinuity. This is the basis of all radar devices:



This same principle explains reflected signals in copper transmission lines as well. Any discontinuities (sudden changes in characteristic impedance) along the length of a transmission line will reflect a portion of the electrical signal's power back to the source. In a transmission line, continuities may be formed by pinches, breaks, or short-circuits. In a radar level measurement system, any sudden change in permittivity is a discontinuity that reflects some of the incident radio energy back to the source.

The ratio of reflected power to incident (transmitted) power at any interface of materials is called the *power reflection factor* (R). This may be expressed as a unitless ratio, or more often as a decibel figure. The relationship between dielectric permittivity and reflection factor is as follows:

$$R = \frac{\left(\sqrt{\epsilon_{r2}} - \sqrt{\epsilon_{r1}}\right)^2}{\left(\sqrt{\epsilon_{r2}} + \sqrt{\epsilon_{r1}}\right)^2}$$

Where,

R = Power reflection factor at interface, as a unitless ratio
ϵ_{r1} = Relative permittivity (dielectric constant) of the first medium
ϵ_{r2} = Relative permittivity (dielectric constant) of the second medium

The fraction of incident power transmitted through the interface ($\frac{P_{forward}}{P_{incident}}$) is, of course, the mathematical complement of the power reflection factor: $1 - R$.

For situations where the first medium is air or some other low-permittivity gas, the formula simplifies to the following form (with ϵ_r being the relative permittivity of the reflecting substance):

$$R = \frac{\left(\sqrt{\epsilon_r} - 1\right)^2}{\left(\sqrt{\epsilon_r} + 1\right)^2}$$

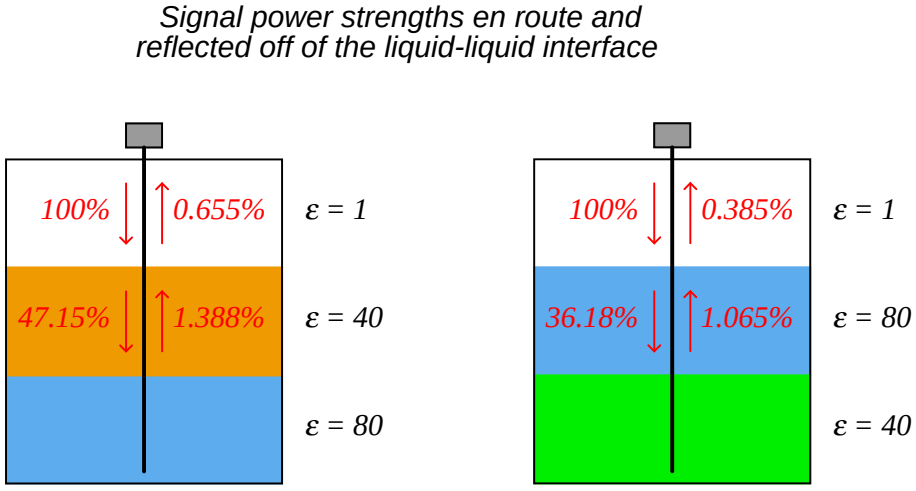
In the previous illustration, the two media were air ($\epsilon_r \approx 1$) and water ($\epsilon_r \approx 80$) – a nearly ideal scenario for strong signal reflection. Given these relative permittivity values, the power reflection factor has a value of 0.638 (63.8%), or -1.95 dB. This means that well over half the incident power reflects off the air/water interface, with the remaining 0.362 (36.2%) of the wave's power making it through the air-water interface and propagating into water. If the liquid in question is gasoline rather than water (having a rather low relative permittivity value of approximately 2), the power reflection ratio will only be 0.0294 (2.94%) or -15.3 dB, with the vast majority of the wave's power successfully penetrating the air-gasoline interface.

The longer version of the power reflection factor formula suggests liquid-liquid interfaces should be detectable using radar, and indeed they are. All that is needed is a sufficiently large difference in relative permittivity between the two liquids to create a strong enough echo to reliably detect. Liquid-liquid interface level measurement with radar works best when the upper liquid has a substantially lesser permittivity value than the lower liquid[23]. A layer of hydrocarbon oil on top of water (or any aqueous solution such as an acid or a caustic) is a good candidate for guided-wave radar level measurement. An example of a liquid-liquid interface that would be very difficult for a radar instrument to detect is water ($\epsilon_r \approx 80$) above glycerin ($\epsilon_r \approx 42$).

If the radar instrument uses a digital network protocol to communicate information with a host system (such as HART or any number of "fieldbus" standards), it may perform as a multi-variable transmitter, transmitting *both* the interface level measurement and the total liquid level measurement simultaneously. This capability is rather unique to guided-wave radar transmitters, and is very useful in some processes because it eliminates the need for multiple instruments measuring multiple levels.

[23] Rosemount's "Replacing Displacers with Guided Wave Radar" technical note states that the difference in dielectric constant between the upper and lower liquids *must* be at least 10.

One reason why a lesser-ϵ fluid above a greater-ϵ fluid is easier to detect than the inverse is due to the necessity of the signal having to travel through a gas-liquid interface above the liquid-liquid interface. With gases and vapors having such small ϵ values, the signal would have to pass through the gas-liquid interface first in order to reach the liquid-liquid interface. This gas-liquid interface, having the greatest difference in ϵ values of any interface within the vessel, will be *most* reflective to radio energy *in both directions*. Thus, only a small portion of the incident wave will ever reach the liquid-liquid interface, and a similarly small portion of the wave reflected off the liquid-liquid interface (which itself is a fraction of the forward wave power that made it through the gas-liquid interface on its way down) will ever make it through the gas-liquid interface on its way *back up* to the instrument. The situation is much improved if the ϵ values of the two liquid layers are inverted, as shown in this hypothetical comparison (all calculations[24] assume no power dissipation along the way, only reflection at the interfaces):



As you can see in the illustration, the difference in power received back at the instrument is nearly two to one, just from the upper liquid having the lesser of two identical ϵ values. Of course, in real life you do not have the luxury of *choosing* which liquid will go on top of the other (this being determined by fluid density), but you do have the luxury of choosing the appropriate liquid-liquid interface level measurement technology, and as you can see here certain orientations of ϵ values are less detectable with radar than others.

Another factor working against radar as a liquid-liquid interface measurement technology for interfaces where the upper liquid has a greater dielectric constant is that fact that many high-ϵ liquids are aqueous in nature, and water readily dissipates microwave energy. This fact is exploited in microwave ovens, where microwave radiation excites water molecules in the food, dissipating energy in the form of heat. For a radar-based level measurement system consisting of gas/vapor over water over some other (heavier) liquid, the echo signal will be extremely weak because the signal must pass through the "lossy" water layer *twice* before it returns to the radar instrument.

Radio energy losses are important to consider in radar level instrumentation, even when the detected interface is simply gas (or vapor) over liquid. The power reflection factor formula only

[24] $R = 0.5285$ for the 1/40 interface; $R = 0.02944$ for the 40/80 interface; and $R = 0.6382$ for the 1/80 interface.

predicts the ratio of reflected power to incident power *at an interface of substances*. Just because an air-water interface reflects 63.8% of the incident power does not mean 63.8% of the incident power will actually return to the transceiver antenna! Any dissipative losses between the transceiver and the interface(s) of concern will weaken the radio signal, to the point where it may become difficult to distinguish from noise.

Another important factor in maximizing reflected power is the degree to which the microwaves spread out on their way to the liquid interface(s) and back to the transceiver. Guided-wave radar instruments receive a far greater percentage of their transmitted power than non-contact radar instruments because the metal probe used to guide the microwave signal pulses help prevent the waves from spreading (and therefore weakening) throughout the liquids as they propagate. In other words, the probe functions as a transmission line to direct and focus the microwave energy, ensuring a straight path from the instrument into the liquid, and a straight echo return path from the liquid back to the instrument. This is why guided-wave radar is the only practical radar technology for measuring liquid-liquid interfaces.

A critically important factor in accurate level measurement using radar instruments is that the relative permittivity of the upper substance(s) (all media between the radar instrument and the interface of interest) be accurately known. The reason for this is rooted in the dependence of electromagnetic wave propagation velocity to relative permittivity. Recalling the wave velocity formula shown earlier:

$$v = \frac{c}{\sqrt{\epsilon_r}}$$

Where,

v = Velocity of electromagnetic wave through a particular substance
c = Velocity of light in a perfect vacuum ($\approx 3 \times 10^8$ meters per second)
ϵ_r = Relative permittivity (dielectric constant) of the substance

In the case of a single-liquid application where nothing but gas or vapor exists above the liquid, the permittivity of that gas or vapor must be precisely known. In the case of a two-liquid interface with gas or vapor above, the relative permittivities of *both* gas and upper liquids must be accurately known in order to accurately measure the liquid-liquid interface. Changes in dielectric constant value of the medium or media through which the microwaves must travel and echo will cause the microwave radiation to propagate at different velocities. Since all radar measurement is based on time-of-flight through the media separating the radar transceiver from the echo interface, changes in wave velocity through this media will affect the amount of time required for the wave to travel from the transceiver to the echo interface, and reflect back to the transceiver. Therefore, changes in dielectric constant are relevant to the accuracy of any radar level measurement[25].

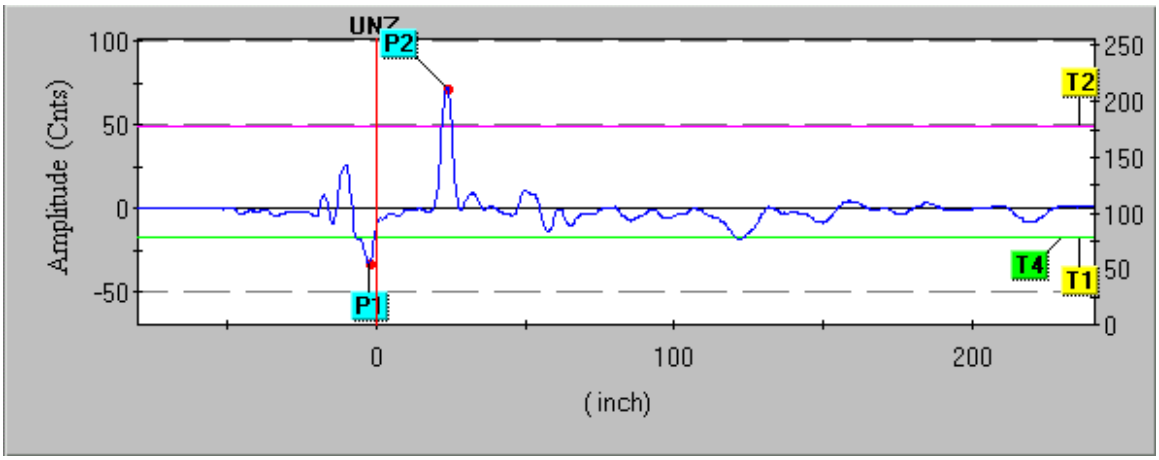
Factors influencing the dielectric constant of gases include pressure and temperature, which means the accuracy of a radar level instrument will vary as gas pressure and/or gas temperature vary! Whether or not this variation is substantial enough to consider for any application depends on the desired measurement accuracy and the degree of permittivity change from one pressure/temperature extreme to another. In no case should a radar instrument be considered for any level measurement

[25] It should be noted that the dielectric constant of the lowest medium (the liquid in a simple, non-interface, level measurement application) is irrelevant for calibration purposes. All we are concerned with is the propagation time of the signal to and from the level of interest, nothing below it.

application unless the dielectric constant value(s) of the upper media are precisely known. This is analogous to the dependence on liquid density that hydrostatic level instruments face. It is futile to attempt level measurement based on hydrostatic pressure if liquid density is unknown, and it is just as futile to attempt level measurement based on radar if the dielectric constants are unknown[26].

As with ultrasonic level instruments, radar level instruments have the ability to measure the level of solid substances in vessels (e.g. powders and granules). The same caveat of repose angle applicable to ultrasonic level measurement (see page 902), however, is a factor for radar measurement as well. When the particulate solid is not very dense (i.e. much air between particles), the dielectric constant may be rather low, making the material surface more difficult to detect.

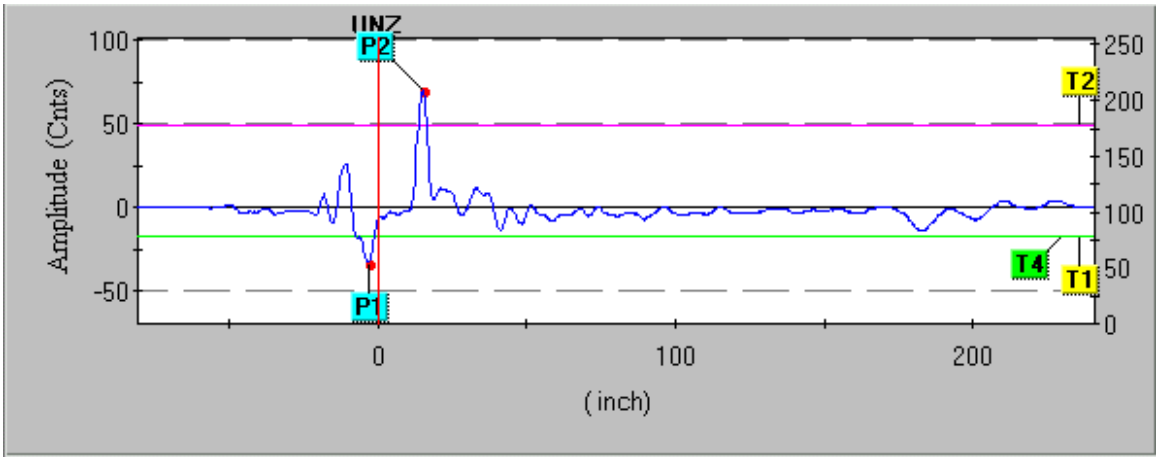
Modern radar level instruments provide a wealth of diagnostic information to aid in troubleshooting. One of the most informative is the *echo curve*, showing each reflected signal received by the instrument along the incident signal's path of travel. The following image is a screen capture of a computer display, from software used to configure a Rosemount model 3301 guided-wave radar level transmitter with a coaxial probe:



Pulse P1 is the *reference* or *fiducial* pulse, resulting from the change in dielectric permittivity between the extended "neck" of the probe (connecting the transmitter to the probe tube) and the coaxial probe itself. This pulse marks the top of the probe, thereby establishing a point of reference for ullage measurement.

[26] For vented-tank level measurement applications where air is the only substance above the point of interest, the relative permittivity is so close to a value of 1 that there is little need for further consideration on this point. Where the relative permittivity of fluids becomes a problem for radar is in high-pressure (non-air) gas applications and liquid-liquid interface applications, especially where the upper substance composition is subject to change.

This next screen capture shows the same level transmitter measuring a water level that is 8 inches higher than before. Note how pulse P2 is further to the left (indicating an echo received sooner in time), indicating a lesser ullage (greater level) measurement:



Several *threshold* settings determine how the transmitter categorizes each received pulse. Threshold T1 for this particular radar instrument defines which pulse is the reference (fiducial). Thus, the first echo in time to exceed the value of threshold T1 is interpreted by the instrument to be the reference point. Threshold T2 defines the upper product level, so the first echo in time to exceed this threshold value is interpreted as the vapor/liquid interface point. Threshold T3 for this particular transmitter is used to define the echo generated by a liquid-liquid interface. However, threshold T3 does not appear in this echo plot because the interface measurement option was disabled during this experiment. The last threshold, T4, defines the end-of-probe detection. Set at a negative value (just like the reference threshold T1), threshold T4 looks for the first pulse in time to exceed that value and interprets that pulse as the one resulting from the signal reaching the probe's end.

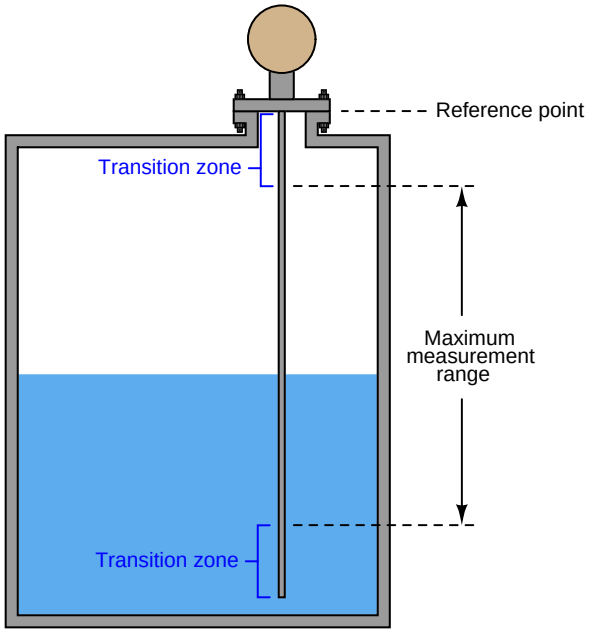
All along the echo curve you can see weak echo signals showing up as bumps. These echoes may be caused by discontinuities along the probe (solid deposits, vent holes, centering spacers, etc.), discontinuities in the process liquid (suspended solids, emulsions, etc.), or even discontinuities in the surrounding process vessel (for non-coaxial probes which exhibit varying degrees of sensitivity to surrounding objects). A challenge in configuring a radar level transmitter is to set the threshold values such that "false" echoes are not interpreted as real liquid or interface levels.

A simple way to eliminate false echoes near the reference point is to set a *null zone* where any echoes are ignored. The upper null zone (UNZ) setting on the Rosemount 3301 radar level transmitter whose screen capture image was shown previously was set to zero, meaning it would be sensitive to any and all echoes near the reference point. If a false echo from a tank nozzle or some other discontinuity near the probe's entry point into the process vessel created a measurement problem, the upper null zone (UNZ) value could be set just beyond that point so the false echo would not be interpreted as a liquid level echo, regardless of the threshold value for T2. A "null zone" is sometimes referred to as a *hold-off distance*.

Some radar level instruments allow thresholds to be set as curves themselves rather than straight lines. Thus, thresholds may be set high during certain periods along the horizontal (time/distance)

axis to ignore false echoes, and set low during other periods to capture legitimate echo signals.

Regardless of how null zones and thresholds are set for any guided-wave radar level transmitter, the technician must be aware of *transition zones* near the extremes of the probe length. Measurements of liquid level or interface level within these zones may not be accurate or even linearly responsive. Thus, it is strongly advised to range the instrument in such a way that the lower- and upper-range values (LRV and URV) lie between the transition zones:



The size of these transition zones depends on both the process substances and the probe type[27]. The instrument manufacturer will provide you with appropriate data for determining transition zone dimensions.

[27] Probe mounting style will also influence the lower transition zone, in the case of flexible probes anchored to the bottom of the process vessel.

19.5.3 Laser level measurement

The least-common form of echo-based level measurement is *laser*, which uses pulses of laser light reflected off the surface of a liquid to detect the liquid level. Perhaps the most limiting factor with laser measurement is the necessity of having a sufficiently reflective surface for the laser light to "echo" off of. Many liquids are not reflective enough for this to be a practical measurement technique, and the presence of dust or thick vapors in the space between the laser and the liquid will disperse the light, weakening the light signal and making the level more difficult to detect.

However, lasers have been applied with great success in measuring distances between objects. Applications of this technology include motion control on large machines, where a laser points at a moving reflector, the laser's electronics calculating distance to the reflector based on the amount of time it takes for the laser "echo" to return. The advent of mass-produced, precision electronics has made this technology practical and affordable for many applications. At the time of this writing (2008), it is even possible for the average American consumer to purchase laser "tape measures" for use in building construction.

19.5.4 Magnetostrictive level measurement

A variation on the theme of echo-based level instruments, where the level of some process material in a vessel is measured by timing the travel of a wave between the instrument and the material interface, is one applied to float-type instruments: *magnetostriction*.

In a magnetostrictive level instrument, liquid level is sensed by a lightweight, donut-shaped float containing a magnet. This float is centered around a long metal rod called a *waveguide*, hung vertically in the process vessel (or hung vertically in a protective cage like the type used for displacement-style level instruments) so that the float may rise and fall with process liquid level. The magnetic field from the float's magnet has an effect on the molecular structure of the metal in the waveguide, such that when an electric current pulse is sent through the rod, a torsional stress pulse[28] is generated at that precise location in the rod where the float magnet's field interacts with the circular magnetic field from the current through the rod. This torsional (twisting) stress travels at the speed of sound through the rod toward either end. At the bottom end is a dampener device designed to absorb the mechanical wave[29]. At the top end of the rod (above the process liquid level) is a sensor and electronics package designed to detect the arrival of the mechanical wave. A precision electronic timing circuit measures the time elapsed between the electric current pulse (called the *interrogation pulse*) and the received mechanical pulse. So long as the speed of sound through the metal waveguide rod remains fixed, the time delay is strictly a function of distance between the float and the sensor, which we already know is called *ullage*.

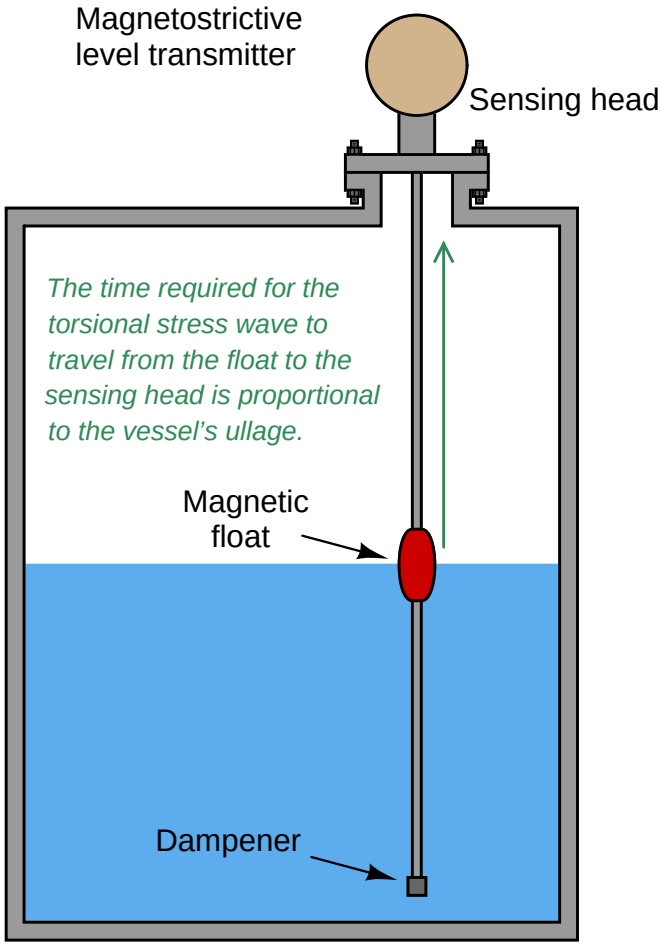
The following photograph (left) and illustration (right) show a magnetostrictive level transmitter[30] propped up against a wall and the same style of transmitter installed in a liquid-

[28] An approximate analogy for understanding the nature of this pulse may be performed using a length of rope. Laying a long piece of rope in a straight line on the ground, pick up one end and quickly move it in a tight circle using a "flip" motion of your wrist. You should be able to see the torsional pulse travel down the length of the rope until it either dies out from dissipation or it reaches the rope's end. Just like the torsional pulse in a magnetostrictive waveguide, this pulse in the rope is mechanical in nature: a movement of the rod's (rope's) molecules. As a mechanical wave, it may be properly understood as a form of *sound*.

[29] This "dampener" is the mechanical equivalent of a *termination resistor* in an electrical transmission line: it makes the traveling wave "think" the waveguide is infinitely long, preventing any reflected pulses. For more information on electrical transmission lines and termination resistors, see section 5.5 beginning on page 255.

[30] This particular transmitter happens to be one of the "M-Series" models manufactured by MTS.

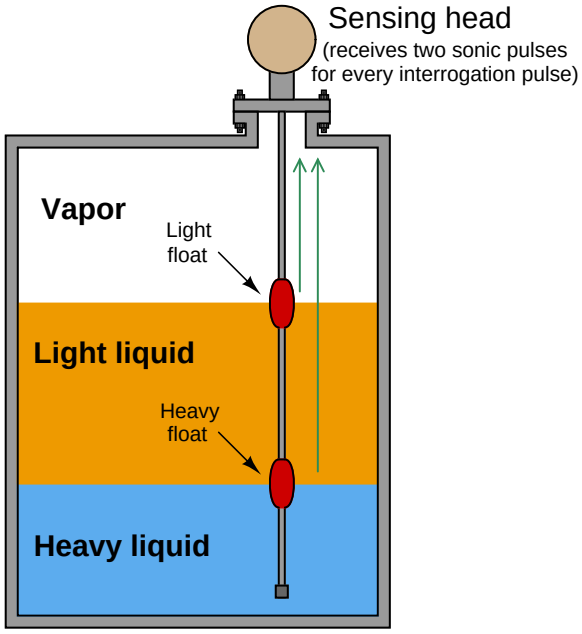
holding vessel, respectively:



The design of this instrument is reminiscent of a guided-wave radar transmitter, where a metal *waveguide* hangs vertically into the process liquid, guiding a pulse to the sensor head where the sensitive electronic components are located. The major difference here is that the pulse we are dealing with is a sonic vibration in the metal of the waveguide rod, not a radio energy (electromagnetic field) pulse as is the case with radar. Like all sound waves, the torsional pulse in a magnetostriction-based level transmitter is much slower-traveling[31] than radio waves.

[31] One reference gives the speed of sound in a magnetostrictive level instrument as 2850 meters per second. Rounding this up to 3×10^3 m/s, we find that the speed of sound in the magnetostrictive waveguide is at least *five orders of magnitude slower* than the speed of light in a vacuum (approximately 3×10^8 m/s). This relative slowness of wave propagation is a good thing for our purposes here, as it gives more time for the electronic timing circuit to count, yielding a more precise measurement of distance traveled by the wave. This fact grants superior resolution of measurement to magnetostrictive level sensors over radar-based and laser-based level sensors. Open-air ultrasonic

It is even possible to measure liquid-liquid interfaces with magnetostrictive instruments. If the waveguide is equipped with a float of such density that it floats on the interface between the two liquids (i.e. the float is denser than the light liquid and less dense than the heavy liquid), the sonic pulse generated in the waveguide by that float's position will represent interface level. Magnetostrictive instruments may even be equipped with two floats: one to sense a liquid-liquid interface, and the other to sense the liquid-vapor interface, so that it may measure both the interface and total levels simultaneously just like a guided-wave radar transmitter:



With such an instrument, each electrical "interrogation" pulse returns *two* sonic pulses to the sensor head: the first pulse representing the total liquid level (upper, light float) and the second pulse representing the interface level (lower, heavy float).

level instruments deal with propagation speeds even slower than this (principally because the density of gases and vapors is far less than that of a solid metal rod) which at first might seem to give these level sensors the upper hand in precision. However, open-air level sensors experience far greater propagation velocity variations caused by changes in pressure and temperature than magnetostrictive sensors. Unlike the speed of sound in gases or liquids, the speed of sound in a solid metal rod is quite stable over a large range of temperatures, and of course is virtually unaffected by the pressure of the surrounding process fluid.

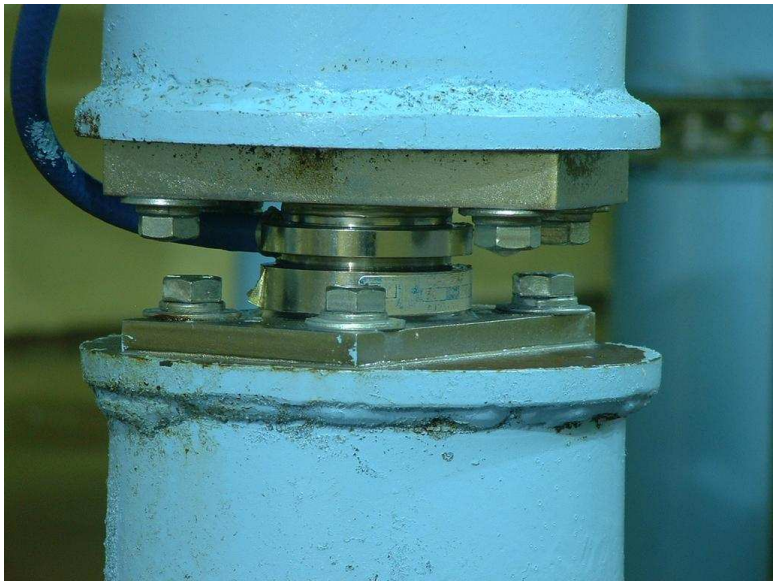
19.6 Weight

Weight-based level instruments sense process level in a vessel by directly measuring the weight of the vessel. If the vessel's empty weight (*tare weight*) is known, process weight becomes a simple calculation of total weight minus tare weight. Obviously, weight-based level sensors can measure both liquid and solid materials, and they have the benefit of providing inherently linear mass storage measurement[32]. *Load cells* (strain gauges bonded to a steel element of precisely known modulus) are typically the primary sensing element of choice for detecting vessel weight. As the vessel's weight changes, the load cells compress or relax on a microscopic scale, causing the strain gauges inside to change resistance. These small changes in electrical resistance become a direct indication of vessel weight.

The following photograph shows three bins used to store powdered milk, each one supported by pillars equipped with load cells near their bases:

[32] Regardless of the vessel's shape or internal structure, the measurement provided by a weight-sensing system is based on the true mass of the stored material. Unlike height-based level measurement technologies (float, ultrasonic, radar, etc.), no characterization will ever be necessary to convert a measurement of height into a measurement of mass.

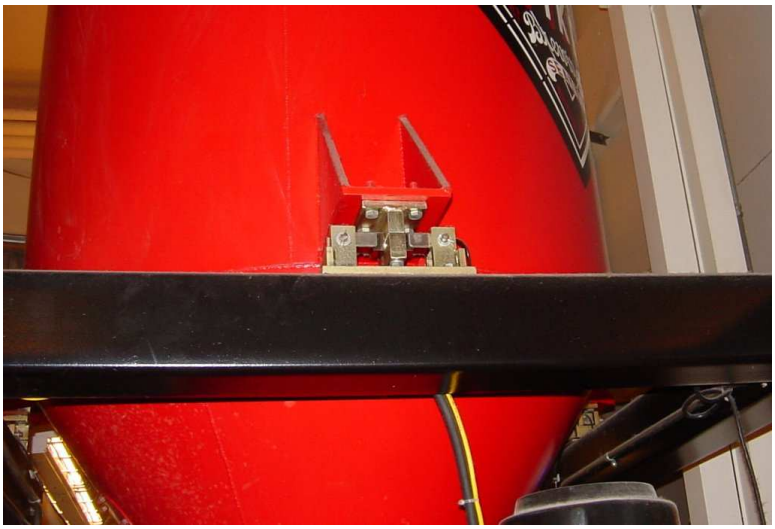
A close-up photograph shows one of the load cell units in detail, near the base of a pillar:



When multiple load cells are used to measure the weight of a storage vessel, the signals from all load cell units must be added together ("summed") to produce a signal representative of the vessel's *total* weight. Simply measuring the weight at one suspension point is insufficient[33], because one can never be sure the vessel's weight is distributed equally amongst all the supports.

[33] If we happened to know, somehow, that the vessel's weight *was* in fact equally shared by all supports, it would be sufficient to simply measure stress at one support to infer total vessel weight. In such an installation, assuming three supports, the total vessel weight would be the stress at any one support multiplied by three.

This next photograph shows a smaller-scale load cell installation used to measure the quantity of material fed into a beer-brewing process[34]:

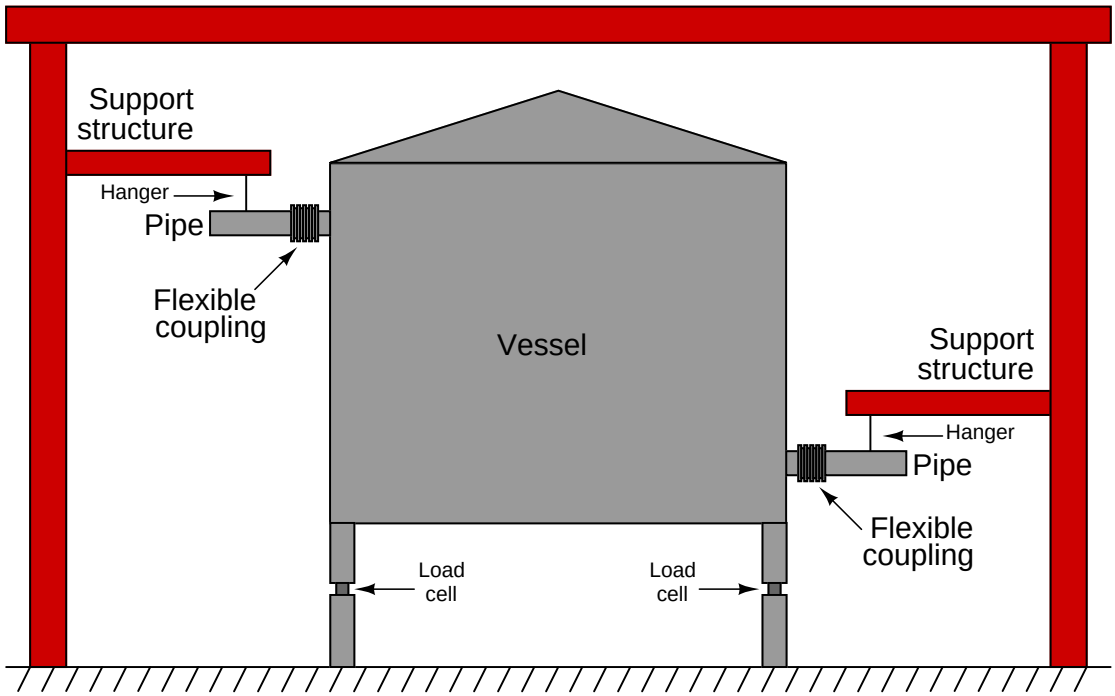


Weight-based measurements are often employed where the true mass of a quantity must be ascertained, rather than the level. So long as the material's density is a known constant, the relationship between weight and level for a vessel of constant cross-sectional area will be linear and predictable. Constant density is not always the case, especially for solid materials, and so weight-based inference of vessel level may be problematic.

In applications where batch mass is more important than height (level), weight-based measurement is often the preferred method for portioning batches. You will find weight-based portion measurements used frequently in the food processing industries (e.g. consistently filling bags and boxes with product), and also for custody transfer of certain materials (e.g. coal and metal ore).

[34] The particular "micro-brewery" process shown here is at the Pike's Place Market in downtown Seattle, Washington. Three load cells measure the weight of a hopper filled with ingredients prior to brewing in the "mash tun" vessel.

One very important caveat for weight-based level instruments is to isolate the vessel from any external mechanical stresses generated by pipes or machinery. The following illustration shows a typical installation for a weight-based measurement system, where all pipes attaching to the vessel do so through flexible couplings, and the weight of the pipes themselves is borne by outside structures through *pipe hangers*:



Stress relief is very important because any forces acting upon the storage vessel will be interpreted by the load cells as more or less material stored in the vessel. The only way to ensure that the load cell's measurement is a direct indication of material held inside the vessel is to ensure that no other forces act upon the vessel except the gravitational weight of the material.

A similar concern for weight-based batch measurement is *vibration* produced by machinery surrounding (or on) the vessel. Vibration is nothing more than oscillatory *acceleration*, and the acceleration of any mass produces a reaction force ($F = ma$). Any vessel suspended by weight-sensing elements such as load cells will induce oscillating forces on those load cells if shaken by vibration. This concern in particular makes it quite difficult to install and operate *agitators* or other rotating machinery on a weighed vessel[35].

An interesting problem associated with load cell measurement of vessel weight arises if there are ever electric currents traveling through the load cell(s). This is not a normal state of affairs,

[35] One practical solution to this problem is to shut down the source of vibration (e.g. agitator motor, pump, etc.) for a long enough time to take a sample weight measurement, then run the machine again between measurements. So long as intermittent weight measurement is adequate for the needs of the process, the interference of machine vibration may be dealt with in this manner.

but it can happen if maintenance workers incorrectly attach arc welding equipment to the support structure of the vessel, or if certain electrical equipment mounted on the vessel such as lights or motors develop ground faults. The electronic amplifier circuits interpreting a load cell's resistance will detect voltage drops created by such currents, interpreting them as changes in load cell resistance and therefore as changes in material level. Sufficiently large currents may even cause permanent damage to load cells, as is often the case when the currents in question are generated by arc welding equipment.

A variation on this theme is the so-called *hydraulic load cell* which is a piston-and-cylinder mechanism designed to translate vessel weight directly into hydraulic (liquid) pressure. A normal pressure transmitter then measures the pressure developed by the load cell and reports it as material weight stored in the vessel. Hydraulic load cells completely bypass the electrical problems associated with resistive load cells, but are more difficult to network for the calculation of total weight (using multiple cells to measure the weight of a large vessel).

19.7 Capacitive

Capacitive level instruments measure electrical capacitance of a conductive rod inserted vertically into a process vessel. As process level increases, capacitance increases between the rod and the vessel walls, causing the instrument to output a greater signal.

The basic principle behind capacitive level instruments is the capacitance equation:

$$C = \frac{\epsilon A}{d}$$

Where,

C = Capacitance

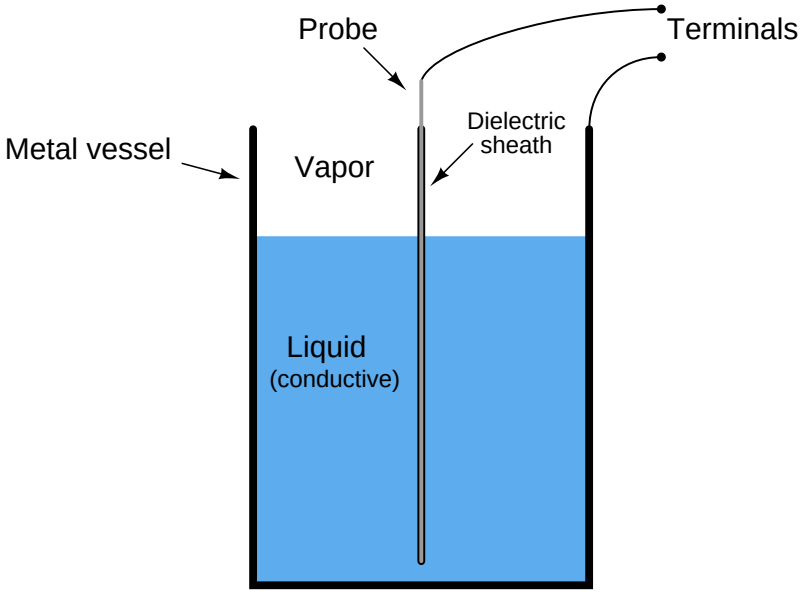
ϵ = Permittivity of dielectric (insulating) material between plates

A = Overlapping area of plates

d = Distance separating plates

The amount of capacitance exhibited between a metal rod inserted into the vessel and the metal walls of that vessel will vary only with changes in permittivity (ϵ), area (A), or distance (d). Since A is constant (the interior surface area of the vessel is fixed, as is the area of the rod once installed), only changes in ϵ or d can affect the probe's capacitance.

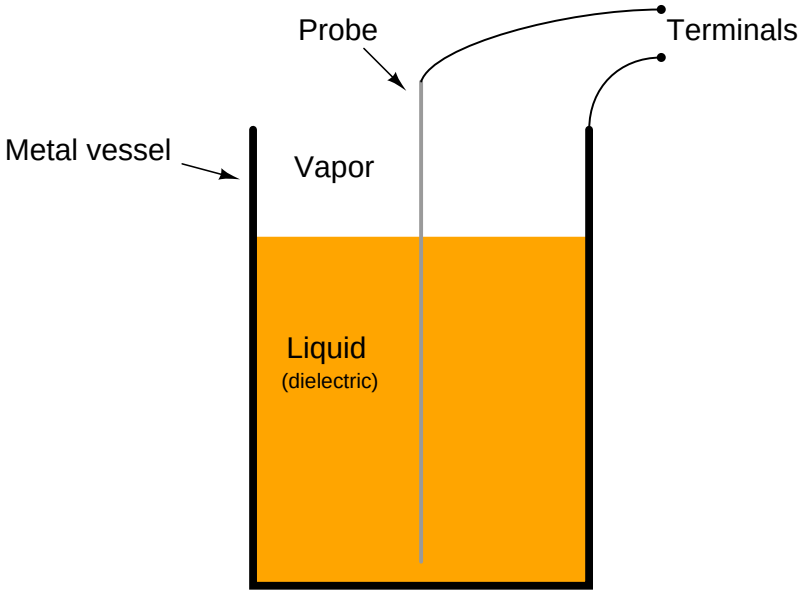
Capacitive level probes come in two basic varieties: one for conductive liquids and one for non-conductive liquids. If the liquid in the vessel is conductive, it cannot be used as the dielectric (insulating) medium of a capacitor. Consequently, capacitive level probes designed for conductive liquids are coated with plastic or some other dielectric substance, so the metal probe forms one plate of the capacitor and the conductive liquid forms the other:



In this style of capacitive level probe, the variables are permittivity (ϵ) and distance (d), since a rising liquid level displaces low-permittivity gas and essentially acts to bring the vessel wall

electrically closer to the probe. This means total capacitance will be greatest when the vessel is full (ϵ is greatest and effective distance d is at a minimum), and least when the vessel is empty (ϵ of the gas is in effect, and over a much greater distance).

If the liquid is non-conductive, it may be used as the dielectric itself, with the metal wall of the storage vessel forming the second capacitor plate:



In this style of capacitive level probe, the variable is permittivity (ϵ), provided the liquid has a substantially greater permittivity than the vapor space above the liquid. This means total capacitance will be greatest when the vessel is full (average permittivity ϵ is at a maximum), and least when the vessel is empty.

Permittivity of the process substance is a critical variable in the non-conductive style of capacitance level probe, and so good accuracy may be obtained with this kind of instrument only if the process permittivity is accurately known. A clever way to ensure good level measurement accuracy when the process permittivity is not stable over time is to equip the instrument with a special *compensating probe* (sometimes called a *composition probe*) below the LRV point in the vessel that will always be submerged in liquid. Since this compensating probe is always immersed, and always experiences the same A and d dimensions, its capacitance is purely a function of the liquid's permittivity (ϵ). This gives the instrument a way to continuously measure process permittivity, which it then uses to calculate level based on the capacitance of the main probe. The inclusion of a compensating probe to measure and compensate for changes in liquid permittivity is analogous to the inclusion of a third pressure transmitter in a hydrostatic *tank expert* system to continuously measure and compensate for liquid density. It is a way to correct for changes in the one remaining system variable that is not related to changes in liquid level.

Capacitive level instruments may be used to measure the level of solids (powders and granules) in addition to liquids. In these applications, and solid substance is almost always non-conductive,

and therefore the permittivity of the substance becomes a factor in measurement accuracy. This can be problematic, as moisture content variations in the solid may greatly affect permittivity, as can variations in granule size. They are not known for great accuracy, though, primarily due to sensitivity to changes in process permittivity and errors caused by stray capacitance in probe cables.

19.8 Radiation

Certain types of nuclear radiation easily penetrates the walls of industrial vessels, but is attenuated by traveling through the bulk of material stored within those vessels. By placing a radioactive source on one side of the vessel and measuring the radiation making it through to the other side of the vessel, an approximate indication of level within that vessel may be obtained. Other types of radiation are *scattered* by process material in vessels, which means the level of process material may be sensed by sending radiation into the vessel through one wall and measuring *back-scattered* radiation returning through the same wall.

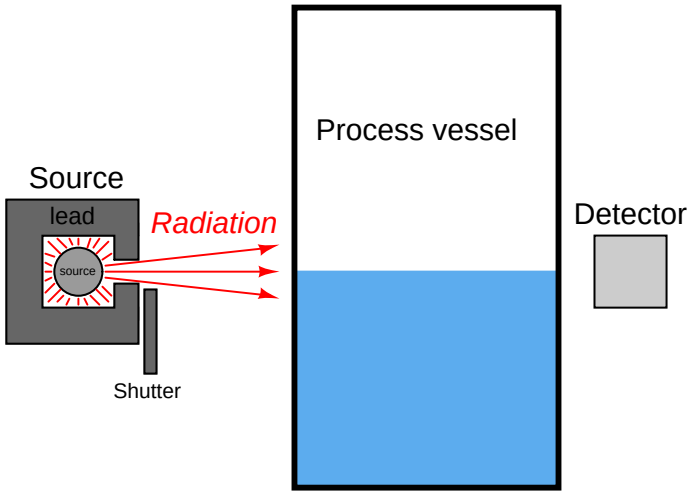
The four most common forms of nuclear radiation are *alpha particles* (α), *beta particles* (β), *gamma rays* (γ), and *neutrons* (n). Alpha particles are helium nuclei (2 protons bound together with 2 neutrons) ejected at high velocity from the nuclei of certain decaying atoms. They are easy to detect, but have very little penetrating power and so are not used for industrial level measurement. Beta particles are electrons[36] ejected at high velocity from the nuclei of certain decaying atoms. Like alpha particles, though, they have little penetrating power and so are not used for industrial level measurement. Gamma rays are electromagnetic in nature (like X-rays and light waves) and have great penetrating power. Neutron radiation also penetrates metal very effectively, but is strongly attenuated and scattered by any substance containing hydrogen (e.g. water, hydrocarbons, and many other industrial fluids), which makes it almost ideal for detecting the presence of a great many process fluids. These latter two forms of radiation (gamma rays and neutrons) are the most common in industrial measurement, with gamma rays used in through-vessel applications and neutrons typically used in backscatter applications.

Nuclear radiation sources consist of radioactive samples contained in a shielded box. The sample itself is a small piece of radioactive substance encased in a double-wall stainless steel cladding, typically resembling a medicinal pill in size and shape. The specific type and quantity of radioactive source material depends on the nature and intensity of radiation required for the application. The basic rule here is that less is better: the smallest source capable of performing the measurement task is the best one for the application.

Common source types for gamma-ray applications are Cesium-137 and Cobalt-60. The numbers represent the *atomic mass* of each isotope: the sum total of protons and neutrons in the nucleus of each atom. These isotopes' nuclei are unstable, decaying over time to become different elements (Barium-137 and Nickel-60, respectively). Cobalt-60 has a relatively short half-life of 5.3 years, whereas Cesium-137 has a much longer half-life of 30 years. This means radiation-based sensors using Cesium will be more stable over time (i.e. less calibration drift) than sensors using Cobalt. The trade-off is that Cobalt emits more powerful gamma rays than Cesium, which makes it better suited to applications where the radiation must penetrate thick process vessels or travel long distances (across wide process vessels).

[36] Beta particles are *not* orbital electrons, but rather than product of elementary particle decay in an atom's nucleus. These electrons are spontaneously generated and subsequently ejected from the nucleus of the atom.

One of the most effective methods of shielding against gamma ray radiation is with very dense substances such as lead or concrete. This is why the source boxes holding gamma-emitting radioactive pellets are lined with lead, so the radiation escapes only in the direction intended:



These "sources" may be locked out for testing and maintenance by moving a lever that hinges a lead shutter over the "window" of the box. This lead shutter acts as an on/off switch for the radioactive source. The lever actuating the shutter typically has provisions for lock-out/tag-out so a maintenance person may place a padlock on the lever and prevent anyone else from "turning on" the source during maintenance. For point-level (level switch) applications, the source shutter acts as a simple simulator for either a full vessel (in the case of a through-vessel installation) or an empty vessel (in the case of a backscatter installation). A full vessel may be simulated for neutron backscatter instruments by placing a sheet of plastic (or other hydrogen-rich substance) between the source box and the process vessel wall.

The detector for a radiation-based instrument is by far the most complex and expensive component of the system. Many different detector designs exist, the most common at the time of this writing being *ionization tubes* such as the Geiger-Muller (G-M) tube. In such devices, a thin metal wire centered in a metal cylinder sealed and filled with inert gas is energized with high voltage DC. Any ionizing radiation such as alpha, beta, or gamma radiation entering the tube causes gas molecules to ionize, allowing a pulse of electric current to travel between the wire and tube wall. A sensitive electronic circuit detects and counts these pulses, with a greater pulse rate corresponding to a greater intensity of detected radiation.

Neutron radiation is notoriously difficult to electronically detect, since neutrons are non-ionizing. Ionization tubes specifically made for neutron radiation detection do exist, and are filled with special substances known to react with neutron radiation. One example of such a detector is the so-called *fission chamber*, which is an ionization chamber lined with a fissile material such as uranium-235 (^{235}U). When a neutron enters the chamber and is captured by a fissile nucleus, that nucleus undergoes fission (splits into separate pieces) with a subsequent emission of gamma rays and charged particles, which are then detected by ionization in the chamber. Another variation on this theme is to fill an ionization tube with boron trifluoride gas. When a boron-10 (^{10}B) nucleus captures a

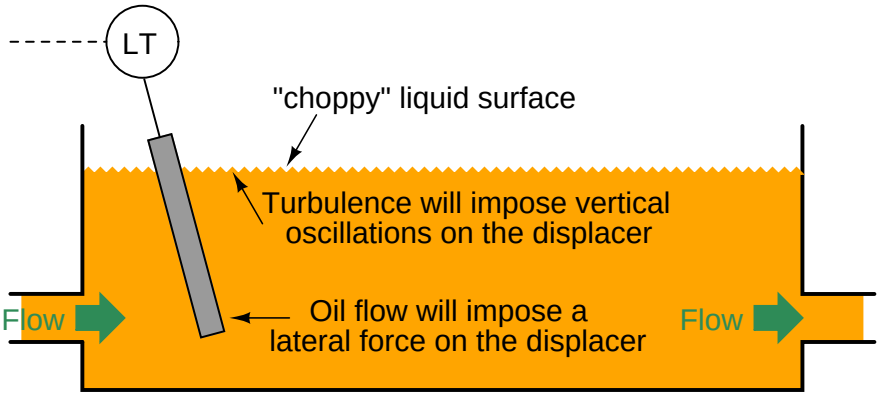
neutron, it transmutates into lithium-7 (^{7}Li) and ejects an alpha particle and several beta particles, both of which cause detectable ionization in the chamber.

The accuracy of radiation-based level instruments varies with the stability of process fluid density, vessel wall coating, source decay rates, and detector drift. Given these error variables and the additional need for NRC (Nuclear Regulatory Commission) licensing to operate such instruments at an industrial facility, radiation instruments are typically used where no other instrument is practical. Examples include the level measurement of highly corrosive or toxic process fluids where penetrations into the vessel must be minimized and where piping requirements make weight-based measurement impractical (e.g. hydrocarbon/acid separators in alkylation processes in the oil refining industry), as well as processes where the internal conditions of the vessel are too physically violent for any instrument to survive (e.g. delayed coking vessels in the oil refining industry, where the coke is "drilled" out of the vessel by a high-pressure water jet).

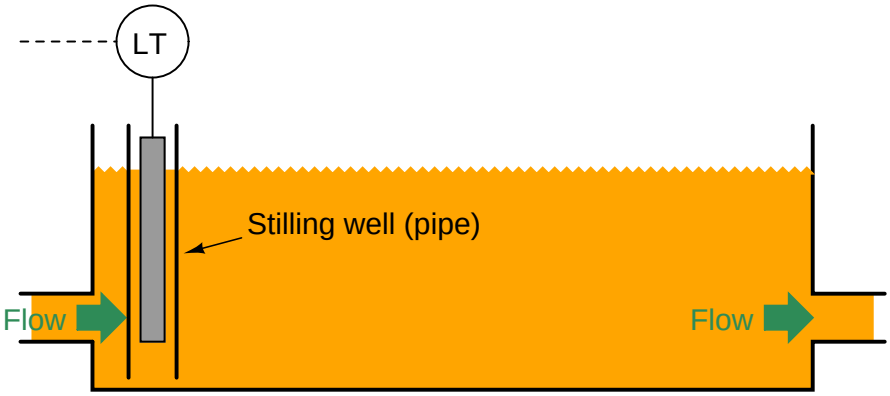
19.9 Level sensor accessories

Disturbances in the liquid tend to complicate liquid level measurement. These disturbances may result from liquid introduced into a vessel above the liquid level (splashing into the liquid's surface), the rotation of agitator paddles, and/or turbulent flows from mixing pumps. Any source of turbulence for the liquid surface (or liquid-liquid interface) is especially problematic for echo-type level sensors, which *only* sense interfaces between vapors and liquids, or liquids and liquids.

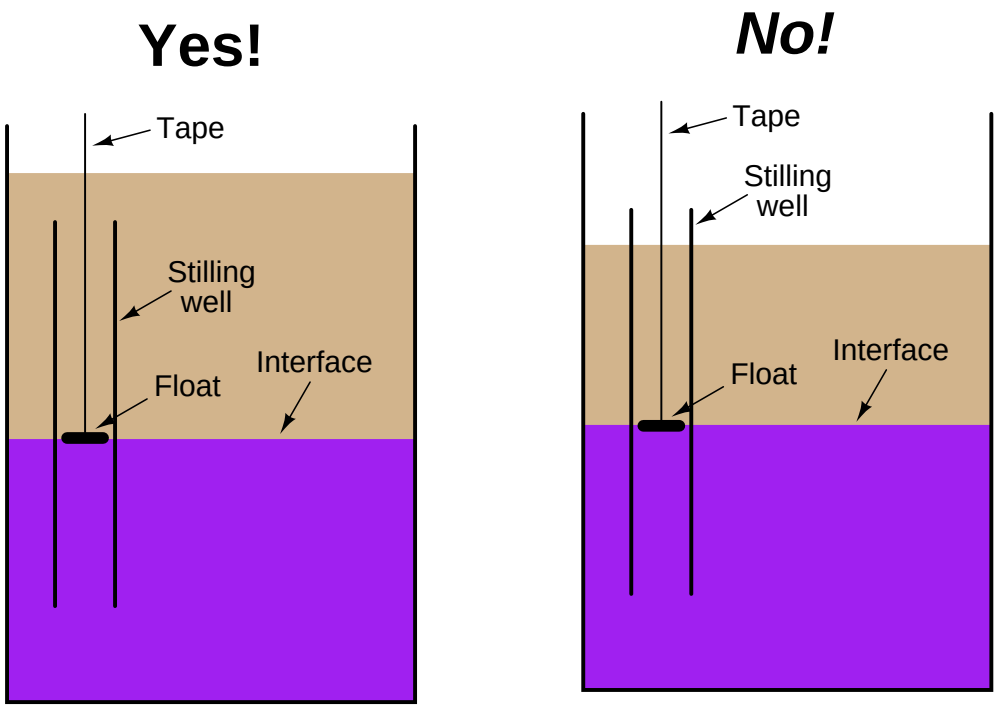
If it is not possible to eliminate disturbances inside the process vessel, a relatively simple accessory one may add to the process vessel is a vertical length of pipe called a *stilling well*. To understand the principle of a stilling well, first consider the application of a hydraulic oil reservoir where we wish to continuously measure oil level. The oil flow in and out of this reservoir will cause problems for the displacer element:



A section of vertical pipe installed in the reservoir around the displacer will serve as a shield to all the turbulence in the rest of the reservoir. The displacer element will no longer be subject to a horizontal blast of oil entering the reservoir, nor any wave action to make it bob up and down. This section of pipe *quiets*, or *stills*, the oil surrounding the displacer, making it easier to measure oil level:

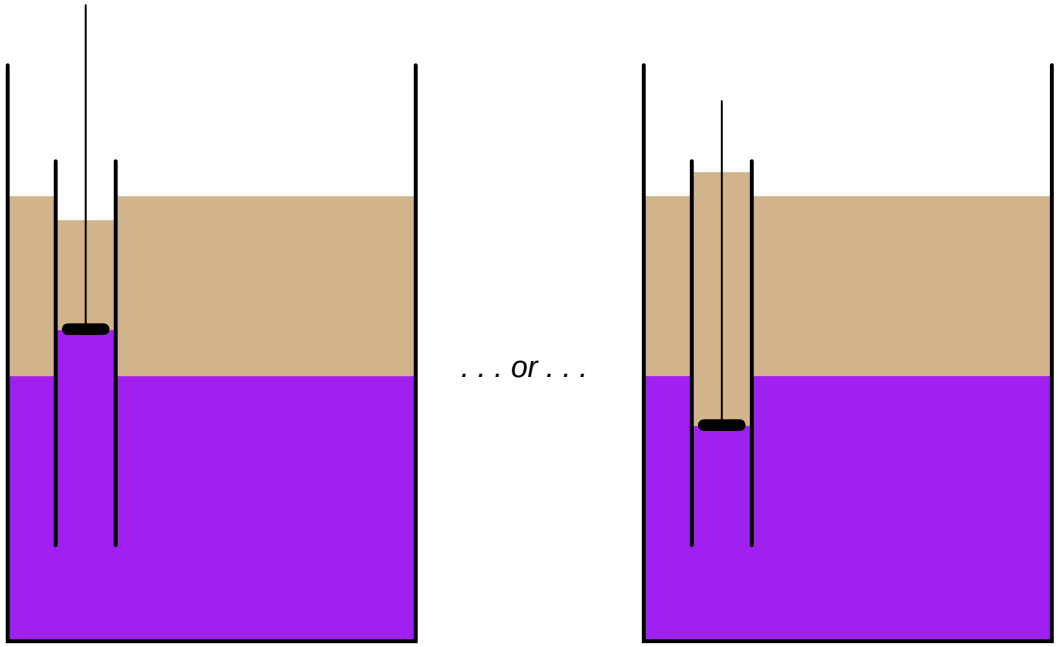


Stilling wells may be used in conjunction with many types of level instruments: floats, displacers, ultrasonic, radar, and laser to name a few. If the process application necessitates liquid-liquid interface measurement, however, the stilling well must be properly installed to ensure the interface level inside the well match the interface levels in the rest of the vessel. Consider this example of using a stilling well in conjunction with a tape-and-float system for interface measurement:



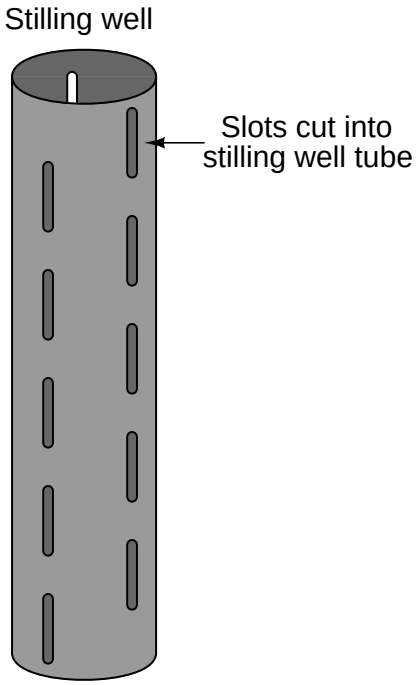
In the left-hand installation where the stilling well is completely submerged, the interface levels will always match. In the right-hand installation where the top of the stilling well extends above the total liquid level, however, the two levels may not always match.

This potential problem for the non-submerged stilling well is graphically illustrated here:



The problem here is analogous to what we see with sightglass-style level gauges: interfaces may be reliably indicated if and only if both ends of the sightglass are submerged (see page 849 for an illustrated description of the problem).

If it is not possible or practical to ensure complete submersion of the stilling well, an alternative technique is to drill holes or cut slots in the well to allow interface levels to equalize inside and outside of the well tube:



Such equalization ports are commonly found as a standard design feature on coaxial probes for guided-wave radar level transmitters, where the outer tube of the coaxial transmission line acts as a sort of stilling well for the fluid. Coaxial probes are typically chosen for liquid-liquid interface radar measurement applications because they do the best job of preventing dispersion of the radio energy[37], but the "stilling well" property of a coaxial probe practically necessitates these equalization ports to ensure the interface level within the probe always matches the interface level in the rest of the vessel.

[37] So much of the incident power is lost as the radar signal partially reflects off the gas-liquid interface, then the liquid-liquid interface, then *again* through the gas-liquid interface on its return trip to the instrument that every care must be taken to ensure optimum received signal strength. While twin-lead probes have been applied in liquid-liquid interface measurement service, the coaxial probe design is still the best for maintaining radar signal integrity.

References

"Autolevel" Application Note AN 01C22A01-01E, Yokogawa Electric Corporation, 2006.

"Boiler Drum Level Transmitter Calibration", application data sheet 00800-0100-3055, Rosemount, Inc., Chanhassen, MN, 2001.

Brumbi, Detlef, *Fundamentals of Radar Technology for Level Gauging*, 4th Edition, Krohne Messtechnik GmbH & Co. KG, Duisburg, Germany, 2003.

"Bubble Tube Installations For Liquid Level, Density, and Interface Measurements", document MI 020-328, The Foxboro Company, Foxboro, MA, 1988.

"DOE Fundamentals Handbook, Instrumentation and Control, Volume 2 of 2", document DOE-HDBK-1013/2-92, U.S. Department of Energy, Washington, D.C., 1992.

Fribance, Austin E., *Industrial Instrumentation Fundamentals*, McGraw-Hill Book Company, New York, NY, 1962.

Kallen, Howard P., *Handbook of Instrumentation and Controls*, McGraw-Hill Book Company, Inc., New York, NY, 1961.

"Level Measurement Technology: Radar", document 00816-0100-3209, revision AA, Rosemount, Inc., Chanhassen, MN, 1999.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Measurement and Analysis Volume I*, Fourth Edition, CRC Press, New York, NY, 2003.

MacBeth, Michael, *IAEA CANDU Instrumentation & Control Course*, SNERDI, Shanghai, 1998.

"Model 1151 Alphaline Pressure Transmitters", product manual 00809-0100-4360, revision AA, Rosemount, Inc., Chanhassen, MN, 1997.

"Replacing Displacers with Guided Wave Radar", technical note 3300_2_02_CA, Rosemount, Inc., Chanhassen, MN, 2003.

"The Art of Tank Gauging For Safety And Precision", IN 4416.650, revision 6, Enraf B.V., The Netherlands.

Chapter 20

Continuous temperature measurement

Temperature is the measure of average molecular kinetic energy within a substance. The concept is easiest to understand for gases under low pressure, where gas molecules randomly shuffle about. The average kinetic (motional) energy of these gas molecules defines temperature for that quantity of gas. There is even a formula expressing the relationship between average kinetic energy ($\overline{E_k}$) and temperature (T) for a monatomic (single-atom molecule) gas:

$$\overline{E_k} = \frac{3kT}{2}$$

Where,

$\overline{E_k}$ = Average kinetic energy of the gas molecules (joules)
k = Boltzmann's constant (1.38×10^{-23} joules/Kelvin)
T = Absolute temperature of gas (Kelvin)

Thermal energy is a different concept: the quantity of *total kinetic energy* for this random molecular motion. If the average kinetic energy is defined as $\frac{3kT}{2}$, then the total kinetic energy for all the molecules in a monatomic gas must be this quantity times the total number of molecules (N) in the gas sample:

$$E_{\text{thermal}} = \frac{3NkT}{2}$$

This may be equivalently expressed in terms of the number of *moles* of gas rather than the number of molecules (a staggeringly large number for any realistic sample):

$$E_{\text{thermal}} = \frac{3nRT}{2}$$

Where,

E_{thermal} = Total thermal energy for a gas sample (joules)
n = Quantity of gas in the sample (moles)
R = Ideal gas constant (8.315 joules per mole-Kelvin)
T = Absolute temperature of gas (Kelvin)

Heat is defined as the exchange of thermal energy from one sample to another, by way of conduction (direct contact), convection (transfer via a moving fluid), or radiation (emitted energy); although you will often find the terms *thermal energy* and *heat* used interchangeably.

Temperature is a more easily detected quantity than heat. There are many different ways to measure temperature, from a simple glass-bulb mercury thermometer to sophisticated infrared optical sensor systems. Like all other areas of measurement, there is no single technology that is best for all applications. Each temperature-measurement technique has its own strengths and weaknesses. One responsibility of the instrument technician is to know these pros and cons so as to choose the best technology for the application, and this knowledge is best obtained through understanding the operational principles of each technology.

20.1 Bi-metal temperature sensors

Solids tend to expand when heated. The amount that a solid sample will expand with increased temperature depends on the size of the sample, the material it is made of, and the amount of temperature rise. The following formula relates linear expansion to temperature change:

$$l = l_0(1 + \alpha\Delta T)$$

Where,
l = Length of material after heating
l_0 = Original length of material
α = Coefficient of linear expansion
ΔT = Change in temperature

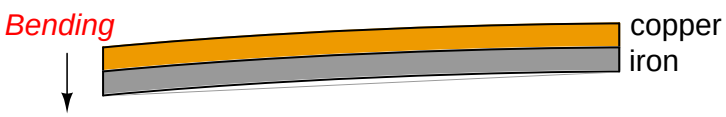
Here are some typical values of α for common metals:

- Aluminum = 25×10^{-6} per degree C
- Copper = 16.6×10^{-6} per degree C
- Iron = 12×10^{-6} per degree C
- Tin = 20×10^{-6} per degree C
- Titanium = 8.5×10^{-6} per degree C

As you can see, the values for α are quite small. This means the amount of expansion (or contraction) for modest temperature changes are almost too small to see unless the sample size (l_0) is huge. We can readily see the effects of thermal expansion in structures such as bridges, where expansion joints must be incorporated into the design to prevent serious problems due to changes in ambient temperature. However, for a sample the size of your hand the change in length from a cold day to a warm day will be microscopic.

One way to amplify the motion resulting from thermal expansion is to bond two strips of dissimilar metals together, such as copper and iron. If we were to take two equally-sized strips of copper and iron, lay them side-by-side, and then heat both of them to a higher temperature, we would see the copper strip lengthen slightly more than the iron strip:

If we bond these two strips of metal together, this differential growth will result in a bending motion that greatly exceeds the linear expansion. This device is called a *bi-metal strip*:



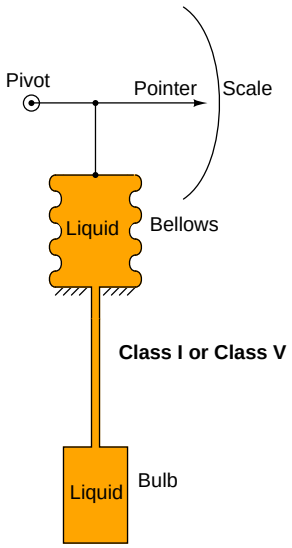
This bending motion is significant enough to drive a pointer mechanism, activate an electromechanical switch, or perform any number of other mechanical tasks, making this a very simple and useful *primary sensing element* for temperature.

If a bi-metallic strip is twisted over a long length, it will tend to un-twist as it heats up. This twisting motion may be used to directly drive the needle of a temperature gauge. This is the operating principle of the temperature gauge shown in the following photograph:

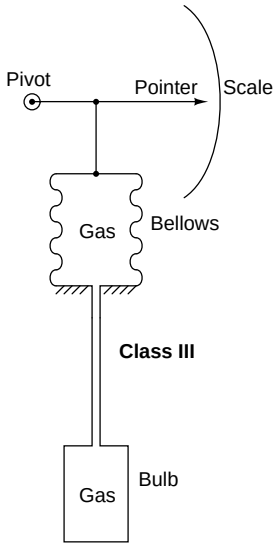
20.2 Filled-bulb temperature sensors

Filled-bulb systems exploit the principle of fluid expansion to measure temperature. If a fluid is enclosed in a sealed system and then heated, the molecules in that fluid will exert a greater pressure on the walls of the enclosing vessel. By measuring this pressure, and/or by allowing the fluid to expand under constant pressure, we may infer the temperature of the fluid.

Class I and Class V systems use a liquid fill fluid (class V is mercury). Here, the volumetric expansion of the liquid drives an indicating mechanism to show temperature:

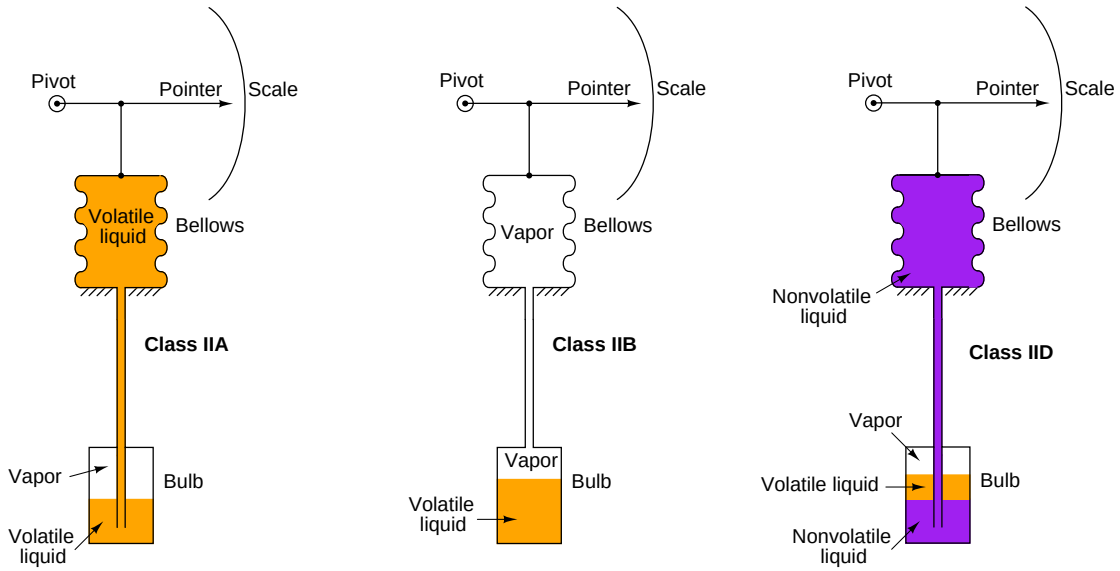


Class III systems use a gas fill fluid instead of liquid. Here, the change in pressure with temperature (as described by the Ideal Gas Law) allows us to sense the bulb's temperature:



In these systems, it is quite critical that the tube connecting the sensing bulb to the indicating element be of minimal volume, so the fluid expansion is primarily due to changes in temperature at the bulb rather than changes in temperature along the length of the tube. It is also important to realize that the fluid volume contained by the bellows (or bourdon tube or diaphragm . . .) is also subject to expansion and contraction due to temperature changes at the indicator. This means the temperature indication varies somewhat as the indicator temperature changes, which is not desirable, since we intend the device to measure temperature (exclusively) at the bulb. Various methods of compensation exist for this effect (for example, a bi-metal spring inside the indicator mechanism to automatically offset the indication as ambient temperature changes), but it may be permanently offset through a simple "zero" adjustment provided that the ambient temperature at the indicator does not change much.

A fundamentally different class of filled-bulb system is the Class II, which uses a volatile liquid/vapor combination to generate a temperature-dependent fluid expansion:

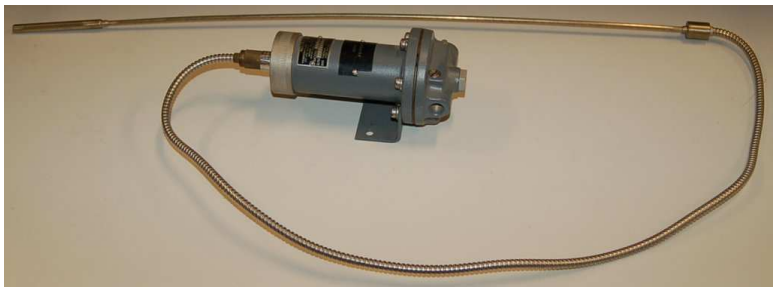


Given that the liquid and vapor are in direct contact with each other, the pressure in the system will be precisely equal to the *saturated vapor pressure* at the vapor/liquid interface. This makes the Class II system sensitive to temperature only at the bulb and nowhere else along the system's volume. Because of this phenomenon, a Class II filled-bulb system requires no compensation for temperature changes at the indicator.

Class II systems do have one notable idiosyncrasy, though: they have a tendency to switch from Class IIA to Class IIB when the temperature of the sensing bulb crosses the ambient temperature at the indicator. Simply put, the liquid tends to seek the colder portion of a Class II system while the vapor tends to seek the warmer portion. This causes problems when the indicator and sensing bulb exchange identities as warmer/colder. The rush of liquid up (or down) the capillary tubing as the system tries to reach a new equilibrium causes intermittent measurement errors. Class II filled-bulb systems designed to operate in either IIA or IIB mode are classified as *IIC*.

One calibration problem common to all systems with liquid-filled capillary tubes is an offset in temperature measurement due to hydrostatic pressure (or suction) resulting from a different in height between the measurement bulb and the indicator. This represents a "zero" shift in calibration, which may be permanently offset by a "zero" adjustment at the time of installation. Class III (gas-filled) and Class IIB (vapor-filled) systems, of course, suffer no such problem because there is no liquid in the capillary tube to generate a pressure due to height.

A photograph of a pneumatic temperature transmitter using a filled-bulb as the sensing element appears here:

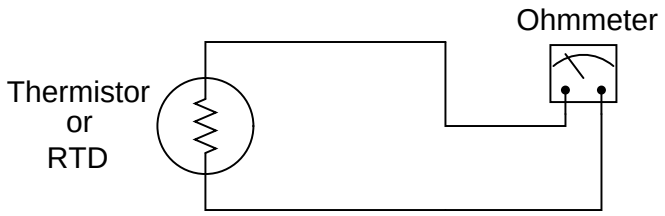


This transmitter happens to be a Moore Products "Nullmatic" model. The capillary tube connecting the fluid-filled bulb to the transmitter mechanism is protected by a spiral-metal jacket. The bulb itself is located at the very end of the stainless steel "wand" which inserts into the process fluid to be measured:

Instead of directly actuating a pointer mechanism, the fluid pressure in this instrument actuates a self-balancing pneumatic mechanism to produce a 3 to 15 PSI air pressure signal representing process temperature.

20.3 Thermistors and Resistance Temperature Detectors (RTDs)

One of the simplest classes of temperature sensor is one where temperature effects a change in electrical resistance. With this type of primary sensing element, a simple ohmmeter is able to function as a thermometer, interpreting the resistance as a temperature measurement:



Thermistors are devices made of metal oxide which either increase in resistance with increasing temperature (a *positive temperature coefficient*) or decrease in resistance with increasing temperature (a *negative temperature coefficient*). *RTD*s are devices made of pure metal (usually platinum or copper) which always increase in resistance with increasing temperature. The major difference between thermistors and RTDs is linearity: thermistors are highly sensitive and nonlinear, whereas RTDs are relatively insensitive but very linear. For this reason, thermistors are typically used where high accuracy is unimportant. Many consumer-grade devices use thermistors for temperature sensors.

20.3.1 Temperature coefficient of resistance (α)

Resistive Temperature Detectors (RTDs) relate resistance to temperature by the following formula:

$$R_T = R_{ref}[1 + \alpha(T - T_{ref})]$$

Where,
R_T = Resistance of RTD at given temperature T (ohms)
R_{ref} = Resistance of RTD at the reference temperature T_{ref} (ohms)
α = Temperature coefficient of resistance (ohms per ohm/degree)

The following example shows how to use this formula to calculate the resistance of a "100 ohm" platinum RTD with a temperature coefficient value of 0.00392 at a temperature of 35 degrees Celsius:

$$R_T = 100\,\Omega[1 + (0.00392)(35^o\text{ C} - 0^o\text{ C})]$$

$$R_T = 100\,\Omega[1 + 0.1372]$$

$$R_T = 100\,\Omega[1.1372]$$

$$R_T = 113.72\,\Omega$$

Due to nonlinearities in the RTD's behavior, this linear RTD formula is only an approximation. A better approximation is the *Callendar-van Dusen formula*, which introduces second, third, and fourth-degree terms for a better fit: $R_T = R_{ref}(1 + AT + BT^2 - 100CT^3 + CT^4)$ for temperatures ranging -200^o C $< T <$ 0^o C and $R_T = R_{ref}(1 + AT + BT^2)$ for temperatures ranging 0^o C $< T <$ 661^o C, both assuming $T_{ref} = 0^o$ C.

Water's melting/freezing point is the standard reference temperature for most RTDs. Here are some typical values of α for common metals:

- Nickel = 0.00672 Ω/Ω^oC
- Tungsten = 0.0045 Ω/Ω^oC
- Silver = 0.0041 Ω/Ω^oC
- Gold = 0.0040 Ω/Ω^oC
- Platinum = 0.00392 Ω/Ω^oC
- Copper = 0.0038 Ω/Ω^oC

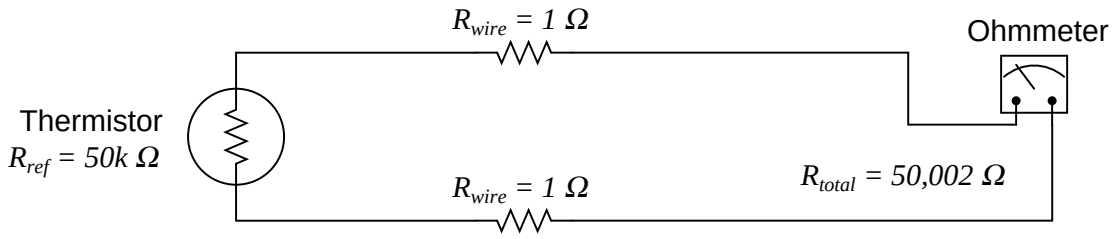
As mentioned previously, platinum is a common wire material for industrial RTD construction. The alpha (α) value for platinum varies according to the alloying of the metal. For "reference grade" platinum wire, the most common alpha value is 0.003902. Industrial-grade platinum alloy RTD wire is commonly available in two different coefficient values: 0.00385 (the "European" alpha value) and

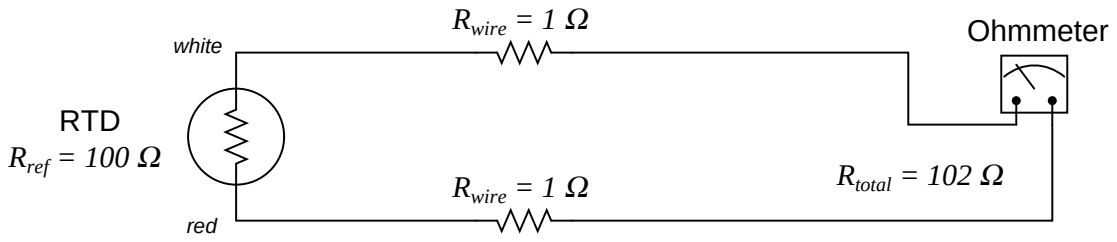
0.00392 (the "American" alpha value), of which the "European" value of 0.00385 is most commonly used (even in the United States!).

100 Ω is a very common reference resistance (R_{ref} at 0 degrees Celsius) for industrial RTDs. 1000 Ω is another common reference resistance, and some industrial RTDs have reference resistances as low as 10 Ω. Compared to thermistors with their tens or even hundreds of thousands of ohms' nominal resistance, an RTD's resistance is comparatively small. This can cause problems with measurement, since the wires connecting an RTD to its ohmmeter possess their own resistance, which will be a more substantial percentage of the total circuit resistance than for a thermistor.

20.3.2 Two-wire RTD circuits

The following schematic diagrams show the relative effects of 2 ohms total wire resistance on a thermistor circuit and on an RTD circuit:





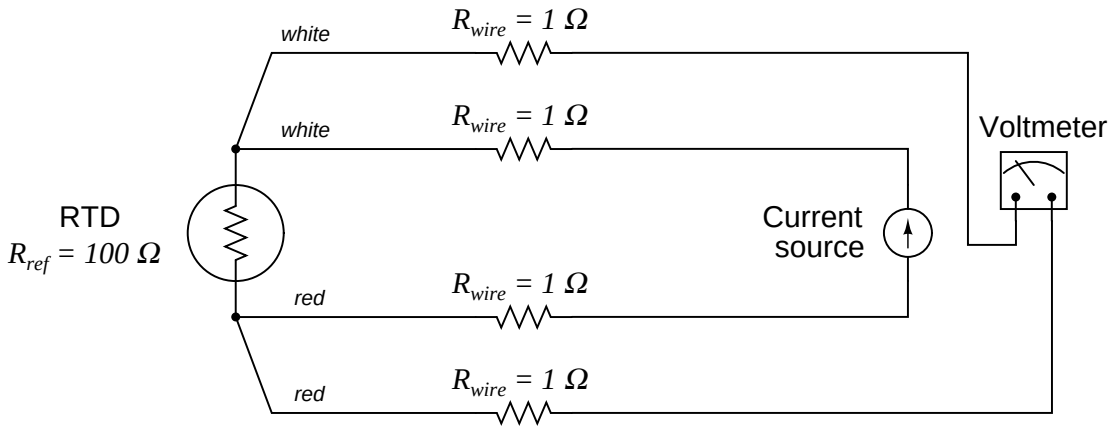
Clearly, wire resistance is more problematic for low-resistance RTDs than for high-resistance thermistors. In the RTD circuit, wire resistance counts for 1.96% of the total circuit resistance. In the thermistor circuit, the same 2 ohms of wire resistance counts for only 0.004% of the total circuit resistance. The thermistor's huge reference resistance value "swamps"[1] the wire resistance to the point that the latter becomes insignificant by comparison.

In HVAC (Heating, Ventilation, and Air Conditioning) systems, where the temperature measurement range is relatively narrow, the nonlinearity of thermistors is not a serious concern and their relative immunity to wire resistance error is a definite advantage over RTDs. In industrial temperature measurement applications where the temperature ranges are usually much wider, the nonlinearity of thermistors becomes a significant problem, so we must find a way to use low-resistance RTDs and deal with the (lesser) problem of wire resistance.

[1] "Swamping" is the term given to the overshadowing of one effect by another. Here, the normal resistance of the high-value RTD greatly overshadows any wire resistance, such that wire resistance becomes negligible.

20.3.3 Four-wire RTD circuits

A very old electrical technique known as the *Kelvin* or *four-wire* method is a practical solution for this problem. Commonly employed to make precise resistance measurements for scientific experiments in laboratory conditions, the four-wire technique uses four wires to connect the resistance under test (in this case, the RTD) to the measuring instrument:



Current is supplied to the RTD from a current source, whose job it is to precisely regulate current regardless of circuit resistance. A voltmeter measures the voltage dropped across the RTD, and Ohm's Law is used to calculate the resistance of the RTD ($R = \frac{V}{I}$).

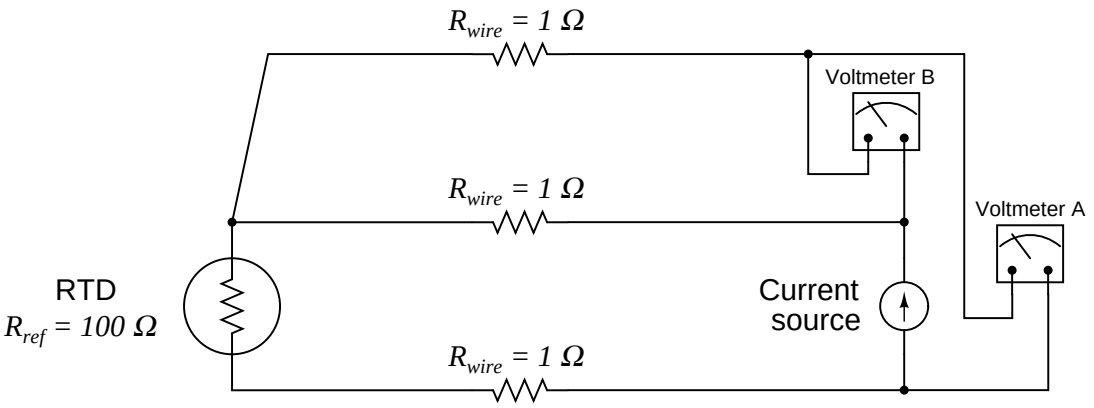
None of the wire resistances are consequential in this circuit. The two wires carrying current to the RTD will drop some voltage along their length, but this is of no concern because the voltmeter only "sees" the voltage dropped across the RTD rather than the voltage drop across the current source. While the two wires connecting the voltmeter to the RTD do have resistance, they drop negligible voltage because the voltmeter draws so little current through them (remember an ideal voltmeters has infinite input impedance, and modern semiconductor-amplified voltmeters have impedances of several mega-ohms or more). Thus, the resistances of the current-carrying wires are of no effect because the voltmeter never senses their voltage drops, and the resistances of the voltmeter's sensing wires are of no effect because they carry practically zero current.

Note how wire colors (*white* and *red*) are used to indicate which wires are common pairs at the RTD. Often, these wire colors will be the technician's only guide to properly connecting a 4-wire RTD to a sensing instrument.

The only disadvantage of the four-wire method is the sheer number of wires necessary. Four wires *per RTD* can add up to a sizeable wire count when many different RTDs are installed in a process area. Wires cost money, and occupy expensive conduit, so there are situations where the four-wire method is a burden.

20.3.4 Three-wire RTD circuits

A compromise between two-wire and four-wire RTD connections is the *three-wire* connection, which looks like this:



In a three-wire RTD circuit, voltmeter "A" measures the voltage dropped across the RTD (plus the voltage dropped across the bottom current-carrying wire). Voltmeter "B" measures just the voltage dropped across the top current-carrying wire. Assuming both current-carrying wires will have (very nearly) the same resistance, subtracting the indication of voltmeter "B" from the indication given by voltmeter "A" yields the voltage dropped across the RTD:

$$V_{RTD} = V_{\text{meter(A)}} - V_{\text{meter(B)}}$$

If the resistances of the two current-carrying wires are precisely identical (and this includes the electrical resistance of any connections within those current-carrying paths, such as terminal blocks), the calculated RTD voltage will be the same as the true RTD voltage, and no wire-resistance error will appear. If, however, one of those current-carrying wires happens to exhibit more resistance than the other, the calculated RTD voltage will not be the same as the actual RTD voltage, and a measurement error will result.

Thus, we see that the three-wire RTD circuit saves us wire cost over a four-wire circuit, but at the "expense" of a potential measurement error. The beauty of the four-wire design was that wire resistances were completely irrelevant: a true determination of RTD voltage (and therefore RTD resistance) could be made regardless of how much resistance each wire had, or even if the wire resistances were different from each other. The error-canceling property of the three-wire circuit, by contrast, hinges on the assumption that the two current-carrying wires have exactly the same resistance, which may or may not actually be true.

It should be understood that real three-wire RTD instruments do not employ direct-indicating voltmeters. Actual RTD instruments use either analog or digital "conditioning" circuits to measure the voltage drops and perform the necessary calculations to compensate for wire resistance. The voltmeters shown in the four-wire and three-wire diagrams serve only to illustrate the basic concepts, not to showcase a practical instrument design.

A photograph of a modern temperature transmitter capable of receiving input from 2-wire, 3-wire, or 4-wire RTDs (as well as thermocouples, another type of temperature sensor entirely) shows the connection points and the labeling:

The rectangle symbol shown on the label represents the resistive element of the RTD. The symbol with the "+" and "−" marks represents a thermocouple junction, and may be ignored for the purposes of this discussion. As shown by the diagram, a two-wire RTD would connect between terminals 2 and 3. Likewise, a three-wire RTD would connect to terminals 1, 2, and 3 (with terminals 1 and 2 being the points of connection for the two common wires of the RTD). Finally, a four-wire RTD would connect to terminals 1, 2, 3, and 4 (terminals 1 and 2 being common, and terminals 3 and 4 being common, at the RTD).

Once the RTD has been connected to the appropriate terminals of the temperature transmitter, the transmitter needs to be electronically configured for that type of RTD. In the case of this particular temperature transmitter, the configuration is performed using a "smart" communicator device using the HART digital protocol to access the transmitter's microprocessor-based settings. Here, the technician would configure the transmitter for 2-wire, 3-wire, or 4-wire RTD connection.

20.3.5 Self-heating error

One problem inherent to both thermistors and RTDs is *self-heating*. In order to measure the resistance of either device, we must pass an electric current through it. Unfortunately, this results in the generation of heat at the resistance according to Joule's Law:

$$P = I^2 R$$

This dissipated power causes the thermistor or RTD to increase in temperature beyond its surrounding environment, introducing a positive measurement error. The effect may be minimized by limiting excitation current to a bare minimum, but this results in less voltage dropped across the device. The smaller the developed voltage, the more sensitive the voltage-measuring instrument must be to accurately sense the condition of the resistive element. Furthermore, a decreased signal voltage means we will have a decreased signal-to-noise ratio, for any given amount of noise induced in the circuit from external sources.

One clever way to circumvent the self-heating problem without diminishing excitation current to the point of uselessness is to *pulse* current through the resistive sensor and digitally sample the voltage only during those brief time periods while the thermistor or RTD is powered. This technique works well when we are able to tolerate slow sample rates from our temperature instrument, which is often the case because most temperature measurement applications are slow-changing by nature. The pulsed-current technique enjoys the further advantage of reducing power consumption for the instrument, an important factor in battery-powered temperature measurement applications.

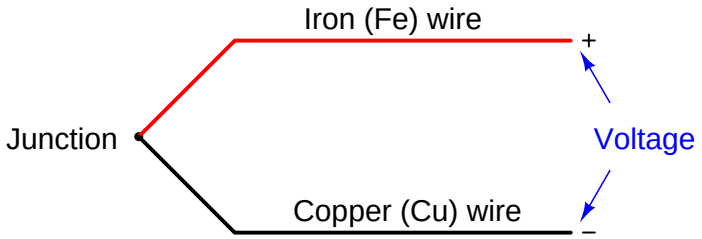
20.4 Thermocouples

RTDs are completely passive sensing elements, requiring the application of an externally-sourced electric current in order to function as temperature sensors. Thermocouples, however, generate their own electric potential. In some ways, this makes thermocouple systems simpler because the device receiving the thermocouple's signal does not have to supply electric power to the thermocouple. The self-powering nature of thermocouples also means they do not suffer from the same "self-heating" effect as RTDs. In other ways, thermocouple circuits are more complex than RTD circuits because the generation of voltage actually occurs in two different locations within the circuit, not simply at the sensing point. This means the receiving circuit must "compensate" for temperature in another location in order to accurately measure temperature in the desired location.

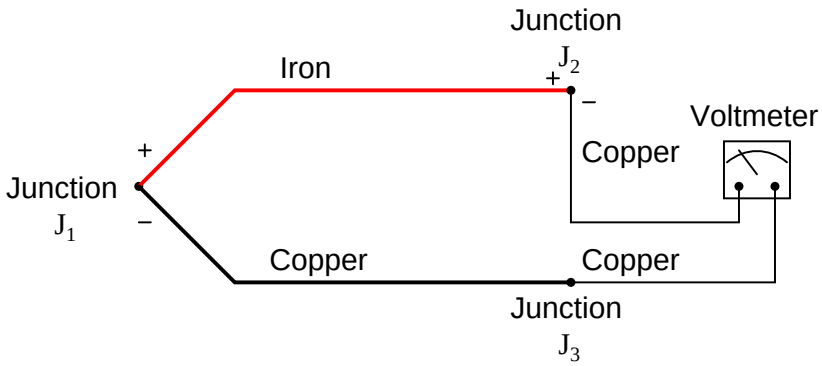
Though typically not as accurate as RTDs, thermocouples are more rugged, have greater temperature measurement spans, and are easier to manufacture in different physical forms.

20.4.1 Dissimilar metal junctions

When two dissimilar metal wires are joined together at one end, a voltage is produced at the other end that is approximately proportional to temperature. That is to say, the junction of two different metals behaves like a temperature-sensitive battery. This form of electrical temperature sensor is called a *thermocouple*:



This phenomenon provides us with a simple way to electrically infer temperature: simply measure the voltage produced by the junction, and you can tell the temperature of that junction. And it would be that simple, if it were not for an unavoidable consequence of electric circuits: when we connect any kind of electrical instrument the iron and copper wires, we inevitably produce another junction of dissimilar metals. The following schematic shows this fact:



Junction J_1 is a junction of iron and copper – two dissimilar metals – which will generate a voltage related to temperature. Note that junction J_2, which is necessary for the simple fact that we must somehow connect our copper-wired voltmeter to the iron wire, is also a dissimilar-metal junction which will generate a voltage related to temperature. Note also how the polarity of junction J_2 stands opposed to the polarity of junction J_1 (iron = positive ; copper = negative). A third junction (J_3) also exists between wires, but it is of no consequence because it is a junction of two identical metals which does not generate a temperature-dependent voltage at all.

The presence of this second voltage-generating junction (J_2) helps explain why the voltmeter registers 0 volts when the entire system is at room temperature: any voltage generated by the iron-copper junctions will be equal in magnitude and opposite in polarity, resulting in a net (series-total) voltage of zero. It is only when the two junctions J_1 and J_2 are at different temperatures that the voltmeter registers any voltage at all.

We may express this relationship mathematically as follows:

$$V_{meter} = V_{J1} - V_{J2}$$

With the measurement (J_1) and reference (J_2) junction voltages opposed to each other, the voltmeter only "sees" the difference between these two voltages.

Thus, thermocouple systems are fundamentally *differential* temperature sensors. That is, they provide an electrical output proportional to the difference in temperature between two different points. For this reason, the wire junction we use to measure the temperature of interest is called the *measurement junction* while the other junction (which we cannot eliminate from the circuit) is called the *reference junction* (or the *cold junction*, because it is typically at a cooler temperature than the process measurement junction).

20.4.2 Thermocouple types

Thermocouples exist in many different types, each with its own color codes for the dissimilar-metal wires. Here is a table showing the more common thermocouple types and their standardized colors[2], along with some distinguishing characteristics of the metal types to aid in polarity identification when the wire colors are not clearly visible:

Type	**Positive wire** *characteristic*	**Negative wire** *characteristic*	**Plug**	**Temp. range**
T	Copper (blue) *yellow colored*	Constantan (red) *silver colored*	Blue	-300 to 700 ^{o}F
J	Iron (white) *magnetic, rusty?*	Constantan (red) *non-magnetic*	Black	32 to 1400 ^{o}F
E	Chromel (violet) *shiny finish*	Constantan (red) *dull finish*	Violet	32 to 1600 ^{o}F
K	Chromel (yellow) *non-magnetic*	Alumel (red) *magnetic*	Yellow	32 to 2300 ^{o}F
N	Nicrosil (orange)	Nisil (red)	Orange	32 to 2300 ^{o}F
S	Pt90% - Rh10% (black)	Platinum (red)	Green	32 to 2700 ^{o}F
B	Pt70% - Rh30% (grey)	Pt94% - Rh6% (red)	Grey	32 to 3380 ^{o}F

Note how the negative ($-$) wire of every thermocouple type is color-coded *red*. While this may seem backward to those familiar with modern electronics (where red and black usually represent the positive and negative poles of a DC power supply, respectively), bear in mind that thermocouple color codes actually pre-date electronic power supply wire coloring!

Aside from having different usable temperature ranges, these thermocouple types also differ in terms of the atmospheres they may withstand at elevated temperatures. Type J thermocouples, for instance, by virtue of the fact that one of the wire types is *iron*, will rapidly corrode in any oxidizing[3] atmosphere. Type K thermocouples are attacked by reducing[4] atmospheres as well as sulfur and cyanide. Type T thermocouples are limited in upper temperature by the oxidation of copper (a very reactive metal when hot), but stand up to both oxidizing and reducing atmospheres quite well at lower temperatures, even when wet.

[2] The colors in this table apply only to the United States and Canada. A stunning diversity of colors has been "standardized" for each thermocouple type depending on where else in the world you go. The British and Czechs use their own color code, as do the Dutch and Germans. France has its own unique color code as well. Just for fun, an "international" color code also exists which does not match any of the others.

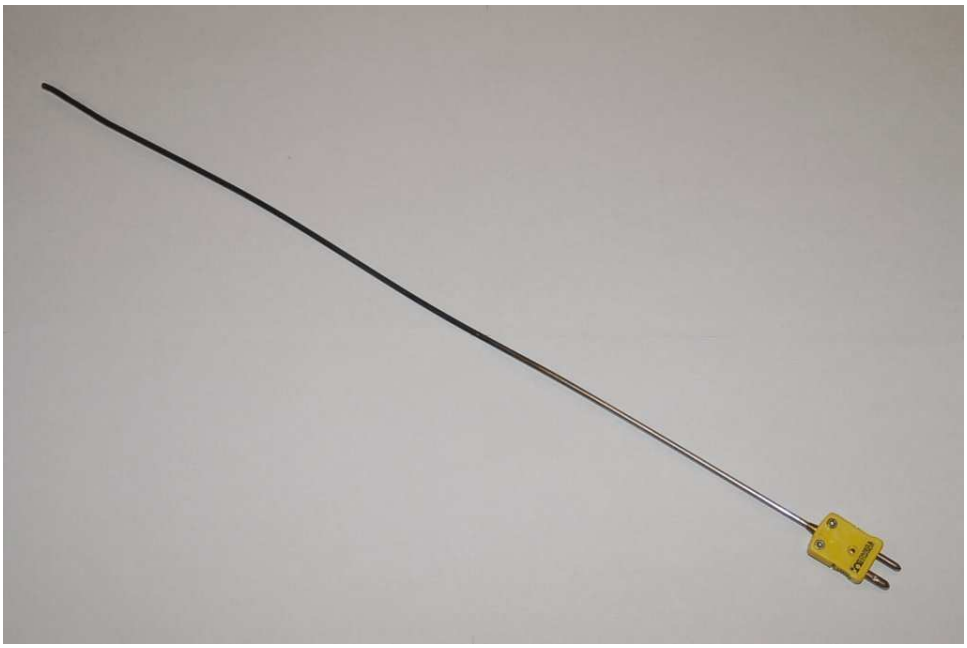
[3] By "oxidizing," what is meant is any atmosphere containing sufficient oxygen molecules or molecules of a similar element such as chlorine or fluorine.

[4] "Reducing" refers to atmospheres rich in elements that readily oxidize. Practically any fuel gas (hydrogen, methane, etc.) will create a reducing atmosphere in sufficient concentration.

20.4.3 Connector and tip styles

In its simplest form, a thermocouple is nothing more than a pair of dissimilar-metal wires joined together. However, in industrial practice, we often need to package thermocouples in a way that optimizes their ruggedness and reliability. For instance, most industrial thermocouples are manufactured in such a way that the dissimilar-metal wires are protected from physical damage by a stainless steel or ceramic *sheath*, and they are often equipped with molded-plastic plugs for quick connection to and disconnection from a thermocouple-based instrument.

A photograph of a type K industrial thermocouple (approximately 20 inches in length) reveals this "sheathed" and "connectorized" construction:



The stainless steel sheath of this particular thermocouple shows signs of discoloration from previous service in a hot process. Note the different diameters of the plug terminals. This "polarized" design makes it difficult[5] to insert backward into a matching socket.

[5] It should be noted that no amount of engineering or design is able to *completely* prevent people from doing the wrong thing. I have seen this style of thermocouple plug forcibly mated the wrong way to a socket. The amount of insertion force necessary to make the plug fit backward into the socket was quite extraordinary, yet this apparently was not enough of a clue for this wayward individual to give them pause.

A miniature version of this same plug (designed to attach to thermocouple wire by screw terminals, rather than be molded onto the end of a sheathed assembly) is shown here, situated next to a ballpoint pen for size comparison:

Industrial-grade thermocouples are available with this miniature style of molded plug end as an alternative to the larger (standard) plug. Miniature plug-ends are often the preferred choice for laboratory applications, while standard-sized plugs are often the preferred choice for field applications.

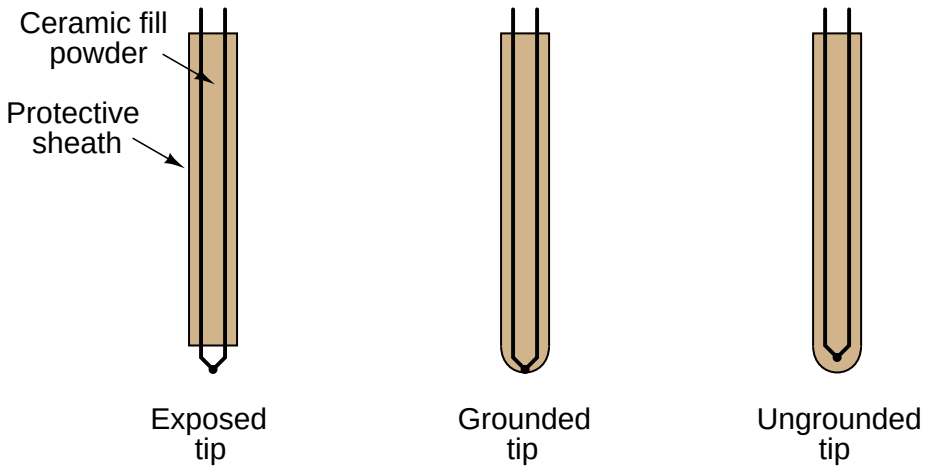
Some industrial thermocouples have no molded plug at all, but terminate simply in a pair of open wire ends. The following photograph shows a type J thermocouple of this construction:

If the electronic measuring instrument (e.g. temperature transmitter) is located near enough for the thermocouple's wires to reach the connection terminals, no plug or socket is needed at all in the circuit. If, however, the distance between the thermocouple and measuring instrument is too far to span with the thermocouple's own wires, a common termination technique is to attach a special terminal block and connection "head" to the top of the thermocouple allowing a pair of thermocouple extension wires to join and carry the millivoltage signal to the measuring instrument.

This next photograph shows a close-up view of such a thermocouple "head":

As you can see from this photograph, the screws directly press against the solid metal wires to make a firm connection between each wire and the brass terminal block. Since the "head" attaches directly to one end of the thermocouple, the thermocouple's wires will be trimmed just long enough to engage with the terminal screws inside the head. A threaded cover provides easy access to these connection points for installation and maintenance, while ensuring the connections are covered and protected from ambient weather conditions the rest of the time.

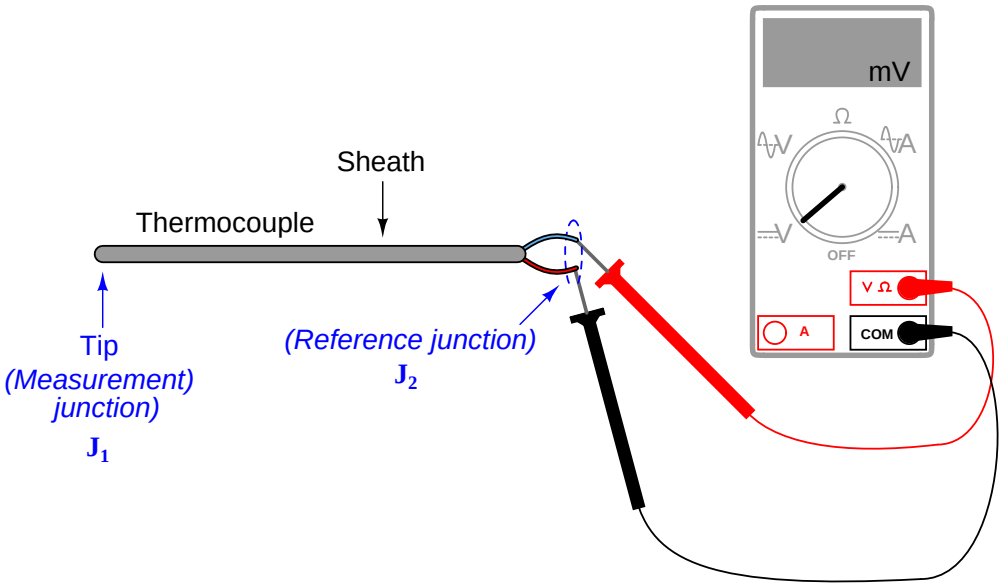
At the other end of the thermocouple, we have a choice of tip styles. For maximum sensitivity and fastest response, the dissimilar-metal junction may be unsheathed (bare). This design, however, makes the thermocouple more fragile. Sheathed tips are typical for industrial applications, available in either *grounded* or *ungrounded* forms:



Grounded-tip thermocouples exhibit faster response times and greater sensitivity than ungrounded-tip thermocouples, but they are vulnerable to *ground loops*: circuitous paths for electric current between the conductive sheath of the thermocouple and some other point in the thermocouple circuit. In order to avoid this potentially troublesome effect, most industrial thermocouples are of the ungrounded design.

20.4.4 Manually interpreting thermocouple voltages

Recall that the amount of voltage indicated by a voltmeter connected to a thermocouple is the *difference* between the voltage produced by the measurement junction (the point where the two dissimilar metals join at the location we desire to sense temperature at) and the voltage produced by the reference junction (the point where the thermocouple wires join to the voltmeter wires):



$$V_{meter} = V_{J1} - V_{J2}$$

This makes thermocouples inherently *differential* sensing devices: they generate a measurable voltage in proportion to the *difference* in temperature between two locations. This inescapable fact of thermocouple circuits complicates the task of interpreting any voltage measurement obtained from a thermocouple.

In order to translate a voltage measurement produced by a voltmeter connected to a thermocouple, we must *add* the voltage produced by the measurement junction (V_{J2}) to the voltage indicated by the voltmeter to find the voltage being produced by the measurement junction (V_{J1}). In other words, we manipulate the previous equation into the following form:

$$V_{J1} = V_{J2} + V_{meter}$$

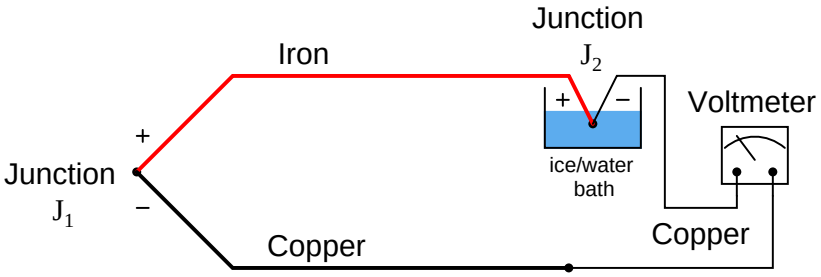
We may ascertain the reference junction voltage by placing a thermometer near that junction (where the thermocouple wire attaches to the voltmeter test leads) and referencing a table of thermocouple voltages for that thermocouple type. Then, we may take the voltage sum for V_{J1} and re-reference that same table, finding the temperature value corresponding to the calculated measurement junction voltage.

To illustrate, suppose we connected a voltmeter to a type K thermocouple and measured 14.30 millivolts. A thermometer situated near the thermocouple wire / voltmeter junction point shows an ambient temperature of 73 degrees Fahrenheit. Referencing a table of voltages for type K thermocouples (in this case, the NIST "ITS-90" reference standard), we see that a type K junction at 73 degrees Fahrenheit corresponds to 0.910 millivolts. Adding this figure to our meter measurement of 14.30 millivolts, we arrive at a sum of 15.21 millivolts for the measurement ("hot") junction. Going back to the same table of values, we see 15.21 millivolts falls between 701 and 702 degrees Fahrenheit. Linearly interpolating between the table values (15.203 mV at 701 ^{o}F and 15.226 mV at 702 ^{o}F), we may more precisely determine the measurement junction to be at 701.3 degrees Fahrenheit.

The process of manually taking voltage measurements, referencing a table of millivoltage values, performing addition, then re-referencing the same table is rather tedious. Compensation for the reference junction's inevitable presence in the thermocouple circuit is something we must do, but it is not something that must always be done by a human being. The next subsection discusses ways to automatically compensate for the effect of the reference junction, which is the only practical alternative for continuous thermocouple-based temperature instruments.

20.4.5 Reference junction compensation

Multiple techniques exist to deal with the influence of the reference junction's temperature[6]. One technique is to physically fix the temperature of that junction at some constant value so it is always stable. This way, any changes in measured voltage *must* be due to changes in temperature at the measurement junction, since the reference junction has been rendered incapable of changing temperature. This may be accomplished by immersing the reference junction in a bath of ice and water:

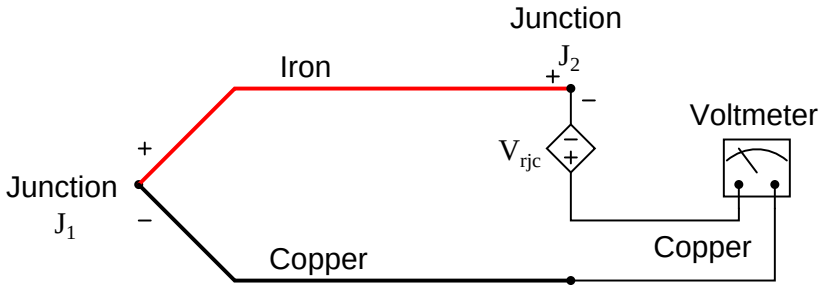


In fact, this is how thermocouple temperature/voltage tables are referenced: describing the amount of voltage produced for given temperatures at the measurement junction with the reference junction held at the freezing point of water (0 ^{o}C = 32 ^{o}F).

However, this is not a very practical solution for dealing with the reference junction's voltage. Instead, we could apply an additional electrical circuit to counter-act the voltage produced by the reference junction. This is called a *reference junction compensation* or *cold junction compensation* circuit:

[6] Early texts on thermocouple use describe multiple techniques for automatic compensation of the reference ("cold") junction. One design placed a mercury bulb thermometer at the reference junction, with a loop of thin platinum wire dipped into the mercury. As junction temperature rose, the mercury column would rise and short past a greater length of the platinum wire loop, causing its resistance to decrease which in turn would electrically bias the measurement circuit to offset the effects of the reference junction's voltage. Another design used a bi-metallic spring to offset the pointer of the meter movement, so that changes in temperature at the indicating instrument (where the reference junction was located) would result in the analog meter's needle becoming offset from its normal "zero" point, thus compensating for the offset in voltage created by the reference junction.

Compensating for the effects of J_2 using a "reference junction compensation" circuit to generate a counter-voltage



Please note that "cold junction" is just a synonymous label for "reference junction." In fact the "cold" reference junction may very well be at a warmer temperature than the so-called "hot" measurement junction! Nothing prevents anyone from using a thermocouple to measure temperatures below freezing.

This compensating voltage source (V_{rjc} in the above schematic) uses some other temperature-sensing device such as a thermistor or RTD to sense the local temperature at the terminal block where junction J_2 is formed, and produce a counter-voltage that is precisely equal and opposite to J_2's voltage ($V_{rjc} = V_{J2}$). Having canceled the effect of the reference junction, the voltmeter now only registers the voltage produced by the measurement junction J_1:

$$V_{meter} = V_{J1} - V_{J2} + V_{rjc}$$

$$V_{meter} = V_{J1} + 0$$

$$V_{meter} = V_{J1}$$

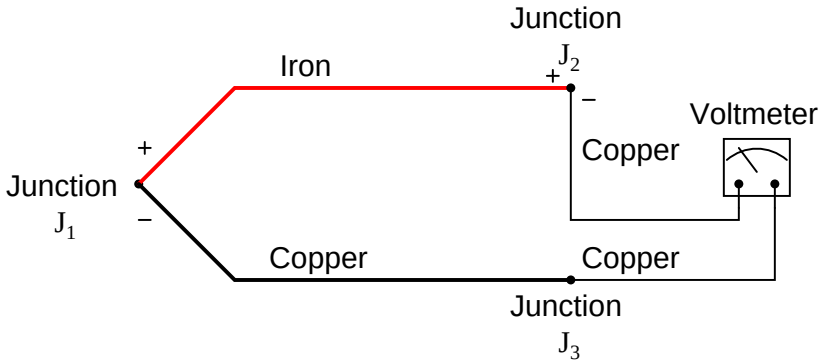
Some instrument manufacturers sell *electronic ice point* modules designed to provide reference junction compensation for un-compensated instruments such as standard voltmeters. The "ice point" circuit performs the function shown by V_{rjc} in the previous diagram: it inserts a counter-acting voltage to cancel the voltage generated by the reference junction, so that the voltmeter only "sees" the measurement junction's voltage. This compensating voltage is maintained at the proper value according to the terminal temperature where the thermocouple wires connect to the ice point module, sensed by a thermistor or RTD.

At first it may seem pointless to go through the trouble of building a reference junction compensation (ice point) circuit, when doing so requires the use of some other temperature-sensing element such as a thermistor or RTD. After all, why bother to do this just to be able to use a thermocouple to accurately measure temperature, when we could simply use this "other" device to directly measure the process temperature? In other words, isn't the usefulness of a thermocouple invalidated if we have to go through the trouble of integrating another type of electrical temperature sensor in the circuit just to compensate for an idiosyncrasy of thermocouples?

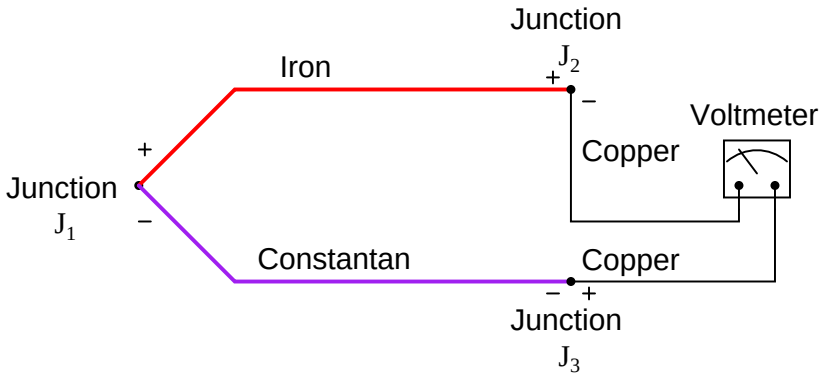
The answer to this very good question is that thermocouples enjoy certain advantages over these other sensor types. Thermocouples are extremely rugged and have far greater temperature-measurement ranges than thermistors, RTDs, and other primary sensing elements. However, if the application does not demand extreme ruggedness or large measurement ranges, a thermistor or RTD would likely be the better choice.

20.4.6 Law of Intermediate Metals

It is critical to realize that the phenomenon of a "reference junction" is an inevitable effect of having to close the electric circuit loop in a circuit made of dissimilar metals. This is true regardless of the number of metals involved. In the last example, only two metals were involved: iron and copper. This formed one iron-copper junction (J_1) at the measurement end and one iron-copper junction (J_2) at the indicator end. Recall that the copper-copper junction J_3 was of no consequence because its identical metallic composition generates no thermal voltage:

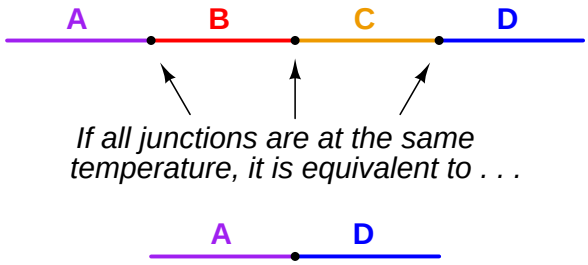


The same thing happens when we form a thermocouple out of two metals, neither one being copper. Take for instance this example of a type J thermocouple:

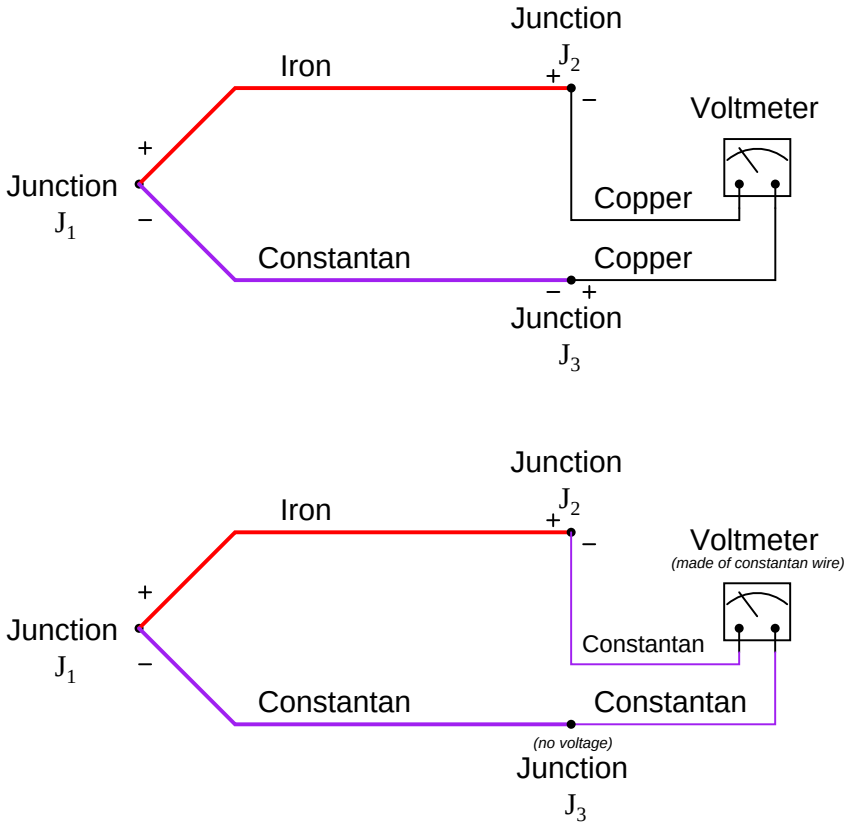


Here we have *three* voltage-generating junctions: J_1 of iron and constantan, J_2 of iron and copper, and J_3 of copper and constantan which just happens to be the metallic combination for a type T thermocouple. Upon first inspection it would seem we have a much more complex situation than we did with just two metals (iron and copper), but the situation is actually just as simple as it was before.

A principle of thermo-electric circuits called the *Law of Intermediate Metals* helps us see this clearly. According to this law, intermediate metals in a series of junctions are of no consequence to the overall (net) voltage so long as those intermediate junctions are all at the same temperature. Representing this pictorially, the net effect of having four different metals (A, B, C, and D) joined together in series is the same as just having the first and last metal in that series (A and D) joined with one junction, if all intermediate junctions are at the same temperature:

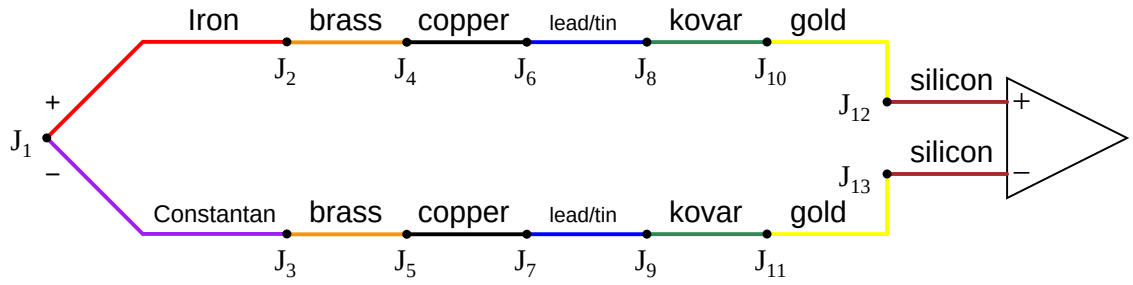


In our Type J thermocouple circuit where iron and constantan both join to copper, we see copper as an intermediate metal so long as junctions J_2 and J_3 are at the same temperature. Since those two junctions are located next to each other on the indicating instrument, identical temperature is a reasonable assumption, and we may treat junctions J_2 and J_3 as a single iron-constantan reference junction. In other words, the Law of Intermediate Metals tells us we can treat these two circuits identically:



The practical importance of this Law is that we can always treat the reference junction(s) as a single junction made from the same two metal types as the measurement junction, so long as all dissimilar metal junctions at the reference location are at the same temperature.

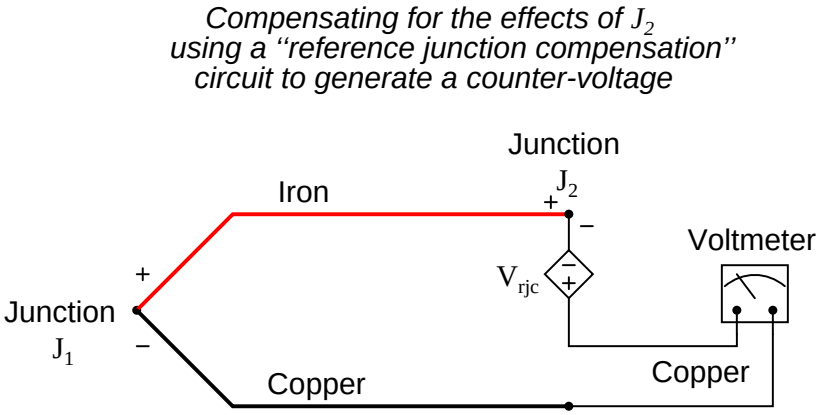
This fact is extremely important in the age of semiconductor circuitry, where the connection of a thermocouple to an electronic amplifier involves many different junctions, from the thermocouple wires to the amplifier's silicon. Here we see a multitude of reference junctions, inevitably formed by the necessary connections from thermocouple wire to the silicon substrate inside the amplifier chip:



It should be obvious that each complementary junction pair cancels if each pair is at the same temperature (e.g. gold-silicon junction J_{12} cancels with silicon-gold junction J_{13} because they generate the exact same amount of voltage with opposing polarities). The Law of Intermediate Metals goes one step further by telling us junctions J_2 through J_{13} taken together in series are of the same effect as a single reference junction of iron and constantan. Automatic reference junction compensation is as simple as counter-acting the voltage produced by this equivalent iron-constantan junction at whatever temperature junctions J_2 through J_{13} happen to be at.

20.4.7 Software compensation

Previously, it was suggested this automatic compensation could be accomplished by intentionally inserting a temperature-dependent voltage source in series with the circuit, oriented in such a way as to oppose the reference junction's voltage:



$$V_{meter} = V_{J1} - V_{J2} + V_{rjc}$$

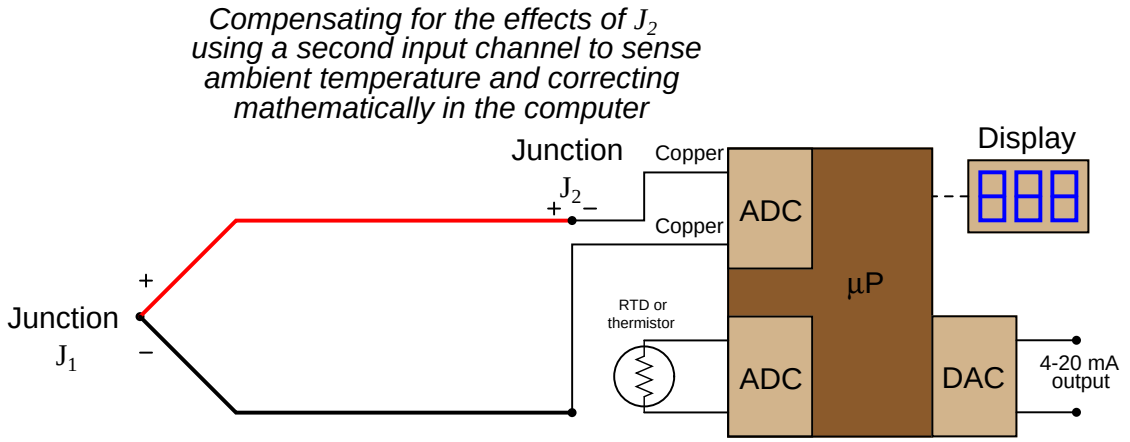
If the series voltage source V_{rjc} is exactly equal in magnitude to the reference junction's voltage (V_{J2}), those two terms cancel out of the equation and lead to the voltmeter measuring only the voltage of the measurement junction J_1:

$$V_{meter} = V_{J1} + 0$$

$$V_{meter} = V_{J1}$$

This technique is known as *hardware* compensation. A stand-alone circuit designed to do this is sometimes called an *ice point*, because it electrically accomplishes the same thing as physically placing the reference junction(s) in a bath of ice-water.

A more modern technique for reference junction compensation is called *software* compensation. This is applicable only where the indicating device is microprocessor-based, and where an additional analog input channel exists:



Instead of canceling the effect of the reference junction electrically, we can cancel the effect mathematically inside the microprocessor. In other words, we let the meter see the difference in voltage between the measurement and reference junctions ($V_{meter} = V_{J1} - V_{J2}$). After digitizing this voltage measurement, the microprocessor adds the equivalent voltage value corresponding to the ambient temperature sensed by the RTD or thermistor (V_{rjc}):

$$\text{Compensated total} = V_{meter} + V_{rjc}$$

$$\text{Compensated total} = (V_{J1} - V_{J2}) + V_{rjc}$$

Since we know the calculated value of V_{rjc} should be equal to the real reference junction voltage (V_{J2}), the result of this digital addition should be a compensated total equal only to the measurement junction voltage V_{J1}:

$$\text{Compensated total} = V_{J1} - V_{J2} + V_{rjc}$$

$$\text{Compensated total} = V_{J1} + 0$$

$$\text{Compensated total} = V_{J1}$$

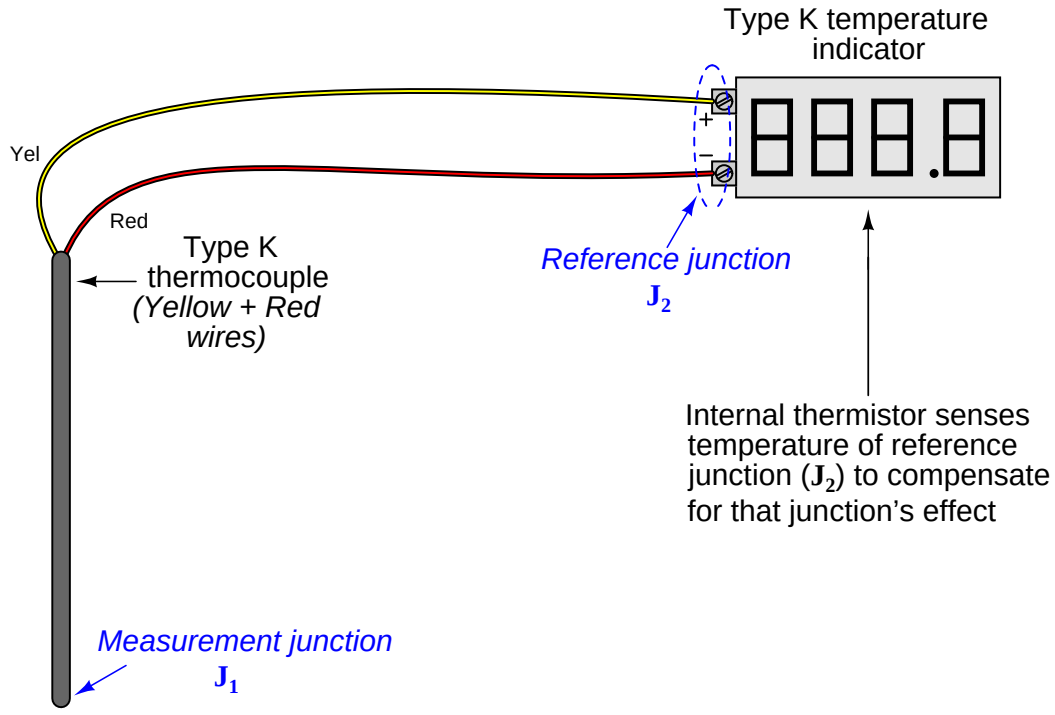
Perhaps the greatest advantage of software compensation is flexibility. Being able to re-program the compensation function means this instrument may easily interpret the voltage output of different thermocouple types with no modifications to the hardware. So long as the microprocessor memory is programmed with look-up tables relating voltage values to temperature values, it may accurately measure (and compensate for the reference junction of) any thermocouple type. With hardware-based compensation (an "ice point" circuit), re-wiring or replacement is necessary to accommodate different thermocouple types.

20.4.8 Extension wire

In every thermocouple circuit there must be both a measurement junction and a reference junction: this is an inevitable consequence of forming a complete circuit (loop) using dissimilar-metal wires. As we already know, the voltage received by the measuring instrument from a thermocouple will be the *difference* between the voltages produced by the measurement and reference junctions. Since the purpose of most temperature instruments is to accurately measure temperature *at a specific location*, the effects of the reference junction's voltage must be "compensated" for by some means, either a special circuit designed to add an additional canceling voltage or by a software algorithm to digitally cancel the reference junction's effect.

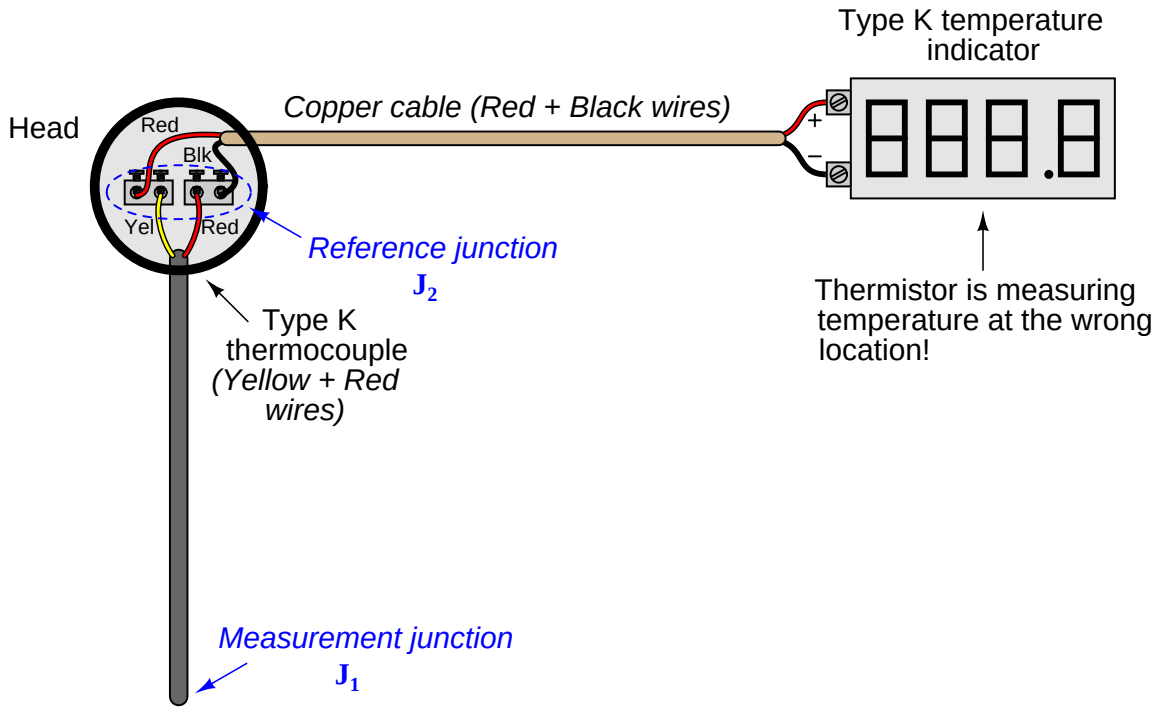
In order for reference junction compensation to be effective, the compensation mechanism must "know" the temperature of the reference junction. This fact is so obvious, it hardly requires statement. However, what is not so obvious is how easily this compensation may be unintentionally defeated simply by installing a different type of wire in a thermocouple circuit.

To illustrate, let us examine a simple type K thermocouple installation, where the thermocouple connects directly to a panel-mounted temperature indicator:



Like all modern thermocouple instruments, the panel-mounted indicator contains its own reference junction compensation, so that it is able to compensate for the temperature of the reference junction formed at its connection terminals, where the internal (copper) wires of the indicator join to the chromel and alumel wires of the thermocouple. The indicator senses this junction temperature using a small thermistor thermally bonded to the connection terminals.

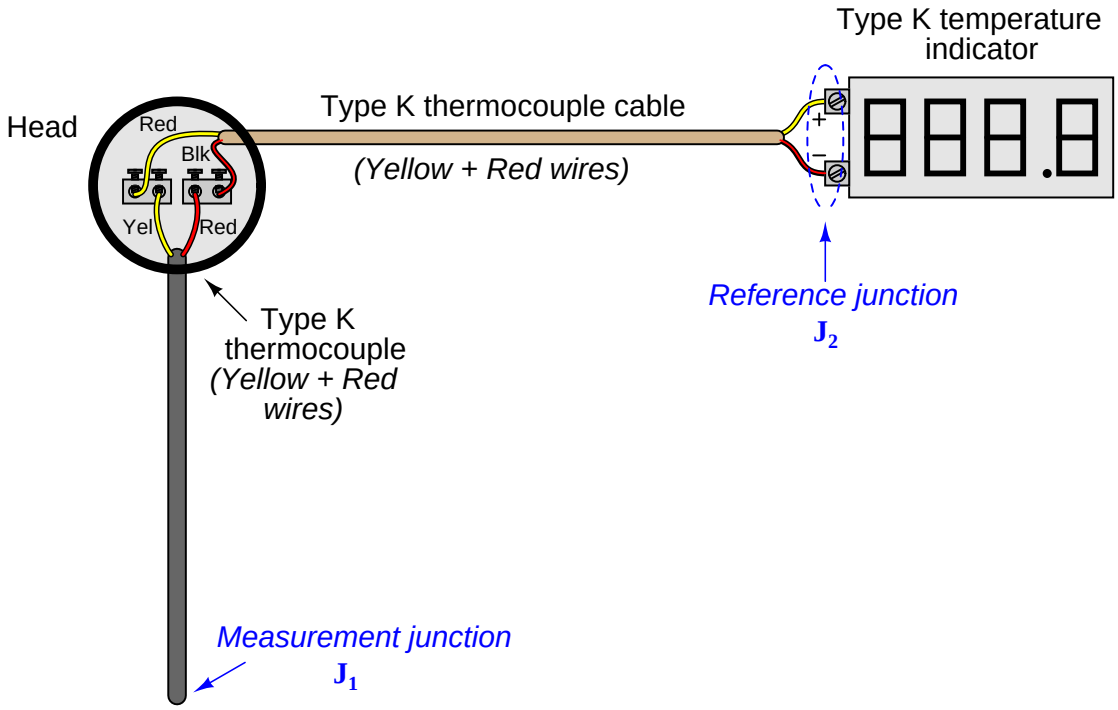
Now let us consider the same thermocouple installation with a length of copper cable (two wires) joining the field-mounted thermocouple to the panel-mounted indicator:



Even though nothing has changed in the thermocouple circuit except for the type of wires joining the thermocouple to the indicator, the reference junction has completely shifted position. What used to be a reference junction (at the indicator's terminals) is no longer, because now we have copper wires joining to copper wires. Where there is no dissimilarity of metals, there can be no thermoelectric potential. At the thermocouple's connection "head," however we now have a joining of chromel and alumel wires to copper wires, thus forming a reference junction *in a novel location*. What is worse, this new location is likely to be at a different temperature than the panel-mounted indicator, which means the indicator's reference junction compensation will be compensating for the wrong temperature.

The only practical way to avoid this problem is to keep the reference junction where it belongs: at the terminals of the panel-mounted instrument where the ambient temperature is measured and the reference junction's effects accurately compensated. If we must install "extension" wire to join a thermocouple to a remotely-located instrument, that wire must be of a type that does not form another dissimilar-metal junction at the thermocouple head, but will form one at the receiving instrument.

An obvious approach is to simply use thermocouple wire of the same type as the installed thermocouple to join the thermocouple to the indicator. For our hypothetical type K thermocouple, this means a type K cable installed between the thermocouple head and the panel-mounted indicator:



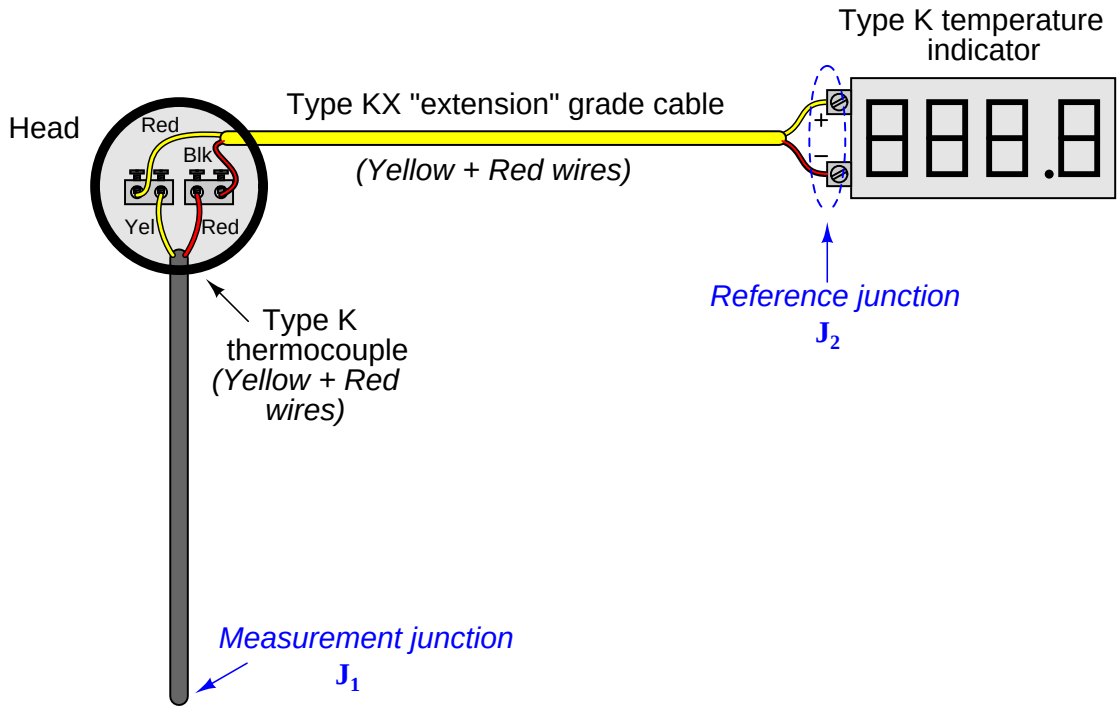
With chromel joining to chromel and alumel joining to alumel at the head, no dissimilar-metal junctions are created at the thermocouple. However, with chromel and alumel joining to copper at the indicator (again), the reference junction has been re-located to its rightful place. This means the thermocouple head's temperature will have no effect on the performance of this measurement system, and the indicator will be able to properly compensate for any ambient temperature changes at the panel as it was designed to do. The only problem with this approach is the potential expense of thermocouple-grade cable. This is especially true with some types of thermocouples, where the metals used are somewhat exotic.

A more economical alternative, however, is to use something called *extension-grade* wire to make the connection between the thermocouple and the receiving instrument. "Extension-grade" thermocouple wire is made less expensive than full "thermocouple-grade" wire by choosing metal alloys similar in thermo-electrical characteristics to the real thermocouple wires within modest temperature ranges. So long as the temperatures at the thermocouple head and receiving instrument terminals never exceed a modest range, the extension wire metals joining to the thermocouple wires and joining to the instrument's copper wires need not be *precisely identical* to the true thermocouple wire alloys. This allows for a wider selection of metal types, some of which may be substantially less expensive than the measurement-grade thermocouple alloys. Also, extension-grade wire may use insulation with a narrower temperature rating than thermocouple-grade wire, reducing cost even

further.

An interesting historical reference to the use of extension-grade wire appears in Charles Robert Darling's 1911 text *Pyrometry – A Practical Treatise on the Measurement of High Temperatures*. On page 61, Darling describes "compensating leads" marketed under the name *Peake* designed to be used with platinum-alloy thermocouples. These "compensating" wires were made of two different copper-nickel alloys, each copper-nickel alloy matched with the respective thermocouple metal (in this case, pure platinum and a 90%-10% platinum-iridium alloy) to generate an equal and opposite millivoltage at any reasonable temperature found at the thermocouple head. Thus, the only reference junction in the thermocouple circuit is where these copper-nickel extension wires joined with the indicating instrument, rather than being located at the thermocouple head as it would be if simple copper extension wires were employed.

Extension-grade cable is denoted by a letter "X" following the thermocouple letter. For our hypothetical type K thermocouple system, this would mean type "KX" extension cable:

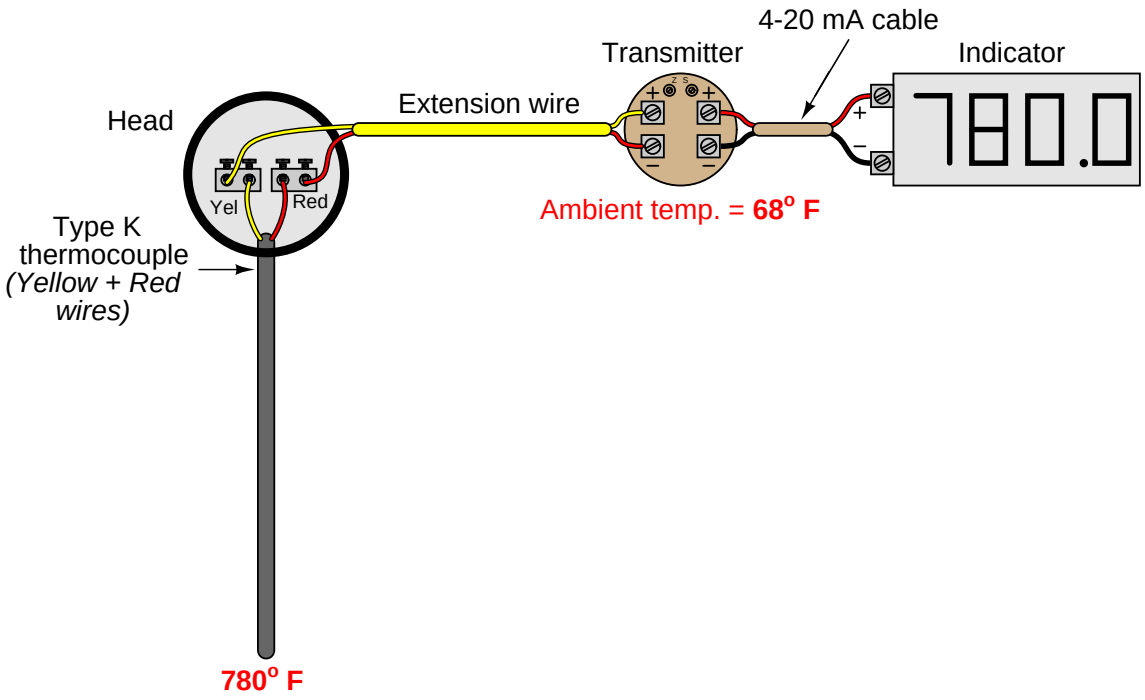


Thermocouple extension cable also differs from thermocouple-grade (measurement) cable in the coloring of its outer jacket. Whereas thermocouple-grade cable is typically brown in exterior color, extension-grade cable is usually colored to match the thermocouple plug (yellow for type K, black for type J, blue for type T, etc.).

20.4.9 Side-effects of reference junction compensation

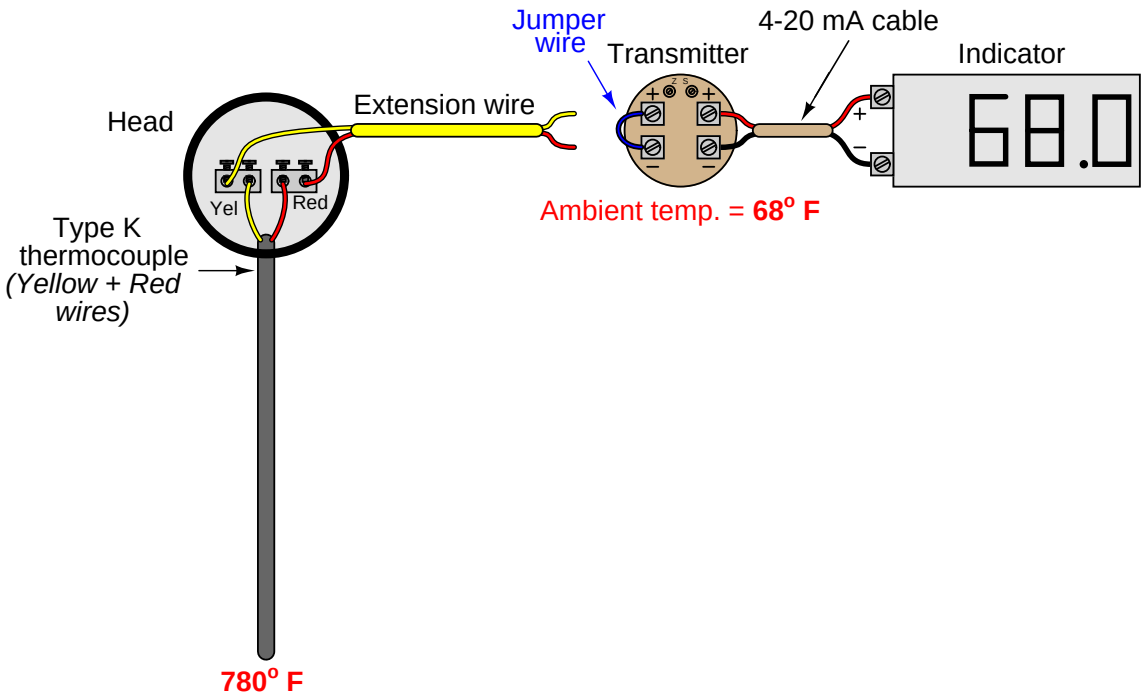
Reference junction compensation is a necessary part of any precision thermocouple circuit, due to the inescapable fact of the reference junction's existence. When you form a complete circuit of dissimilar metals, you *will* form both a measurement junction and a reference junction, with those two junctions' polarities opposed to one another. This is why reference junction compensation – whether it takes the form of a hardware circuit or an algorithm in software – must exist within every precision thermocouple instrument.

The presence of reference junction compensation in every precision thermocouple instrument results in an interesting phenomenon: *if you directly short-circuit the thermocouple input terminals of such an instrument, it will always register ambient temperature, regardless of the thermocouple type the instrument is built or configured for.* This behavior may be illustrated by example, first showing a normal operating temperature measurement system and then with that same system short-circuited. Here we see a temperature indicator receiving a 4-20 mA current signal from a temperature transmitter, which is receiving a millivoltage signal from a type "K" thermocouple sensing a process temperature of 780 degrees Fahrenheit:



The transmitter's internal reference junction compensation feature compensates for the ambient temperature of 68 degrees Fahrenheit. If the ambient temperature rises or falls, the compensation will automatically adjust for the change in reference junction potential, such that the output will still register the process (measurement junction) temperature of 780 degrees F. This is what the reference junction compensation is designed to do.

Now, we disconnect the thermocouple from the temperature transmitter and short-circuit the transmitter's input:



With the input short-circuited, the transmitter "sees" no voltage at all from the thermocouple circuit. There is no measurement junction nor a reference junction to compensate for, just a piece of wire making both input terminals electrically common. This means the reference junction compensation inside the transmitter no longer performs a useful function. However, the transmitter does not "know" it is no longer connected to the thermocouple, so the compensation keeps on working even though it has nothing to compensate for. Recall the voltage equation relating measurement, reference, and compensation voltages in a hardware-compensated thermocouple instrument:

$$V_{meter} = V_{J1} - V_{J2} + V_{rjc}$$

Disconnecting the thermocouple wire and connecting a shorting jumper to the instrument eliminates the V_{J1} and V_{J2} terms, leaving only the compensation voltage to be read by the meter[7]:

$$V_{meter} = 0 + V_{rjc}$$

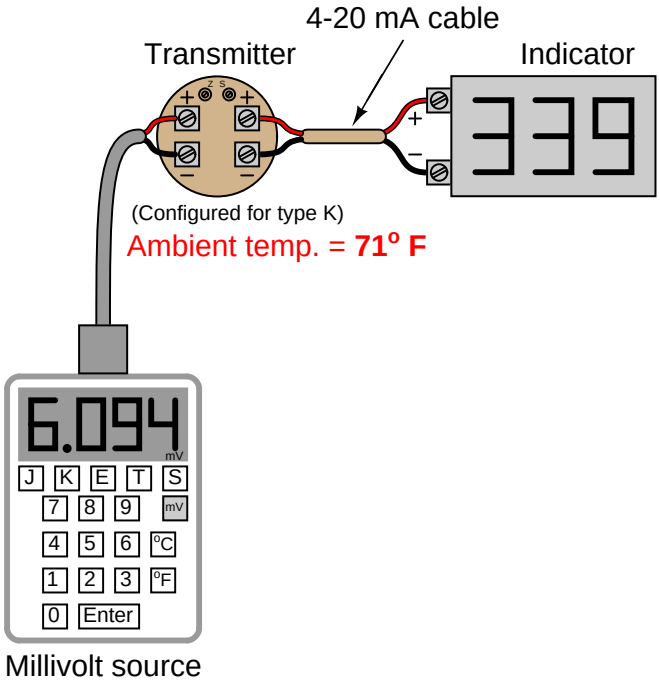
$$V_{meter} = V_{rjc}$$

[7] The effect will be exactly the same for an instrument with software compensation rather than hardware compensation. With software compensation, there is no literal V_{rjc} voltage source, but the equivalent millivolt value is digitally added to the zero input measured at the thermocouple connection terminals, resulting in the same effect of measuring ambient temperature.

This is why the instrument registers the equivalent temperature created by the reference junction compensation feature: this is the only signal it "sees" with its input short-circuited. This phenomenon is true regardless of which thermocouple type the instrument is configured for, which makes it a convenient "quick test" of instrument function in the field. If a technician short-circuits the input terminals of any thermocouple instrument, it should respond as though it is sensing ambient temperature.

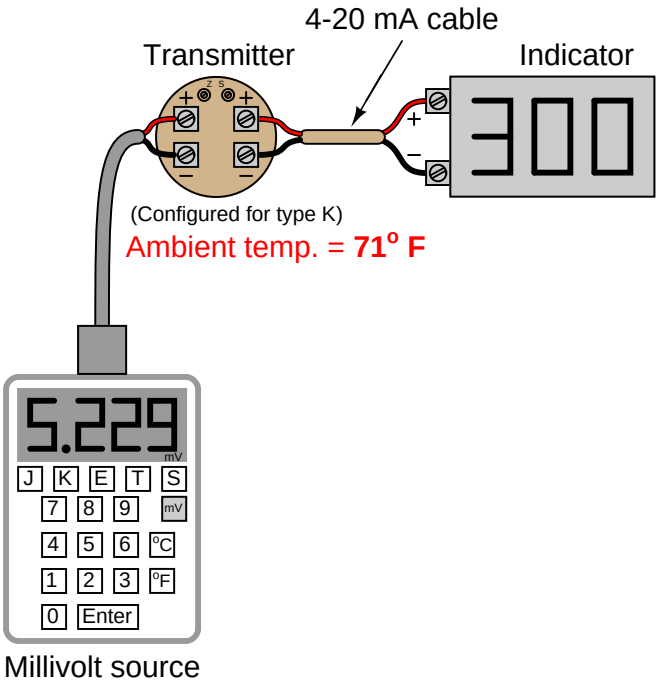
While this interesting trait is a somewhat useful side-effect of reference junction compensation in thermocouple instruments, there are other effects that are not quite so useful. The presence of reference junction compensation becomes quite troublesome, for example, if one tries to simulate a thermocouple using a precision millivoltage source. Simply setting the millivoltage source to the value corresponding to the desired (simulation) temperature given in a thermocouple table will yield an incorrect result for any ambient temperature other than the freezing point of water!

Suppose, for example, a technician wished to simulate a type K thermocouple at 300 degrees Fahrenheit by setting a millivolt source to 6.094 millivolts (the voltage corresponding to 300^o F for type K thermocouples according to the ITS-90 standard). Connecting the millivolt source to the instrument will *not* result in an instrument response appropriate for 300 degrees F:



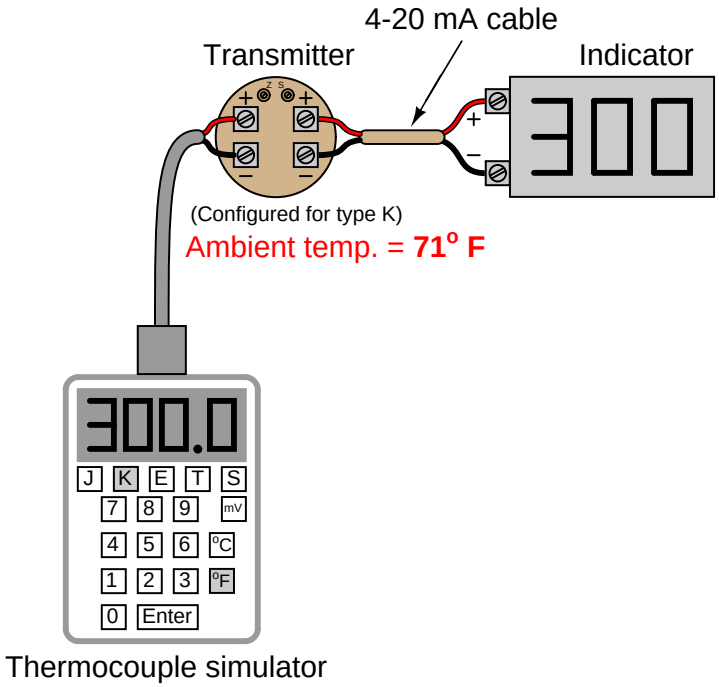
Instead, the instrument registers 339 degrees because its internal reference junction compensation feature is still active, compensating for a reference junction voltage that no longer exists. The millivolt source's output of 6.094 mV gets *added* to the compensation voltage (inside the transmitter) of 0.865 mV – the necessary millivolt value to compensate for a type K reference junction at 71^o F – with the result being a larger millivoltage (6.959 mV) interpreted by the transmitter as a temperature of 339^o F.

The only way to properly use a millivoltage source to simulate a desired temperature is for the instrument technician to "out-think" the transmitter's compensation feature by specifying a millivolt signal that is offset by the amount of equivalent voltage generated by the transmitter's compensation. In other words, instead of setting the millivolt source to a value of 6.094 mV, the technician should set the source to only 5.229 mV so the transmitter will add 0.865 mV to this value to arrive at 6.094 mV and register as 300 degrees Fahrenheit:



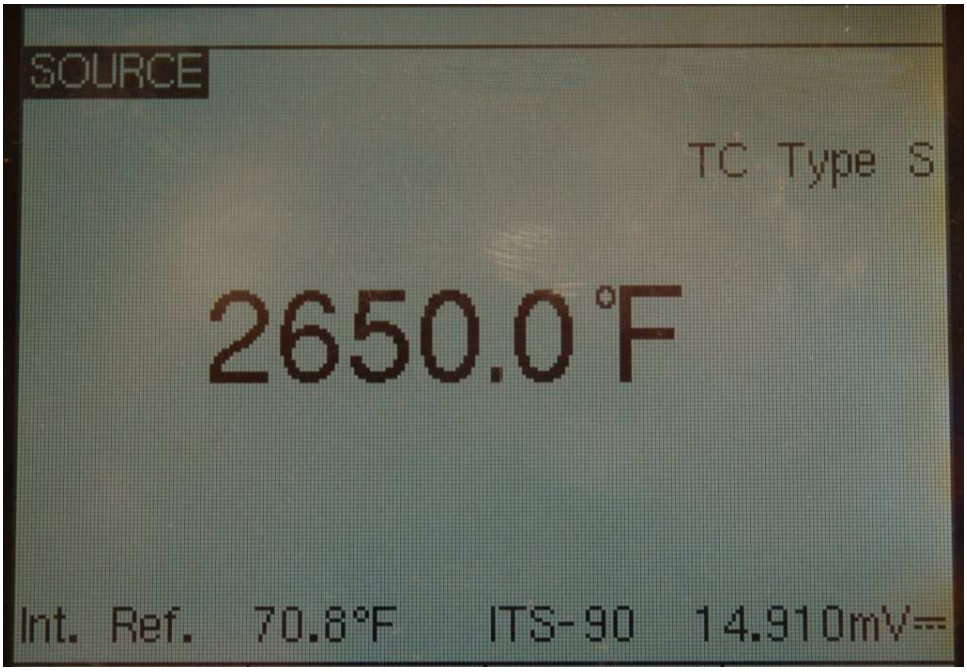
Years ago, the only suitable piece of test equipment available for generating the precise millivoltage signals necessary to calibrate thermocouple instruments was a device called a *precision potentiometer*. These "potentiometers" used a stable *mercury cell* battery (sometimes called a *standard cell*) as a voltage reference and a potentiometer with a calibrated knob to output low-voltage signals. Photographs of two vintage precision potentiometers are shown here:

Of course, modern thermocouple calibrators also provide direct entry of temperature and automatic compensation to "un-compensate" the transmitter such that any desired temperature may be easily simulated:



In this example, when the technician sets the calibrator for 300^o F (type K), it measures the ambient temperature and automatically subtracts 0.865 mV from the output signal, so only 5.229 mV is sent to the transmitter terminals instead of the full 6.094 mV. The transmitter's internal reference junction compensation adds the 0.865 mV offset value (thinking it must compensate for a reference junction that in reality is not there) and "sees" a total signal voltage of 6.094 mV, interpreting this properly as 300 degrees Fahrenheit.

The following photograph shows the display of a modern thermocouple calibration device (a Fluke model 744 documenting process calibrator) being used to generate a thermocouple signal. In this particular example, the thermocouple type is set to type "S" (Platinum-Rhodium/Platinum) at a temperature of 2650 degrees Fahrenheit:



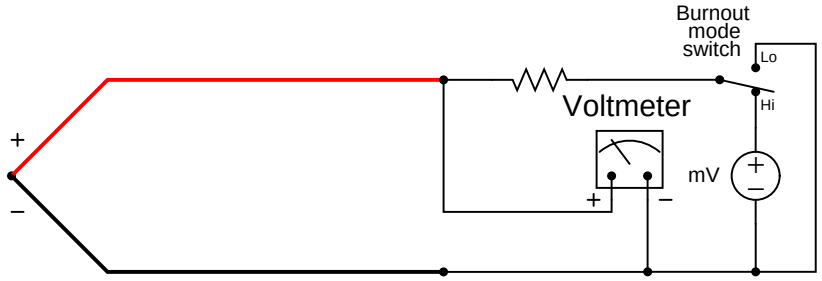
The ITS-90 thermocouple standard declares a millivoltage signal value of 15.032 mV for a type S thermocouple junction at 2650 degrees F (with a reference junction temperature of 32 degrees F). Note how the calibrator does *not* output 15.032 mV even though the simulated temperature has been set to 2650 degrees F. Instead, it outputs 14.910 mV, which is 0.122 mV less than 15.032 mV. This offset of 0.122 mV corresponds to the calibrator's local temperature of 70.8 degrees F (according to the ITS-90 standard for type S thermocouple junctions).

When the calibrator's 14.910 mV signal reaches the thermocouple instrument being calibrated (be it an indicator, transmitter, or even a controller equipped with a type S thermocouple input), the instrument's own internal reference junction compensation will add 0.122 mV to the received signal of 14.910 mV, "thinking" it needs to compensate for a real reference junction. The result will be a perceived measurement junction signal of 15.032 mV, which is exactly what we want the instrument to "think" it sees if our goal is to simulate connection to a real type S thermocouple at a temperature of 2650 degrees F.

20.4.10 Burnout detection

Another consideration for thermocouples is *burnout detection*. The most common failure mode for thermocouples is to fail open, otherwise known as "burning out." An open thermocouple is problematic for any voltage-measuring instrument with high input impedance because the lack of a complete circuit on the input makes it possible for electrical noise from surrounding sources (power lines, electric motors, variable-frequency motor drives) to be detected by the instrument and falsely interpreted as a wildly varying temperature.

For this reason it is prudent to design into the thermocouple instrument some provision for generating a consistent state in the absence of a complete circuit. This is called the *burnout mode* of a thermocouple instrument.



The resistor in this circuit provides a path for current in the event of an open thermocouple. It is sized in the mega-ohm range to minimize its effect during normal operation when the thermocouple circuit is complete. Only when the thermocouple fails open will the miniscule current through the resistor have any substantial effect on the voltmeter's indication. The SPDT switch provides a selectable burnout mode: in the event of a burnt-out thermocouple, we can configure the meter to either read high temperature (sourced by the instrument's internal milli-voltage source) or low temperature (grounded), depending on what failure mode we deem safest for the application.

20.5 Non-contact temperature sensors

Virtually any mass above absolute zero temperature emits electromagnetic radiation (photons, or light) as a function of that temperature. This basic fact makes possible the measurement of temperature by analyzing the light emitted by an object. The *Stefan-Boltzmann Law* of radiated energy quantifies this fact, declaring that the rate of heat lost by radiant emission from a hot object is proportional to the fourth power of the absolute temperature:

$$\frac{dQ}{dt} = e\sigma AT^4$$

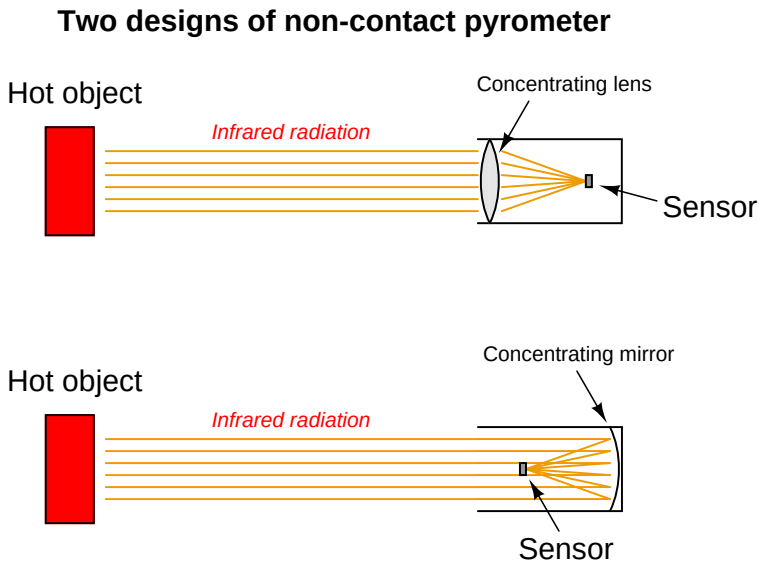
Where,
$\frac{dQ}{dt}$ = Radiant heat loss rate (watts)
e = Emissivity factor (unitless)
σ = Stefan-Boltzmann constant (5.67×10^{-8} W / m^2 $\cdot$ K^4)
A = Surface area (square meters)
T = Absolute temperature (Kelvin)

The primary advantage of non-contact thermometry (or *pyrometry* as high-temperature measurement is often referred) is rather obvious: with no need to place a sensor in direct contact with the process, a wide variety of temperature measurements may be made that are either impractical or impossible to make using any other technology.

It may surprise some readers to discover that non-contact pyrometry is nearly as old as thermocouple technology[8], the first non-contact pyrometer being constructed in 1892.

[8] Although Seebeck discovered thermo-electricity in 1822, the technique of measuring temperature by sensing the voltage produced at a dissimilar-metal junction was delayed in practical development until 1886 when rugged and accurate electrical meters became available for industrial use.

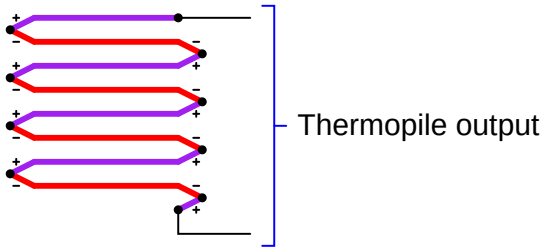
A time-honored design for non-contact pyrometers is to concentrate incident light from a heated object onto a small temperature-sensing element. A rise in temperature at the sensor reveals the intensity of the infrared optical energy falling upon it, which as discussed previously is a function of the target object's temperature (absolute temperature to the fourth power):



The fourth-power characteristic of Stefan-Boltzmann's law means that a doubling of absolute temperature at the hot object results in sixteen times as much radiant energy falling on the sensor, and therefore a sixteen-fold increase in the sensor's temperature rise over ambient. A tripling of target temperature (absolute) yields *eighty one times as much radiant energy*, and therefore an 81-fold increase in sensor temperature rise. This extreme nonlinearity limits the practical application of non-contact pyrometry to relatively narrow ranges of target temperature wherever good accuracy is required.

Thermocouples were the first type of sensor used in non-contact pyrometers, and they still find application in modern versions of the same technology. Since the sensor does not become nearly as hot as the target object, the output of any single thermocouple junction at the sensor area will be quite small. For this reason, instrument manufacturers often employ a series-connected array of thermocouples called a *thermopile* to generate a stronger electrical signal.

The basic concept of a thermopile is to connect multiple thermocouple junctions in series so their voltages will add:



Examining the polarity marks of each junction (type E thermocouple wires are assumed in this example: chromel and constantan), we see that all the "hot" junctions' voltages aid each other, as do all the "cold" junctions' voltages. Like all thermocouple circuits, though, the each "cold" junction voltage opposes each the "hot" junction voltage. The example thermopile shown in this diagram, with four hot junctions and four cold junctions, will generate four times the potential difference that a single type E thermocouple hot/cold junction pair would generate, assuming all the hot junctions are at the same temperature and all the cold junctions are at the same temperature.

When used as the detector for a non-contact pyrometer, the thermopile is oriented so all the concentrated light falls on the hot junctions, while the cold junctions face away from the focal point to a region of ambient temperature. Thus, the thermopile acts like a multiplied thermocouple, generating more voltage than a single thermocouple junction could under the same temperature conditions.

A popular design of non-contact pyrometer manufactured for years by Honeywell was the *Radiamatic*[9], using ten thermocouple junction pairs arrayed in a circle. All the "hot" junctions were placed at the center of this circle where the focal point of the concentrated light fell, while all the "cold" junctions were situated around the circumference of the circle away from the heat of the focal point. A table of values showing the approximate relationship between target temperature and millivolt output for one model of Radiamatic sensing unit reveals the fourth-power function:

Target temperature (K)	**Millivolt output**
4144 K	34.8 mV
3866 K	26.6 mV
3589 K	19.7 mV
3311 K	14.0 mV
3033 K	9.9 mV
2755 K	6.6 mV
2478 K	4.2 mV
2200 K	2.5 mV
1922 K	1.4 mV
1644 K	0.7 mV

We may test the basic[10] validity of the Stefan-Boltzmann law by finding the ratio of temperatures for any two temperature values in this table, raising that ratio to the fourth power, and seeing if the millivolt output signals for those same two temperatures match the new ratio. The operating theory here is that increases in target temperature will produce fourth-power increases in sensor temperature rise, since the sensor's temperature rise should be a direct function of radiation *power* impinging on it.

For example, if we were to take 4144 K and 3033 K as our two test temperatures, we find that the ratio of these two temperature values is 1.3663. Raising this ratio to the fourth power gives us 3.485 for a ratio of millivolt values. Multiplying the 3033 K millivoltage value of 9.9 mV by 3.485 gives us 34.5 mV, which is quite close to the value of 34.8 mV advertised by Honeywell.

If accuracy is not terribly important, and if the range of measured temperatures for the process is modest, we may take the millivoltage output of such a sensor and interpret it *linearly*. When used in this fashion, a non-contact pyrometer is often referred to as an *infrared thermocouple*, with the output voltage intended to connect directly to a thermocouple-input instrument such as an indicator, transmitter, recorder, or controller. An example of this usage is the OS-36 line of infrared thermocouples manufactured by Omega.

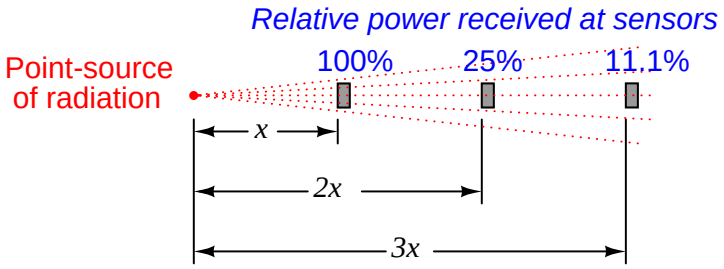
Infrared thermocouples are manufactured for a narrow range of temperature (most OS-36 models limited to a calibration span of 100 ^{o}F or less), their thermopiles designed to produce millivolt signals corresponding to a standard thermocouple type (T, J, K, etc.) over that narrow range.

[9] Later versions of the Radiamatic (dubbed the *Radiamatic II*) were more than just a bare thermopile and optical concentrator, containing electronic circuitry to output a linearized 4-20 mA signal representing target temperature.

[10] Comparing temperature ratios versus thermopile millivoltage ratios assumes linear thermocouple behavior, which we know is not exactly true. Even if the thermopile focal point temperatures precisely followed the ratios predicted by the Stefan-Boltzmann law, we would still expect some inconsistencies due to the non-linearities of thermocouple voltages. There will also be variations from predicted values due to shifts in radiated light frequencies, changes in emissivity factor, thermal losses within the sensing head, and other factors that refuse to remain constant over wide ranges of received radiation intensity. The lesson here is to not expect perfect agreement with theory!

A counter-intuitive characteristic of non-contact pyrometers is that their calibration does *not* depend on the distance separating the sensor from the target object[11]. This is counter-intuitive to anyone who has ever stood to an intense radiative heat source: standing in close proximity to a bonfire, for example, results in much hotter skin temperature than standing far away from it. Why wouldn't a non-contact pyrometer register cooler target temperatures when it was far away, given the fact that infrared radiation from the object spreads out with increased separation distance? The fact that an infrared pyrometer does not suffer from this limitation is good for our purposes in measuring temperature, but it doesn't seem to make sense at first.

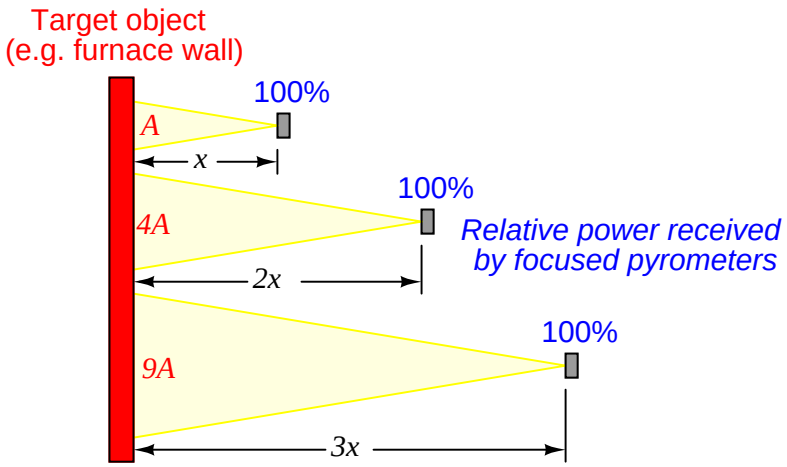
One key to understanding this paradox is to quantify the bonfire experience, where perceived temperature falls off with increased distance. In physics, this is known as the *inverse square law*: the intensity of radiation falling on an object from a point-source decreases with the *square* of the distance separating the radiation source from the object. Backing away to twice the distance from a bonfire results in a *four-fold* decrease in received infrared radiation; backing way to three times the distance results in a *nine-fold* decrease in received radiation. Placing a sensor at three integer distances (x, $2x$, and $3x$) from a radiation point-source results in relative power levels of 100%, 25% (one-quarter), and 11.1% (one-ninth) received at those locations, respectively:



This is a basic physical principle for all kinds of radiation, grounded in simple geometry. If we examine the radiation flux emanating from a point-source, we find that it must spread out as it travels in straight lines, and that the spreading-out happens at a rate defined by the square of the distance. An analogy for this phenomenon is to imagine a spherical latex balloon expanding as air is blown into it. The surface area of the balloon is proportional to the square of its radius. Likewise, the radiation flux emanating from a point-source spreads out in straight lines, in all directions, reaching a total area proportional to the square of the distance from the point (center). The total flux measured as a sphere will be the same no matter what the distance from the point-source, but the area it is divided over increases with the square of the distance, and so any object of fixed area backing away from a point-source of radiation encounters a smaller and smaller fraction of that flux.

[11] An important caveat to this rule is *so long as the target object completely fills the sensor's field of view (FOV)*. The reason for this caveat will become clear at the conclusion of the explanation.

If non-contact pyrometers really were "looking" at a point-source of infrared radiation, their signals *would* decrease with distance. The saving grace here is that non-contact pyrometers are focused-optic devices, with a definite *field of view*, and that field of view should always be completely filled by the target object. As distance between the pyrometer and the target object changes, the cone-shaped field of view covers a surface area on that object proportional to the square of the distance. Backing the pyrometer away to twice the distance increases the viewing area on the target object by a factor of four; backing away to three times the distance increases the viewing area nine times:



So, even though the inverse square law correctly declares that radiation emanating from the hot wall (which may be thought of as a collection of point-sources) decreases in intensity with the square of the distance, this attenuation is perfectly balanced by an increased viewing area of the pyrometer. Doubling the separation distance does result in the flux from any given area of the wall spreading out by a factor of four, but the pyrometer's view now covers four times as much area on the object as it did previously. The result is a perfect cancellation, with the pyrometer providing the exact same temperature measurement at *any* distance from the target where the target fills the entire field of view.

Perhaps the main disadvantage of non-contact temperature sensors is their inaccuracy. The emissivity factor (e) in the Stefan-Boltzmann equation varies with the composition of a substance, but beyond that there are several other factors (surface finish, shape, etc.) that affect the amount of radiation a sensor will receive from an object. For this reason, emissivity is not a very practical way to gauge the effectiveness of a non-contact pyrometer. Instead, a more comprehensive measure of an object's "thermal-optical measureability" is *emittance*.

A perfect emitter of thermal radiation is known as a *blackbody*. Emittance for a blackbody is unity (1), while emittance figures for any real object is a value between 1 and 0. The only certain way to know the emittance of an object is to test that object's thermal radiation at a known temperature. This assumes we have the ability to measure that object's temperature by direct contact, which of course renders void one of the major purposes of non-contact thermometry: to be able to measure an object's temperature without having to touch it. Not all hope is lost, though: all we have to do is obtain an emittance value for that object *one time*, and then we may calibrate any non-contact

pyrometer for that object's particular emittance so as to measure its temperature in the future without contact.

Beyond the issue of emittance, other idiosyncrasies plague non-contact pyrometers. Objects also have the ability to *reflect* and *transmit* radiation from other bodies, which taints the accuracy of any non-contact device sensing the radiation from that body. An example of the former is trying to measure the temperature of a silver mirror using an optical pyrometer: the radiation received by the pyrometer is mostly from other objects, merely *reflected* by the mirror. An example of the latter is trying to measure the temperature of a gas or a clear liquid, and instead primarily measuring the temperature of a solid object in the background (*through* the gas or liquid).

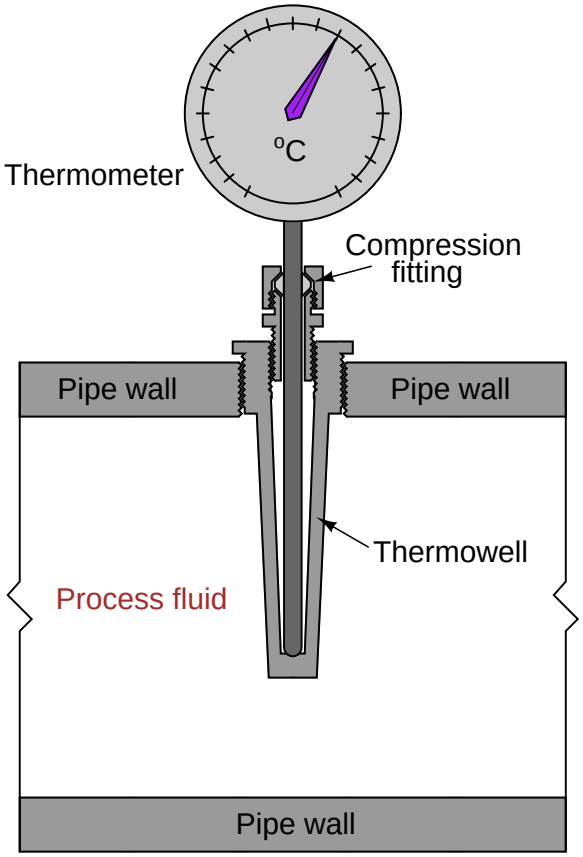
Nevertheless, non-contact pyrometers have been and will continue to be useful in specific applications where other, contact-based temperature measurement techniques are impractical.

A very useful application of non-contact sensor technology is *thermal imaging*, where a dense array of infrared radiation sensors provides a graphic display of objects in its view according to their temperatures. Each object shown on the digital display of a thermal imager is artificially colored in the display on a chromatic scale that varies with temperature, hot objects typically registering as red tones and cold objects typically registering as blue tones. Thermal imaging is very useful in the electric power distribution industry, where technicians can check power line insulators and other objects at elevated potential for "hot spots" without having to make physical contact with those objects. Thermal imaging is also useful in performing "energy audits" of buildings and other heated structures, providing a means of revealing points of heat escape through walls, windows, and roofs. In such applications, relative differences in temperature are often more important to detect than specific temperature values. "Hot spots" readily appear on a thermal imager display, and may give useful data on the test subject even in the absence of accurate temperature measurement at any one spot.

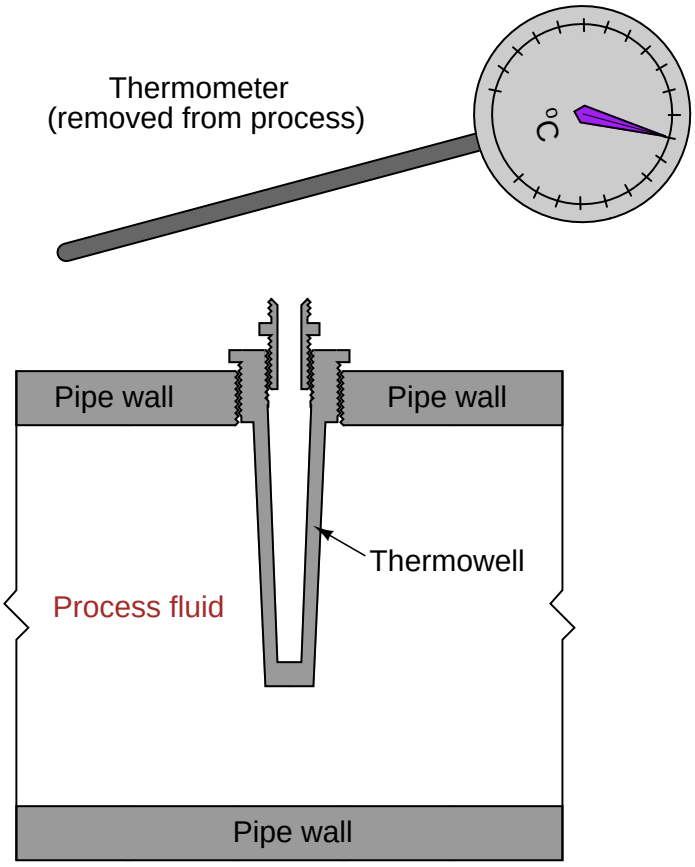
20.6 Temperature sensor accessories

One of the most important accessories for any temperature-sensing element is a pressure-tight sheath known as a *thermowell*. This may be thought of as a thermally conductive protrusion into a process vessel or pipe that allows a temperature-sensitive instrument to detect process temperature without opening a hole in the vessel or pipe. Thermowells are critically important for installations where the temperature element (RTD, thermocouple, thermometer, etc.) must be replaceable without de-pressurizing the process.

Thermowells may be made out of any material that is thermally conductive, pressure-tight, and not chemically reactive with the process. A simple diagram showing a thermowell in use with a temperature gauge is shown here:



If the temperature gauge is removed for maintenance or replacement, the thermowell maintains pressure integrity of the pipe (no process fluid leaking out, and no air leaking in):



Photographs of a real (stainless steel) thermowell are shown here, the left-hand photo showing the entire length of the thermowell, and the right-hand photo showing the end where the temperature-sensing device is inserted:

A photo of a complete RTD assembly (connection head, RTD, and thermowell) appears in the next photograph:

Another photo shows an RTD installed in a thermowell on the side of a commercial freezer, using a Rosemount model 3044C temperature transmitter to output a 4-20 mA signal to an operator display:

As useful as thermowells are, they are not without their caveats. First and foremost is the first-order time lag they add to the temperature measurement system by virtue of their mass and specific heat value. It should be intuitively obvious that one or more pounds of metal will not heat up and cool down as fast as a few ounces' worth of RTD or thermocouple, and therefore that the presence of a thermowell will decrease the response time of any temperature-sensing element.

A potential problem with thermowells is incorrect installation of the temperature-sensing element. The element *must* be inserted with full contact at the bottom of the thermowell's blind hole. If any air gap is allows to exist between the end of the temperature element and the bottom of the thermowell's hole, this will add a *second* time lag to the measurement system[12]. Some thermowells include a spring clip in the bottom of the blind hole to help maintain constant metal-to-metal contact between the sensing element and the thermowell wall.

[12] The air gap acts as a thermal *resistance* while the mass of the element itself acts as a thermal *capacitance*. Thus, the inclusion of an air gap forms a thermal "RC time constant" delay network secondary to the thermal delay incurred by the thermowell.

20.7 Process/instrument suitability

The primary consideration for selecting a proper temperature sensing element for any application is the expected temperature range. Mechanical (bi-metal) and filled-system temperature sensors are limited to relatively low process temperatures, and cannot relay signals very far from the point of measurement.

Thermocouples are by far the most rugged and wide-ranging of the contact-type temperature sensors. Accuracies vary with thermocouple type and installation quality.

RTDs are more fragile than thermocouples, but they require no reference compensation and are inherently more linear.

Optical sensors lack the ability to measure temperature of fluids inside vessels unless a transparent window is provided in the vessel for light emissions to reach the sensor. Otherwise, the best an optical sensor can do is report the skin temperature of a vessel. For monitoring surface temperatures of solid objects, especially objects that would be impractical or even dangerous to contact (e.g. electrical insulators on high-voltage power lines), optical sensors are the only appropriate solution.

Chemical reactivity is a concern for contact-type sensors. If the sensing element is held inside a thermowell, that thermowell must be selected for minimum reaction with the process fluid(s). Bare thermocouples are particularly vulnerable to chemical reactions given the nature of most thermocouple metals (iron, nickel, copper, etc.), and must be carefully chosen for the particular process chemistry to avoid reliability problems later.

References

Beckerath, Alexander von; Eberlein, Anselm; Julien, Hermann; Kersten, Peter; and Kreutzer, Jochem, *WIKA-Handbook, Pressure and Temperature Measurement*, WIKA Alexander Wiegand GmbH & Co., Klingenberg, Germany, 1995.

Darling, Charles Robert, *Pyrometry – A Practical Treatise on the Measurement of High Temperatures*, E. & F.N. Spon, Ltd, London, 1911.

Fribance, Austin E., *Industrial Instrumentation Fundamentals*, McGraw-Hill Book Company, New York, NY, 1962.

Irwin, J. David, *The Industrial Electronics Handbook*, CRC Press, Boca Raton, FL, 1997.

Kallen, Howard P., *Handbook of Instrumentation and Controls*, McGraw-Hill Book Company, Inc., New York, NY, 1961.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Measurement and Analysis Volume I*, Fourth Edition, CRC Press, New York, NY, 2003.

"Model 444 Alphaline Temperature Transmitters", Document 00809-0100-4263, Revision AA, Rosemount, Inc., 1998

"Radiamatic Detectors and Accessories", Specification document 23-75-03-03, Honeywell, Inc., Fort Washington, PA, 1992.

"Temperature - Electromotive Force (EMF) Tables for Standardized Thermocouples", Pyromation, Inc.

"Temperature Measurement – Thermocouples", ISA-MC96.1-1982, Instrument Society of America, Research Triangle Park, NC, 1982.

Chapter 21

Continuous fluid flow measurement

The measurement of fluid flow is arguably the single most complex type of process variable measurement in all of industrial instrumentation[1]. Not only is there a bewildering array of technologies one might use to measure fluid flow – each one with its own limitations and idiosyncrasies – but the very nature of the variable itself lacks a singular definition. "Flow" may refer to volumetric flow (the number of fluid *volumes* passing by per unit time), mass flow (the number of fluid mass units passing by per unit time), or even *standardized* volumetric flow (the number of gas volumes flowing, supposing different pressure and temperature values than what the actual process line operates at). Flowmeters configured to work with gas or vapor flows often are unusable on liquid flows. The dynamic properties of the fluids themselves change with flow rates. Most flow measurement technologies cannot achieve respectable measurement linearity from the maximum rated flow all the way to zero flow, no matter how well matched they might be to the process application.

Furthermore, the performance of most flowmeter technologies critically depends on proper installation. One cannot simply hang a flowmeter at any location in a piping system and expect it to function as designed. This is a constant source of friction between piping (mechanical) engineers and instrumentation (controls) engineers on large industrial projects. What might be considered excellent piping layout from the perspective of process equipment function and economy is often poor (at best) for good flow measurement, and visa-versa. In many cases the flowmeter equipment gets installed improperly and the instrument technicians have to deal with the resulting measurement problems during process unit start-up.

Even after a flowmeter has been properly selected for the process application and properly installed in the piping, problems may arise due to changes in process fluid properties (density, viscosity, conductivity), or the presence of impurities in the process fluid. Flowmeters are also subject to far more "wear and tear" than most other primary sensing elements, given the fact that a flowmeter's sensing element(s) must lie directly in the path of potentially abrasive fluid streams.

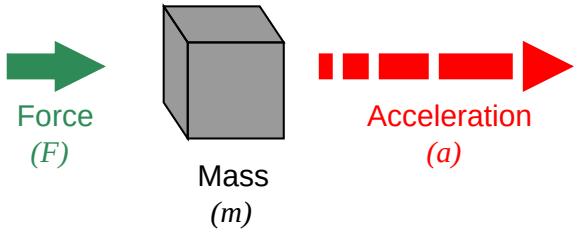
Given all these complications, it is imperative for instrumentation professionals to understand the complexities of flow measurement. What matters most is that you thoroughly understand the *physical principles* upon which each flowmeter depends. If the "first principles" of each technology are understood, the appropriate applications and potential problems become much easier

[1] Analytical (chemical composition) measurement is undeniably more complex and diverse than flow measurement, but analytical measurement covers a great deal of specific measurement types. As a *single* process variable, flow measurement is probably the most complex.

to recognize.

21.1 Pressure-based flowmeters

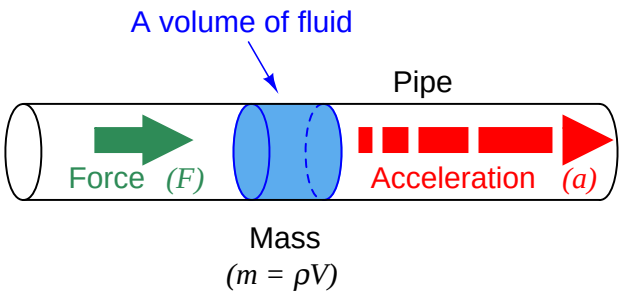
All masses require force to accelerate (we can also think of this in terms of the mass generating a reaction force as a result of being accelerated). This is quantitatively expressed by Newton's Second Law of Motion:



Newton's Second Law formula

$F = ma$

All fluids possess mass, and therefore require force to accelerate just like solid masses. If we consider a quantity of fluid confined inside a pipe[2], with that fluid quantity having a mass equal to its volume multiplied by its mass density ($m = \rho V$, where ρ is the fluid's mass per unit volume), the force required to accelerate that fluid "plug" would be calculated just the same as for a solid mass:



Newton's Second Law formula

$F = ma$ $F = \rho Va$

[2]Sometimes referred to as a *plug* of fluid.

Since this accelerating force is applied on the cross-sectional area of the fluid plug, we may express it as a *pressure*, the definition of pressure being force per unit area:

$$F = \rho V a$$

$$\frac{F}{A} = \rho \frac{V}{A} a$$

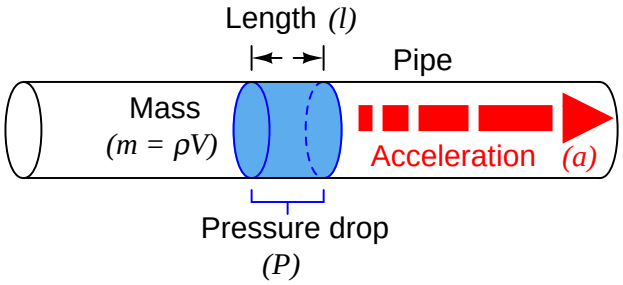
$$P = \rho \frac{V}{A} a$$

Since the rules of algebra required we divide *both* sides of the force equation by area, it left us with a fraction of volume over area ($\frac{V}{A}$) on the right-hand side of the equation. This fraction has a physical meaning, since we know the volume of a cylinder divided by the area of its circular face is simply the length of that cylinder:

$$P = \rho \frac{V}{A} a$$

$$P = \rho l a$$

When we apply this to the illustration of the fluid mass, it makes sense: the pressure described by the equation is actually a *differential*[3] pressure drop from one side of the fluid mass to the other, with the length variable (l) describing the spacing between the differential pressure ports:

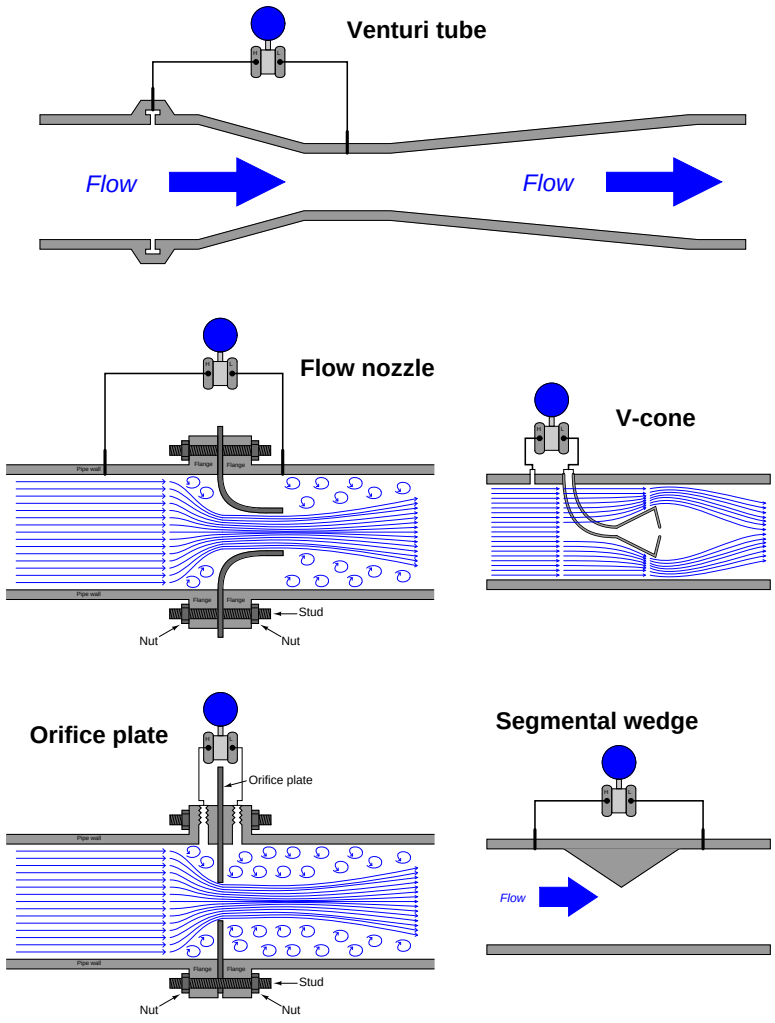


This tells us we can accelerate a "plug" of fluid by applying a difference of pressure across its length. The amount of pressure we apply will be in direct proportion to the density of the fluid and its rate of acceleration. Conversely, we may measure a fluid's rate of acceleration by measuring the pressure developed across a distance over which it accelerates.

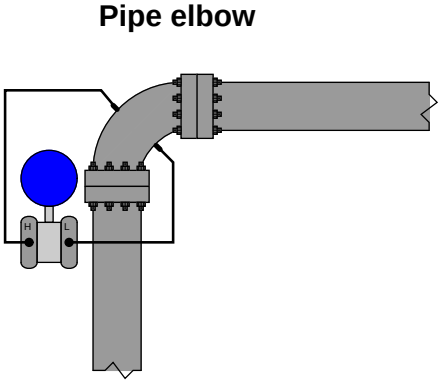
We may easily force a fluid to accelerate by altering its natural flow path. The difference of pressure generated by this acceleration will indirectly indicate the rate of acceleration. Since the acceleration we see from a change in flow path is a direct function of how fast the fluid was originally moving, the acceleration (and therefore the pressure drop) indirectly indicates fluid flow rate.

[3] What really matters in Newton's Second Law equation is the *resultant* force causing the acceleration. This is the vector sum of all forces acting on the mass. Likewise, what really matters in this scenario is the *resultant* pressure acting on the fluid plug, and this resultant pressure is the difference of pressure between one face of the plug and the other, since those two pressures impart two forces on the fluid mass in direct opposition to each other.

A very common way to cause linear acceleration in a moving fluid is to pass the fluid through a constriction in the pipe, thereby increasing its velocity (remember that the definition of acceleration is a change in velocity). The following illustrations show several devices used to linearly accelerate moving fluids when placed in pipes, with differential pressure transmitters connected to measure the pressure drop resulting from this acceleration:

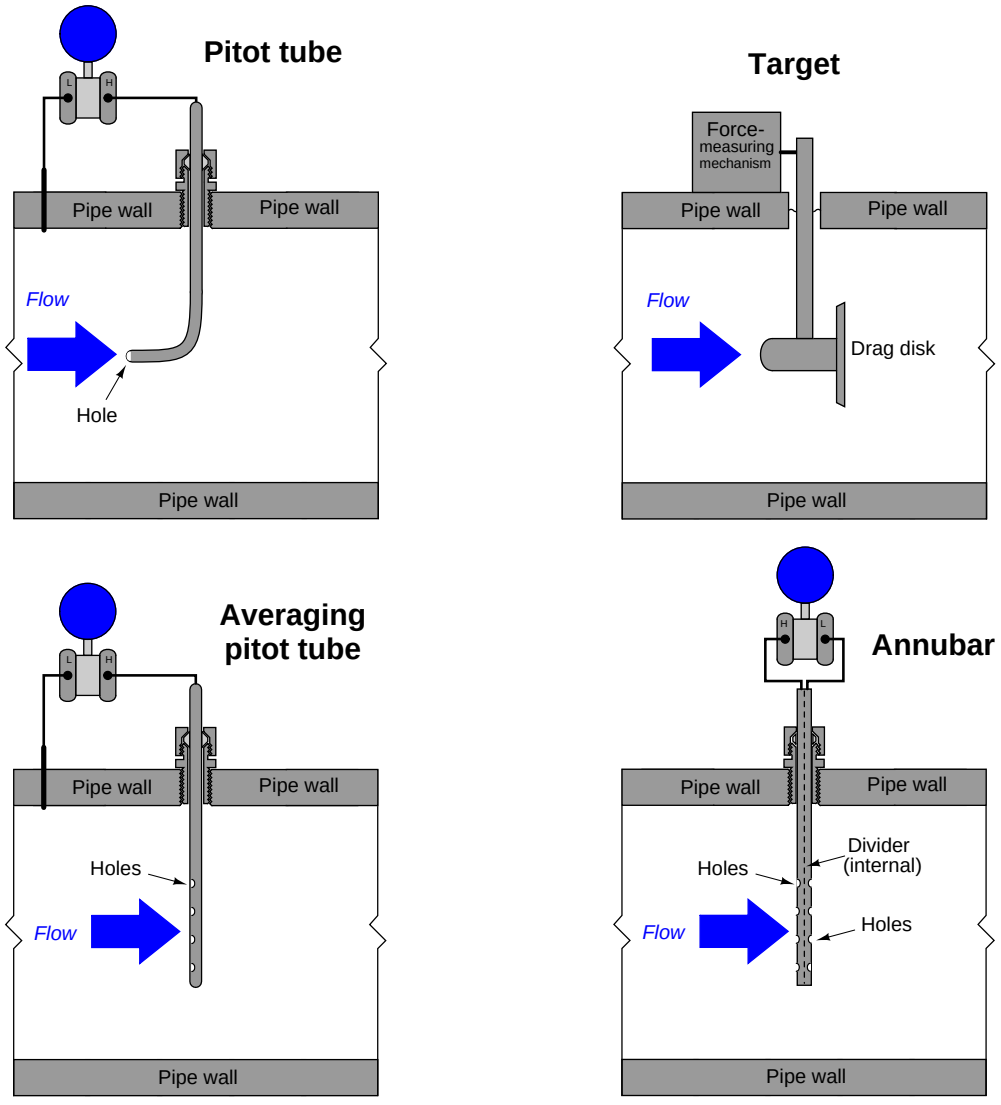


Another way we may accelerate a fluid is to force it to turn a corner through a pipe fitting called an *elbow*. This will generate radial acceleration, causing a pressure difference between the outside and inside of the elbow which may be measured by a differential pressure transmitter:



The pressure tap located on the outside of the elbow's turn registers a greater pressure than the tap located on the inside of the elbow's turn, due to the inertial force of the fluid's mass being "flung" to the outside of the turn as it rounds the corner.

Yet another way to cause a change in fluid velocity is to force it to *decelerate* by bringing a portion of it to a full stop. The pressure generated by this deceleration (called the *stagnation pressure*) tells us how fast it was originally flowing. A few devices working on this principle are shown here:



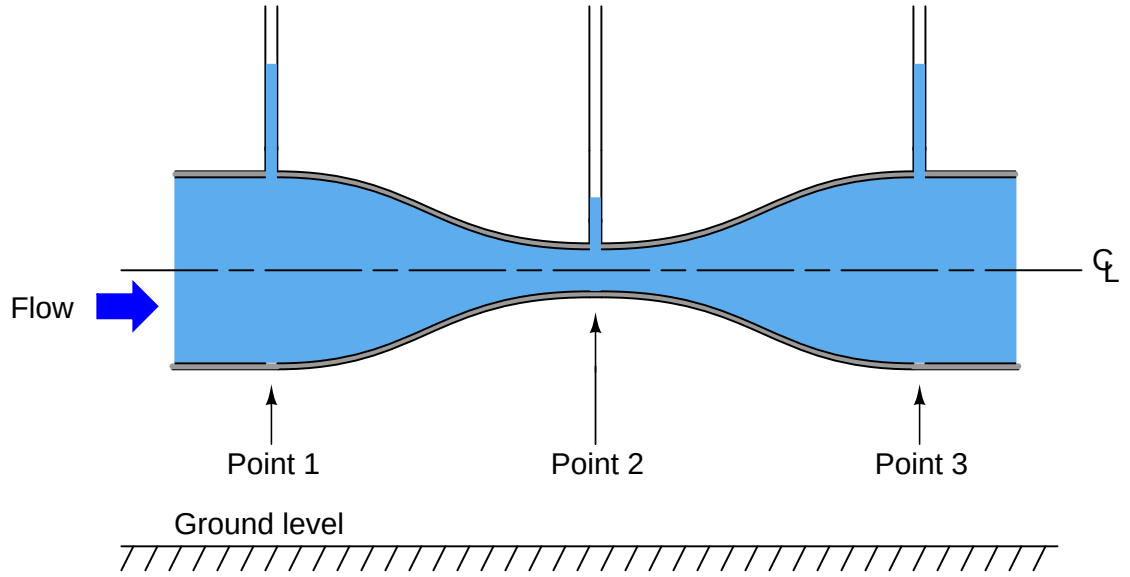
The following subsections in this flow measurement chapter explore different primary sensing elements (PSE's) used to generate differential pressure in a moving fluid stream. Despite their very different designs, they all operate on the same fundamental principle: causing a fluid to accelerate or decelerate by forcing a change in its flow path, and thus generating a measurable pressure difference. The following subsection will introduce a device called a *venturi tube* used to measure fluid flow

rates, and derive mathematical relationships between fluid pressure and flow rate starting from basic physical conservation laws.

21.1.1 Venturi tubes and basic principles

The standard "textbook example" flow element used to create a pressure change by accelerating a fluid stream is the *venturi tube*: a pipe purposefully narrowed to create a region of low pressure. As shown previously, venturi tubes are not the only structure capable of producing a flow-dependent pressure drop. You should keep this in mind as we proceed to derive equations relating flow rate with pressure change: although the venturi tube is the canonical form, the exact same mathematical relationship applies to all flow elements generating a pressure drop by accelerating fluid, including orifice plates, flow nozzles, V-cones, segmental wedges, pipe elbows, pitot tubes, etc.

If the fluid going through the venturi tube is a liquid under relatively low pressure, we may vividly show the pressure at different points in the tube by means of *piezometers*, which are transparent tubes allowing us to view liquid column heights. The greater the height of liquid column in the piezometer, the greater the pressure at that point in the flowstream:



As indicated by the piezometer liquid heights, pressure at the constriction (point 2) is the least, while pressures at the wide portions of the venturi tube (points 1 and 3) are the greatest. This is a counter-intuitive result, but it has a firm grounding in the physics of mass and energy conservation. If we assume no energy is added (by a pump) or lost (due to friction) as fluid travels through this pipe, then the Law of Energy Conservation describes a situation where the fluid's energy must remain constant at all points in the pipe as it travels through. If we assume no fluid joins this flowstream from another pipe, or is lost from this pipe through any leaks, then the Law of Mass Conservation describes a situation where the fluid's mass flow rate must remain constant at all points in the pipe as it travels through.

So long as fluid density remains fairly constant[4], fluid velocity must increase as the cross-sectional

[4] This is a very sound assumption for liquids, and a fair assumption for gases when pressure changes through the venturi tube are modest.

area of the pipe decreases, as described by the Law of Continuity (see section 2.9.10 on page 117 for more details on this concept):

$$A_1 \overline{v_1} = A_2 \overline{v_2}$$

Rearranging variables in this equation to place velocities in terms of areas, we get the following result:

$$\frac{\overline{v_2}}{\overline{v_1}} = \frac{A_1}{A_2}$$

This equation tells us that the ratio of fluid velocity between the narrow throat (point 2) and the wide mouth (point 1) of the pipe is the same ratio as the mouth's area to the throat's area. So, if the mouth of the pipe had an area 5 times as great as the area of the throat, then we would expect the fluid velocity in the throat to be 5 times as great as the velocity at the mouth. Simply put, the narrow throat causes the fluid to accelerate from a lower velocity to a higher velocity.

We know from our study of energy in physics that kinetic energy is proportional to the square of a mass's velocity ($E_k = \frac{1}{2}mv^2$). If we know the fluid molecules increase velocity as they travel through the venturi tube's throat, we may safely conclude that those molecules' kinetic energies must increase as well. However, we also know that the total energy at any point in the fluid stream must remain constant, because no energy is added to or taken away from the stream in this simple fluid system[5]. Therefore, if kinetic energy increases at the throat, potential energy must correspondingly decrease to keep the total amount of energy constant at any point in the fluid.

Potential energy may be manifest as height above ground, or as pressure in a fluid system. Since this venturi tube is level with the ground, there cannot be a height change to account for a change in potential energy. Therefore, there *must* be a change of pressure (P) as the fluid travels through the venturi throat. The Laws of Mass and Energy Conservation invariably lead us to this conclusion: fluid pressure must decrease as it travels through the narrow throat of the venturi tube[6].

Conservation of energy at different points in a fluid stream is neatly expressed in *Bernoulli's Equation* as a constant sum of elevation, pressure, and velocity "heads" (see section 2.9.12 on page 120 for more details on this concept):

$$z_1 \rho g + \frac{v_1^2 \rho}{2} + P_1 = z_2 \rho g + \frac{v_2^2 \rho}{2} + P_2$$

Where,

z = Height of fluid (from a common reference point, usually ground level)
ρ = Mass density of fluid
g = Acceleration of gravity
v = Velocity of fluid
P = Pressure of fluid

[5] One of the simplifying assumptions we make in this derivation is that friction plays no significant role in the fluid's behavior as it moves through the venturi tube. In truth, no industrial fluid flow is totally frictionless (especially through more primitive flow elements such as orifice plates), and so our "theoretical" equations must be adjusted a bit to match real life.

[6] To see a graphical relationship between fluid acceleration and fluid pressures in a venturi tube, examine the illustration found on page 1739.

We will use Bernoulli's equation to develop a precise mathematical relationship between pressure and flow rate in a venturi tube. To simplify our task, we will hold to the following assumptions for our venturi tube system:

- No energy lost or gained in the venturi tube (all energy is conserved)
- No mass lost or gained in the venturi tube (all mass is conserved)
- Fluid is incompressible
- Venturi tube centerline is level (no height changes to consider)

Applying the last two assumptions to Bernoulli's equation, we see that the "elevation head" term drops out of both sides, since z, ρ, and g are equal at all points in the system:

$$\frac{v_1^2\rho}{2} + P_1 = \frac{v_2^2\rho}{2} + P_2$$

Now we will algebraically re-arrange this equation to show pressures at points 1 and 2 in terms of velocities at points 1 and 2:

$$\frac{v_2^2\rho}{2} - \frac{v_1^2\rho}{2} = P_1 - P_2$$

Factoring $\frac{\rho}{2}$ out of the velocity head terms:

$$\frac{\rho}{2}(v_2^2 - v_1^2) = P_1 - P_2$$

The Continuity equation shows us the relationship between velocities v_1 and v_2 and the areas at those points in the venturi tube, assuming constant density (ρ):

$$A_1 v_1 = A_2 v_2$$

Specifically, we need to re-arrange this equation to define v_1 in terms of v_2 so we may substitute into Bernoulli's equation:

$$v_1 = \left(\frac{A_2}{A_1}\right) v_2$$

Performing the algebraic substitution:

$$\frac{\rho}{2}(v_2^2 - \left[\left(\frac{A_2}{A_1}\right) v_2\right]^2) = P_1 - P_2$$

Distributing the "square" power:

$$\frac{\rho}{2}(v_2^2 - \left(\frac{A_2}{A_1}\right)^2 v_2^2) = P_1 - P_2$$

Factoring v_2^2 out of the outer parentheses set:

$$\frac{\rho v_2^2}{2}(1 - \left(\frac{A_2}{A_1}\right)^2) = P_1 - P_2$$

Solving for v_2, step by step:

$$\frac{\rho v_2^2}{2} = \left(\frac{1}{1 - \left(\frac{A_2}{A_1}\right)^2}\right)(P_1 - P_2)$$

$$\rho v_2^2 = 2\left(\frac{1}{1 - \left(\frac{A_2}{A_1}\right)^2}\right)(P_1 - P_2)$$

$$v_2^2 = 2\left(\frac{1}{1 - \left(\frac{A_2}{A_1}\right)^2}\right)\left(\frac{P_1 - P_2}{\rho}\right)$$

$$v_2 = \sqrt{2}\frac{1}{\sqrt{1 - \left(\frac{A_2}{A_1}\right)^2}}\sqrt{\frac{P_1 - P_2}{\rho}}$$

The result shows us how to solve for fluid velocity at the venturi throat (v_2) based on a difference of pressure measured between the mouth and the throat ($P_1 - P_2$). We are only one step away from a volumetric flow equation here, and that is to convert velocity (v) into flow rate (Q). Velocity is expressed in units of length per time (feet or meters per second or minute), while volumetric flow is expressed in units of volume per time (cubic feet or cubic meters per second or minute). Simply multiplying throat velocity (v_2) by throat area (A_2) will give us the result we seek:

General flow/area/velocity relationship:

$$Q = Av$$

Equation for throat velocity:

$$v_2 = \sqrt{2}\frac{1}{\sqrt{1 - \left(\frac{A_2}{A_1}\right)^2}}\sqrt{\frac{P_1 - P_2}{\rho}}$$

Multiplying both sides of the equation by throat area:

$$A_2 v_2 = \sqrt{2}A_2\frac{1}{\sqrt{1 - \left(\frac{A_2}{A_1}\right)^2}}\sqrt{\frac{P_1 - P_2}{\rho}}$$

Now we have an equation solving for volumetric flow in terms of pressures and areas:

$$Q = \sqrt{2}A_2 \frac{1}{\sqrt{1 - \left(\frac{A_2}{A_1}\right)^2}} \sqrt{\frac{P_1 - P_2}{\rho}}$$

Please note how many constants we have in this equation. For any given venturi tube, the mouth and throat areas (A_1 and A_2) will be fixed. This means nearly half the variables found within this rather long equation are actually constant for any given venturi tube, and therefore do not change with pressure, density, or flow rate. Knowing this, we may re-write the equation as a simple proportionality:

$$Q \propto \sqrt{\frac{P_1 - P_2}{\rho}}$$

To make this a more precise mathematical statement, we may insert a *constant of proportionality* (k) and once more have a true equation to work with:

$$Q = k\sqrt{\frac{P_1 - P_2}{\rho}}$$

21.1.2 Volumetric flow calculations

As we saw in the previous subsection, we may derive a relatively simple equation for predicting flow through a fluid-accelerating element given the pressure drop generated by that element and the density of the fluid flowing through it:

$$Q = k\sqrt{\frac{P_1 - P_2}{\rho}}$$

This equation is a simplified version of the one derived from the physical construction of a venturi tube:

$$Q = \sqrt{2}A_2 \frac{1}{\sqrt{1 - \left(\frac{A_2}{A_1}\right)^2}} \sqrt{\frac{P_1 - P_2}{\rho}}$$

As you can see, the constant of proportionality (k) shown in the simpler equation is nothing more than a condensation of the first half of the longer equation: k represents the geometry of the venturi tube. If we define k by the mouth and throat areas (A_1, A_2) of any particular venturi tube, we must be very careful to express the pressures and densities in compatible units of measurement. For example, with k strictly defined by flow element geometry (tube areas measured in square *feet*), the calculated flow rate (Q) must be in units of cubic *feet* per second, the pressure values P_1 and P_2 must be in units of pounds per square *foot*, and mass density must be in units of *slugs* per cubic *foot*. We cannot arbitrarily choose different units of measurement for these variables, because the units must "agree" with one another. If we wish to use more convenient units of measurement such as inches of water column for pressure and specific gravity (unitless) for density, the original (longer) equation simply will not work.

However, if we happen to know the differential pressure produced by any particular flow element tube with any particular fluid density at a specified flow rate (real-life conditions), we may *calculate* a value for k in the short equation that makes all those measurements "agree" with one another. In other words, we may use the constant of proportionality (k) as a *unit-of-measurement correction factor* as well as a definition of element geometry. This is a useful property of all proportionalities: simply plug in values (expressed in any unit of measurement) determined by physical experiment and solve for the proportionality constant's value to satisfy the expression as an equation. If we do this, the value we arrive at for k will automatically compensate for whatever units of measurement we arbitrarily choose for pressure and density.

For example, if we know a particular orifice plate develops 45 inches of water column differential pressure at a flow rate of 180 gallons per minute of water (specific gravity = 1), we may plug these values into the equation and solve for k:

$$Q = k\sqrt{\frac{P_1 - P_2}{\rho}}$$

$$180 = k\sqrt{\frac{45}{1}}$$

$$k = \frac{180}{\sqrt{\frac{45}{1}}} = 26.83$$

Now we possess a value for k (26.83) that yields a flow rate in units of "gallons per minute" given differential pressure in units of "inches of water column" and density expressed as a specific gravity for this particular orifice plate. From the known fact of all accelerating flow elements' behavior (flow rate proportional to the square root of pressure divided by density) and from a set of values experimentally determined for this particular orifice plate, we now have an equation useful for calculating flow rate given any set of pressure and density values we may happen to encounter with this particular orifice plate:

$$\left[\frac{\text{gal}}{\text{m}}\right] = 26.83\sqrt{\frac{[\text{"W.C.}]}{\text{Specific gravity}}}$$

Applying our new equation to this orifice plate, we see that 60 inches of water column differential pressure generated by a flow of water (specific gravity = 1) equates to 207.8 gallons per minute of flow:

$$Q = 26.83\sqrt{\frac{60}{1}}$$

$$Q = 207.8 \text{ GPM}$$

If we were to measure 110 inches of water column differential pressure across this orifice plate as gasoline (specific gravity = 0.657) flowed through it, we could calculate the flow rate to be 347 gallons per minute:

$$Q = 26.83\sqrt{\frac{110}{0.657}}$$

$$Q = 347 \text{ GPM}$$

Suppose, though, we wished to have an equation for calculating the flow rate through this same orifice plate given pressure and density data in different units (say, kPa instead of inches water column, and kilograms per cubic meter instead of specific gravity). In order to do this, we would need to re-calculate the constant of proportionality (k) to accommodate those new units of

measurement. To do this, all we would need is a single set of experimental data for the orifice plate relating flow in GPM, pressure in kPa, and density in kg/m^3.

Applying this to our original data where a water flow rate of 180 GPM resulted in a pressure drop of 45 inches water column, we could convert the pressure drop of 45 "W.C. into 11.21 kPa and express the density as 1000 kg/m^3 to solve for a new value of k:

$$Q = k\sqrt{\frac{P_1 - P_2}{\rho}}$$

$$180 = k\sqrt{\frac{11.21}{1000}}$$

$$k = \frac{180}{\sqrt{\frac{11.21}{1000}}} = 1700$$

Nothing about the orifice plate's geometry has changed from before, only the units of measurement we have chosen to work with. Now we possess a value for k (1700) for the same orifice plate yielding a flow rate in units of "gallons per minute" given differential pressure in units of "kilopascals" and density in units of "kilograms per cubic meter."

$$\left[\frac{\text{gal}}{\text{m}}\right] = 1700\sqrt{\frac{[\text{kPa}]}{\text{kg/m}^3}}$$

If we were to be given a pressure drop in kPa and a fluid density in kg/m^3 for this orifice plate, we could calculate the corresponding flow rate (in GPM) with our new value of k (1700) just as easily as we could with the old value of k (26.83) given pressure in "W.C. and specific gravity.

21.1.3 Mass flow calculations

Measurements of *mass* flow are preferred over measurements of *volumetric* flow in process applications where mass balance (monitoring the rates of mass entry and exit for a process) is important. Whereas volumetric flow measurements express the fluid flow rate in such terms as *gallons per minute* or *cubic meters per second*, mass flow measurements always express fluid flow rate in terms of actual mass units over time, such *pounds (mass) per second* or *kilograms per minute*. Applications for mass flow measurement include custody transfer (where a fluid product is bought or sold by its mass), chemical reaction processes (where the mass flow rates of reactants must be maintained in precise proportion in order for the desired chemical reactions to occur), and steam boiler control systems (where the out-flow of vaporous steam must be balanced by an equivalent in-flow of liquid water to the boiler – here, volumetric comparisons of steam and water flow would be useless because one cubic foot of steam is certainly not the same number of H_2O molecules as one cubic foot of water).

If we wish to calculate *mass* flow instead of volumetric flow, the equation does not change much. The relationship between volume (V) and mass (m) for a sample of fluid is its mass density (ρ):

$$\rho = \frac{m}{V}$$

Similarly, the relationship between a volumetric *flow rate* (Q) and a mass *flow rate* (W) is also the fluid's mass density (ρ):

$$\rho = \frac{W}{Q}$$

Solving for W in this equation leads us to a product of volumetric flow rate and mass density:

$$W = \rho Q$$

A quick dimensional analysis check using common metric units confirms this fact. A mass flow rate in kilograms per second will be obtained by multiplying a mass density in kilograms per cubic meter by a volumetric flow rate in cubic meters per second:

$$\left[\frac{\text{kg}}{\text{s}}\right] = \left[\frac{\text{kg}}{\text{m}^3}\right]\left[\frac{\text{m}^3}{\text{s}}\right]$$

Therefore, all we have to do to turn our general volumetric flow equation into a mass flow equation is multiply both sides by fluid density (ρ):

$$Q = k\sqrt{\frac{P_1 - P_2}{\rho}}$$

$$\rho Q = k\rho\sqrt{\frac{P_1 - P_2}{\rho}}$$

$$W = k\rho\sqrt{\frac{P_1 - P_2}{\rho}}$$

It is generally considered "inelegant" to show the same variable more than once in an equation if it is not necessary, so let's try to consolidate the two densities (ρ) using algebra. First, we may write ρ as the product of two square-roots:

$$W = k\sqrt{\rho}\sqrt{\rho}\sqrt{\frac{P_1 - P_2}{\rho}}$$

Next, we will break up the last radical into a quotient of two separate square roots:

$$W = k\sqrt{\rho}\sqrt{\rho}\frac{\sqrt{P_1 - P_2}}{\sqrt{\rho}}$$

Now we see how one of the square-rooted ρ terms cancels out the one in the denominator of the fraction:

$$W = k\sqrt{\rho}\sqrt{P_1 - P_2}$$

It is also considered "inelegant" to have multiple radicands in an equation where one will suffice, so we will re-write our equation for aesthetic improvement[7]:

$$W = k\sqrt{\rho(P_1 - P_2)}$$

As with the volumetric flow equation, all we need in order to arrive at a suitable k value for any particular flow element is a set of values taken from that real element in service, expressed in whatever units of measurement we desire.

[7] This re-write is solidly grounded in the rules of algebra. We know that $\sqrt{a}\sqrt{b} = \sqrt{ab}$, which is what allows us to do the re-write.

For example, if we had a venturi tube generating a differential pressure of 2.30 kilo-Pascals (kPa) at a mass flow rate of 500 kilograms per minute of naphtha (a petroleum product having a density of 0.665 kilograms per liter), we could solve for the k value of this venturi tube as such:

$$W = k\sqrt{\rho(P_1 - P_2)}$$

$$500 = k\sqrt{(0.665)(2.3)}$$

$$k = \frac{500}{\sqrt{(0.665)(2.3)}}$$

$$k = 404$$

Now that we know a value of 404 for k will yield kilograms per minute of liquid flow through this venturi tube given pressure in kPa and density in kilograms per liter, we may readily predict the mass flow rate through this tube for any other pressure drop and fluid density we might happen to encounter. The value of 404 for k relates the disparate units of measurement for us:

$$\left[\frac{\text{kg}}{\text{m}}\right] = 404\sqrt{\left[\frac{\text{kg}}{\text{l}}\right] [\text{kPa}]}$$

As with volumetric flow calculations, the calculated value for k neatly accounts for any set of measurement units we may arbitrarily choose. The key is first knowing the proportional relationship between flow rate, pressure drop, and density. Once we combine that proportionality with a specific set of data experimentally gathered from a particular flow element, we have a true equation properly relating all the variables together in our chosen units of measurement.

If we happened to measure 6.1 kPa of differential pressure across this same venturi tube as it flowed sea water (density = 1.03 kilograms per liter), we could calculate the mass flow rate quite using the same equation (with the k factor of 404):

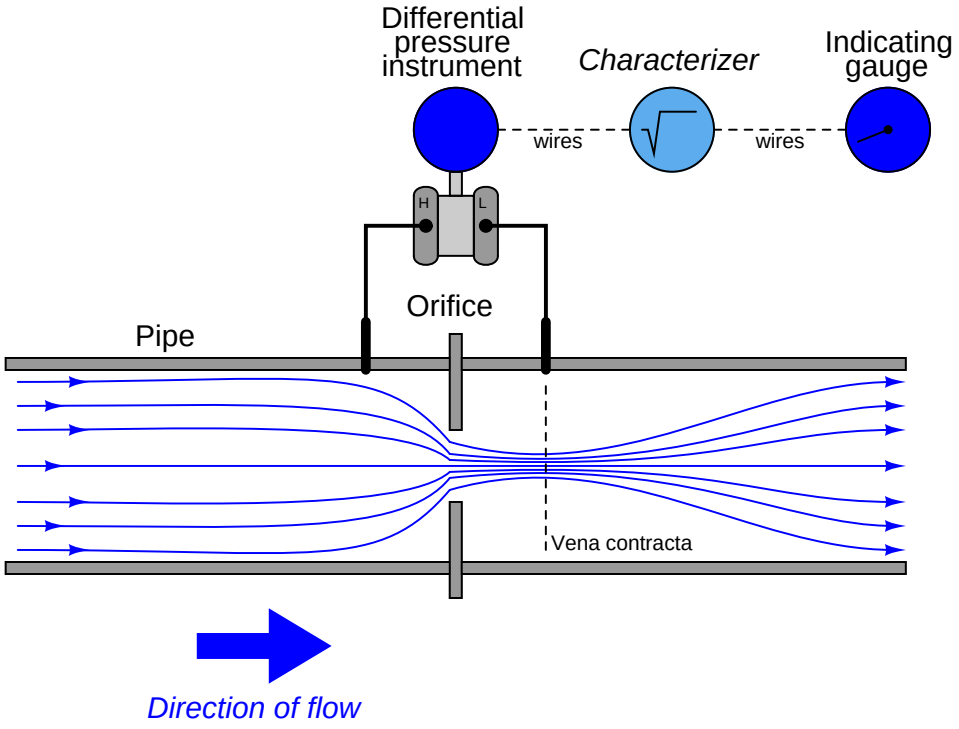
$$W = 404\sqrt{(1.03)(6.1)}$$

$$W = 1012\ \frac{\text{kg}}{\text{m}}$$

21.1.4 Square-root characterization

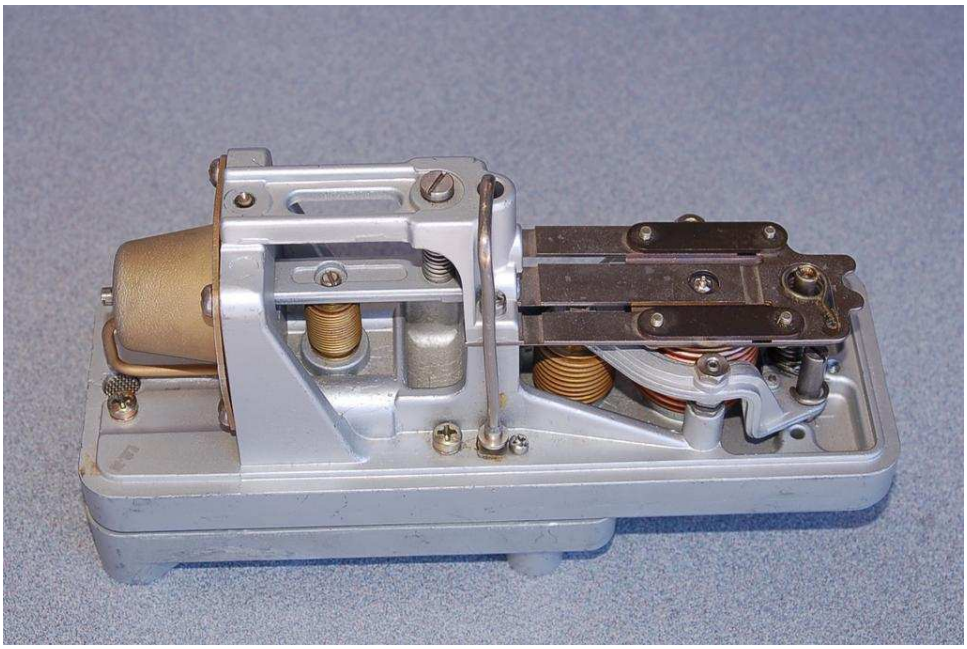
It should be apparent by now that the relationship between flow rate (whether it be volumetric or mass) and differential pressure for any fluid-accelerating flow element is non-linear: a doubling of flow rate will *not* result in a doubling of differential pressure. Rather, a doubling of flow rate will result in a *quadrupling* of differential pressure.

This quadratic relationship between flow and pressure drop due to fluid acceleration requires us to mathematically "condition" or "characterize" the pressure signal sensed by the differential pressure instrument in order to arrive at an expressed value for flow rate. The traditional solution to this problem was to incorporate a "square root" function relay between the transmitter and the flow indicator, as shown in the following diagram:



The modern solution to this problem is to incorporate square-root signal characterization either inside the transmitter or inside the receiving instrument (e.g. indicator, recorder, or controller).

In the days of pneumatic instrumentation, this square-root function was performed in a separate device called a *square root extractor*. The Foxboro corporation model 557 pneumatic square root extractor was a classic example of this technology[8]:



Pneumatic square root extraction relays approximated the square-root function by means of triangulated force or motion. In essence, they were *trigonometric* function relays, not square-root relays. However, for small angular motions, certain trigonometric functions were close enough to a square-root function that the relays were able to serve their purpose in characterizing the output signal of a pressure sensor to yield a signal representing flow rate.

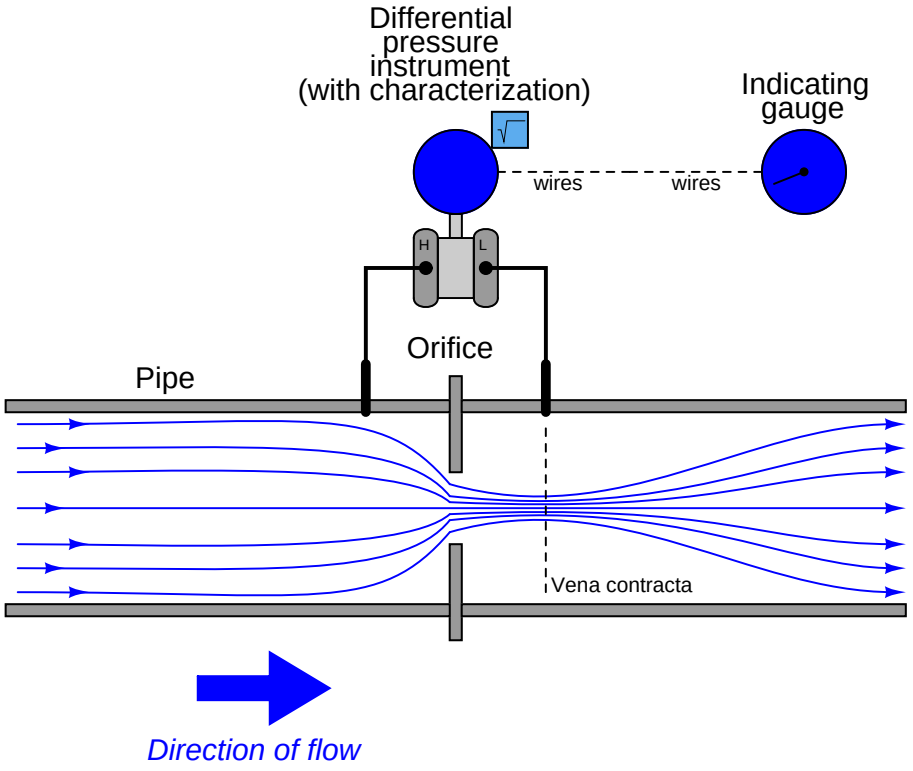
[8]Despite the impressive craftsmanship and engineering that went into the design of pneumatic square root extractors, their obsolescence is mourned by no one. These devices were notoriously difficult to set up and calibrate accurately, especially as they aged.

The following table shows the ideal response of a pneumatic square root relay:

Input signal	Input %	Output %	Output signal
3 PSI	0%	0%	3 PSI
4 PSI	8.33%	28.87%	6.464 PSI
5 PSI	16.67%	40.82%	7.899 PSI
6 PSI	25%	50%	9 PSI
7 PSI	33.33%	57.74%	9.928 PSI
8 PSI	41.67%	64.55%	10.75 PSI
9 PSI	50%	70.71%	11.49 PSI
10 PSI	58.33%	76.38%	12.17 PSI
11 PSI	66.67%	81.65%	12.80 PSI
12 PSI	75%	86.60%	13.39 PSI
13 PSI	83.33%	91.29%	13.95 PSI
14 PSI	91.67%	95.74%	14.49 PSI
15 PSI	100%	100%	15 PSI

As you can see from the table, the square-root relationship is most evident in comparing the input and output *percentage* values. For example, at an input signal pressure of 6 PSI (25%), the output signal percentage will be the square root of 25%, which is 50% ($0.5 = \sqrt{0.25}$) or 9 PSI as a pneumatic signal. At an input signal pressure of 10 PSI (58.33%), the output signal percentage will be 76.38%, because $0.7638 = \sqrt{0.5833}$, yielding an output signal pressure of 12.17 PSI.

Although analog electronic square-root relays have been built and used in industry for characterizing the output of 4-20 mA electronic transmitters, a far more common application of electronic square-root characterization is found in DP transmitters with the square-root function built in. This way, an external relay device is not necessary to characterize the DP transmitter's signal into a flow rate signal:



Using a characterized DP transmitter, any 4-20 mA sensing instrument connected to the transmitter's output wires will directly interpret the signal as flow rate rather than as pressure. A calibration table for such a DP transmitter (with an input range of 0 to 150 inches water column) is shown here:

Differential pressure	% of input span	Output %	Output signal
0 "W.C.	0%	0%	4 mA
37.5 "W.C.	25%	50%	12 mA
75 "W.C.	50%	70.71%	15.31 mA
112.5 "W.C.	75%	86.60%	17.86 mA
150 "W.C.	100%	100%	20 mA

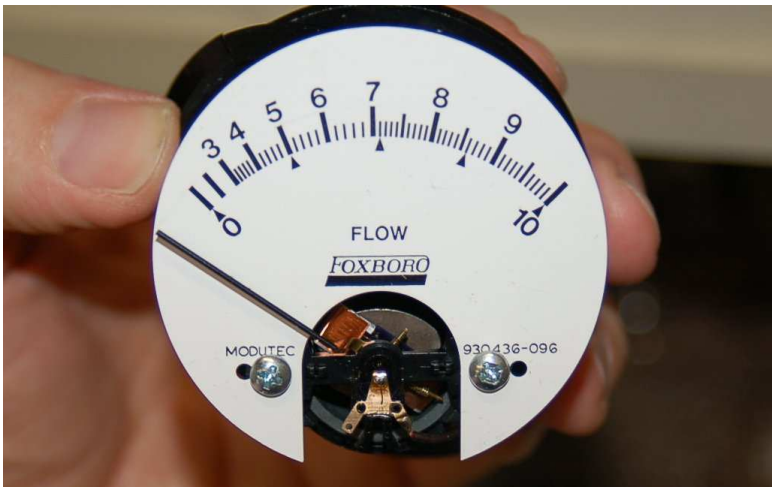
Once again, we see how the square-root relationship is most evident in comparing the input and output *percentages*. Note how the four sets of percentages in this table precisely match the same

four percentage sets in the pneumatic relay table: 0% input gives 0% output; 25% input gives 50% output, 50% input gives 70.71% output, etc.

An ingenious solution to the problem of square-root characterization, more commonly applied before the advent of DP transmitters with built-in characterization, is to use an indicating device with a square-root indicating scale. For example, the following photograph shows a 3-15 PSI "receiver gauge" designed to directly sense the output of a pneumatic DP transmitter:

Note how the gauge mechanism responds directly and linearly to a 3-15 PSI input signal range (note the "3 PSI" and "15 PSI" labels in small print at the extremes of the scale, and the linearly-spaced marks around the outside of the scale arc representing 1 PSI each), but how the flow markings (0 through 10 on the inside of the scale arc) are spaced in a non-linear fashion.

An electronic variation on this theme is to draw a square-root scale on the face of a meter movement driven by the 4-20 mA output signal of an electronic DP transmitter:



As with the square-root receiver gauge, the meter movement's response to the transmitter signal is linear (note the evenly-spaced, triangular marks on the bottom of the scale arc representing increments of 4 mA each), but the markings drawn on the top of the scale are spaced in a non-linear (square-root) fashion. This makes it possible for a human operator to read the scale in terms of (characterized) flow units. Instead of using complicated mechanisms or circuitry to characterize the transmitter's signal, a non-linear scale "performs the math" necessary to interpret flow.

A major disadvantage to the use of these non-linear indicator scales is that the transmitter signal itself remains un-characterized. Any other instrument receiving this un-characterized signal will either require its own square-root characterization or simply not interpret the signal in terms of flow at all. An un-characterized flow signal input to a process controller can cause loop instability at high flow rates, where small changes in actual flow rate result in huge changes in differential pressure sensed by the transmitter. A fair number of flow control loops operating without characterization have been installed in industrial applications (usually with square-root scales drawn on the face of the indicators, and square-root paper installed in chart recorders), but these loops are notorious for achieving good flow control at only one setpoint value. If the operator raises or lowers the setpoint value, the "gain" of the control loop changes thanks to the nonlinearities of the flow element, resulting in either under-responsive or over-responsive action from the controller.

Despite the limited practicality of non-linear indicating scales, they hold significant value as teaching tools. Closely examine the scales of both the receiver gauge and the 4-20 mA indicating meter, comparing the linear marks (one mark for every 1 PSI increment on the gauge, and one mark for every 4 mA increment on the meter), then compare what you see on the scales against figures in the tables provided earlier for characterized instruments (the square-root extractor and the characterized DP transmitter). Do you see the correspondence? Note how the points marked by the linear divisions match with points on the square-root scale in the exact same manner as the input percentage values in the characterizer tables correspond with the output (square-root) percentage values. At an input value of 25% (6 PSI on the receiver gauge, and the first non-zero

linear mark on the meter) match precisely with the 50% point on the square-root scale. Note also how a linear value of 50% (the half-way point on the needle's sweep for both the receiver gauge and the meter movement) points to just under 71% on each indicator's square-root scale. A few checks like this verify the fact that the square-root function is *encoded* in the spacing of the numbers on each instrument's non-linear scale.

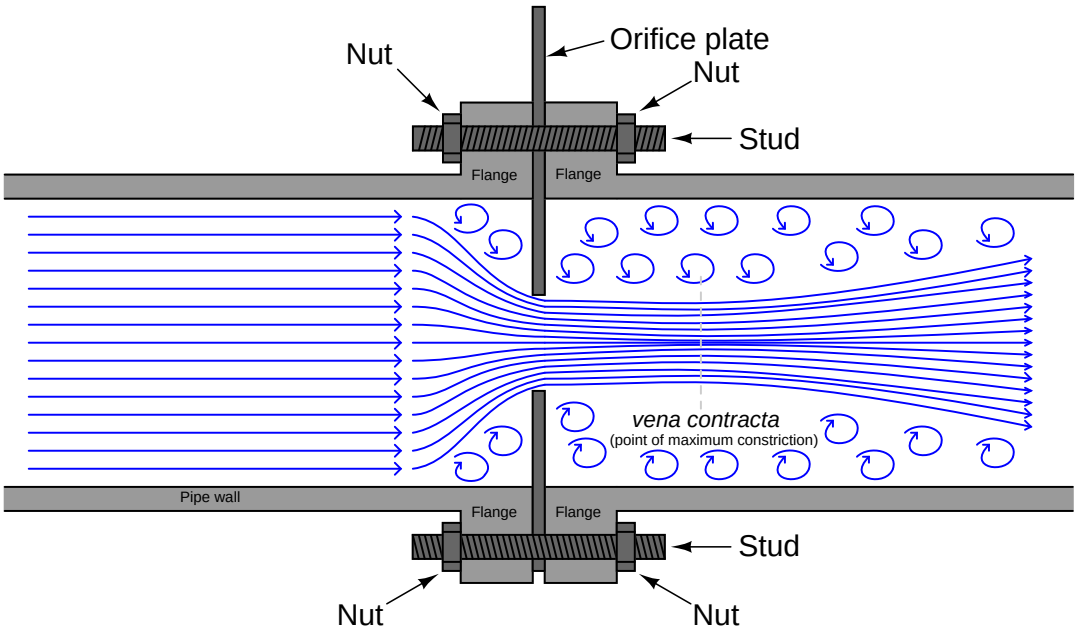
Another valuable lesson we may take from the faces of these indicating instruments is how uncertain the flow measurement becomes at the low end of the scale. Note how for each indicating instrument (both the receiver gauge and the meter movement), the square-root scale is "compressed" at the low end, to the point where it becomes impossible to interpret fine increments of flow at that end of the scale. At the high end of each scale, it's a different situation entirely: the numbers are spaced so far apart that it's easy to read fine distinctions in flow values (e.g. 94% flow versus 95% flow). However, the scale is so crowded at the low end that it's really impossible to clearly distinguish two different flow values such as 4% from 5%.

This "crowding" is not just an artifact of a visual scale; it is a reflection of a fundamental limitation in measurement certainty with this type of flow measurement. The amount of differential pressure separating different low-range values of flow for a flow element is so little, even small amounts of pressure-measurement error equate to large amounts of flow-measurement error. In other words, it becomes more and more difficult to precisely resolve flow rate as the flow rate decreases toward the low end of the scale. The "crowding" that we see on an indicator's square-root scale is a visual reflection of this fundamental problem: even a small error in interpreting the pointer's position at the low end of the scale can yield major errors in flow interpretation.

A technical term used to quantify this problem is *turndown*. "Turndown" refers to the ratio of high-range measurement to low-range measurement possible for an instrument while maintaining reasonable accuracy. For pressure-based flowmeters, which must deal with the non-linearities of Bernoulli's Equation, the practical turndown is often no more than 3 to 1 (3:1). This means a flowmeter ranged for 0 to 300 GPM might only read accurately down to a flow of 100 GPM. Below that, the accuracy becomes so poor that the measurement is almost useless. Advances in DP transmitter technology have pushed this ratio further, perhaps as far as 10:1 for certain installations. However, the fundamental problem is not transmitter resolution, but rather the nonlinearity of the flow element itself. This means *any* source of pressure-measurement error – whether originating in the transmitter's pressure sensor or not – compromises our ability to accurately measure flow at low rates. Even with a *perfectly* calibrated transmitter, errors resulting from wear of the flow element (e.g. a dulled edge on an orifice plate) or from uneven liquid columns in the impulse tubes connecting the transmitter to the element, will cause large flow-measurement errors at the low end of the instrument's range where the flow element hardly produces a differential pressure at all. Everyone involved with the technical details of flow measurement needs to understand this fact: the fundamental problem of limited turndown is grounded in the physics of turbulent flow and potential/kinetic energy exchange for these flow elements. Technological improvements will help, but they cannot overcome the limitations imposed by physics. If better turndown is required for a particular flow-measurement application, a wholly different type of flowmeter should be considered.

21.1.5 Orifice plates

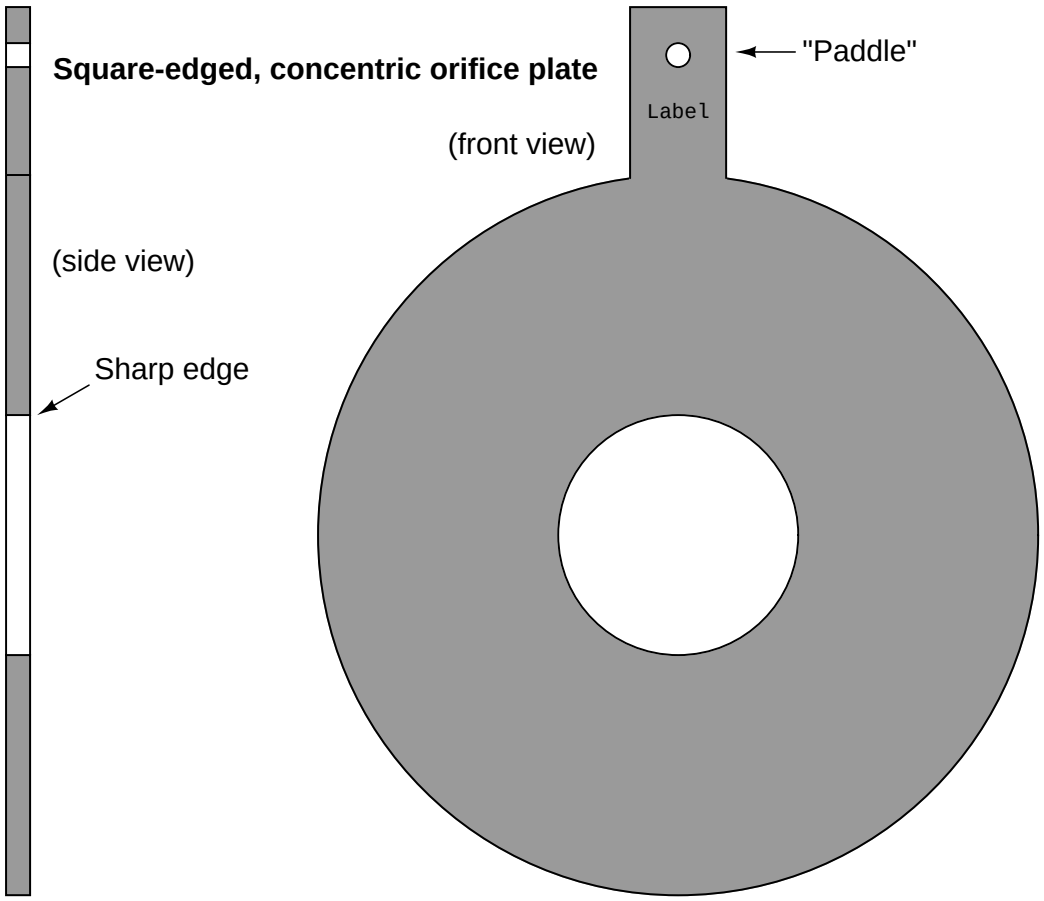
Of all the pressure-based flow elements in existence, the most common is the *orifice plate*. This is simply a metal plate with a hole in the middle for fluid to flow through. Orifice plates are typically sandwiched between two flanges of a pipe joint, allowing for easy installation and removal:



The point where the fluid flow profile constricts to a minimum cross-sectional area after flowing through the orifice is called the *vena contracta*, and it is the area of minimum fluid pressure. The vena contracta corresponds to the narrow throat of a venturi tube. The precise location of the vena contracta for an orifice plate installation will vary with flow rate, and also with the *beta ratio* (β) of the orifice plate, defined as the ratio of bore diameter (d) to inside pipe diameter (D):

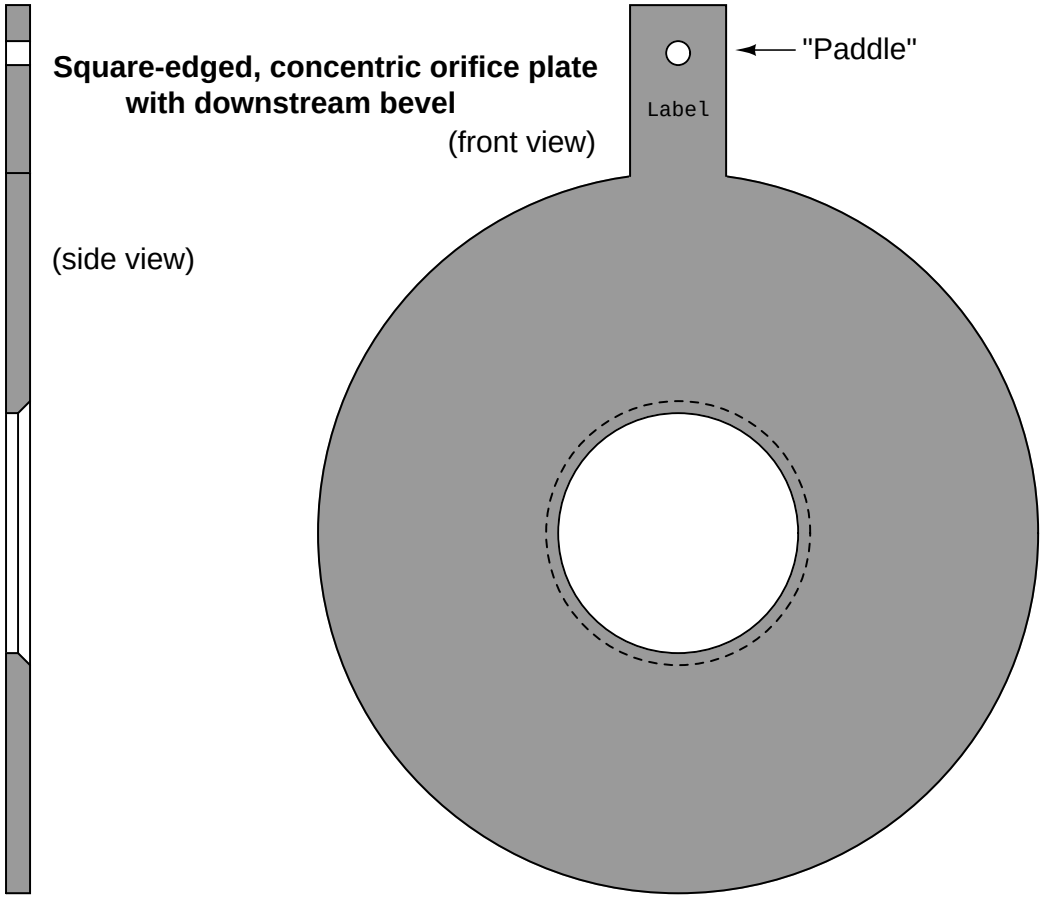
$$\beta = \frac{d}{D}$$

The simplest design of orifice plate is the *square-edged, concentric* orifice. This type of orifice plate is manufactured by machining a precise, straight hole in the middle of a thin metal plate. Looking at a side view of a square-edged concentric orifice plate reveals sharp edges (90^o corners) at the hole:



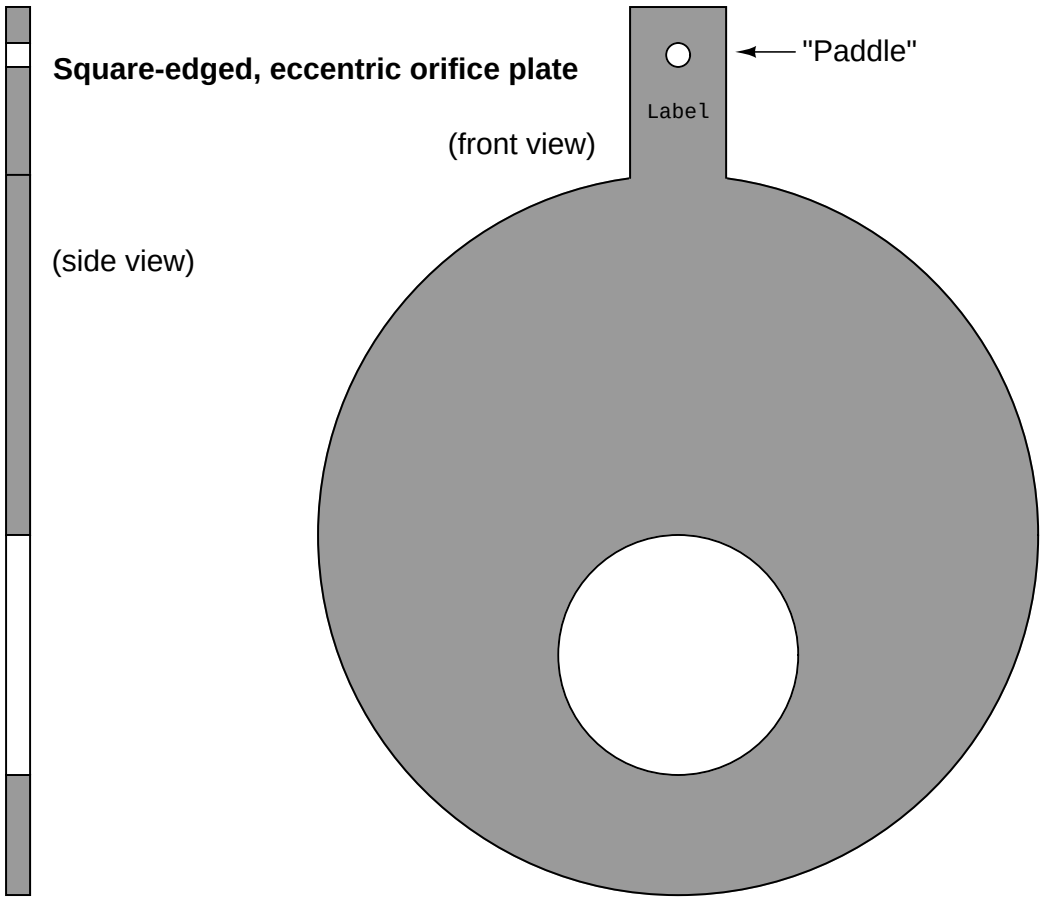
Square-edged orifice plates may be installed in either direction, since the orifice plate "appears" exactly the same from either direction of fluid approach. In fact, this allows square-edged orifice plates to be used for measuring bidirectional flow rates (where the fluid flow direction reverses itself from time to time). A text label printed on the "paddle" of any orifice plate customarily identifies the upstream side of that plate, but in the case of the square-edged orifice plate it does not matter.

The purpose of having a square edge on the hole in an orifice plate is to minimize contact with the fast-moving moving fluid stream going through the hole. Ideally, this edge will be knife-sharp. If the orifice plate is relatively thick (1/8 or an inch or more), it may be necessary to bevel the downstream side of the hole to further minimize contact with the fluid stream:



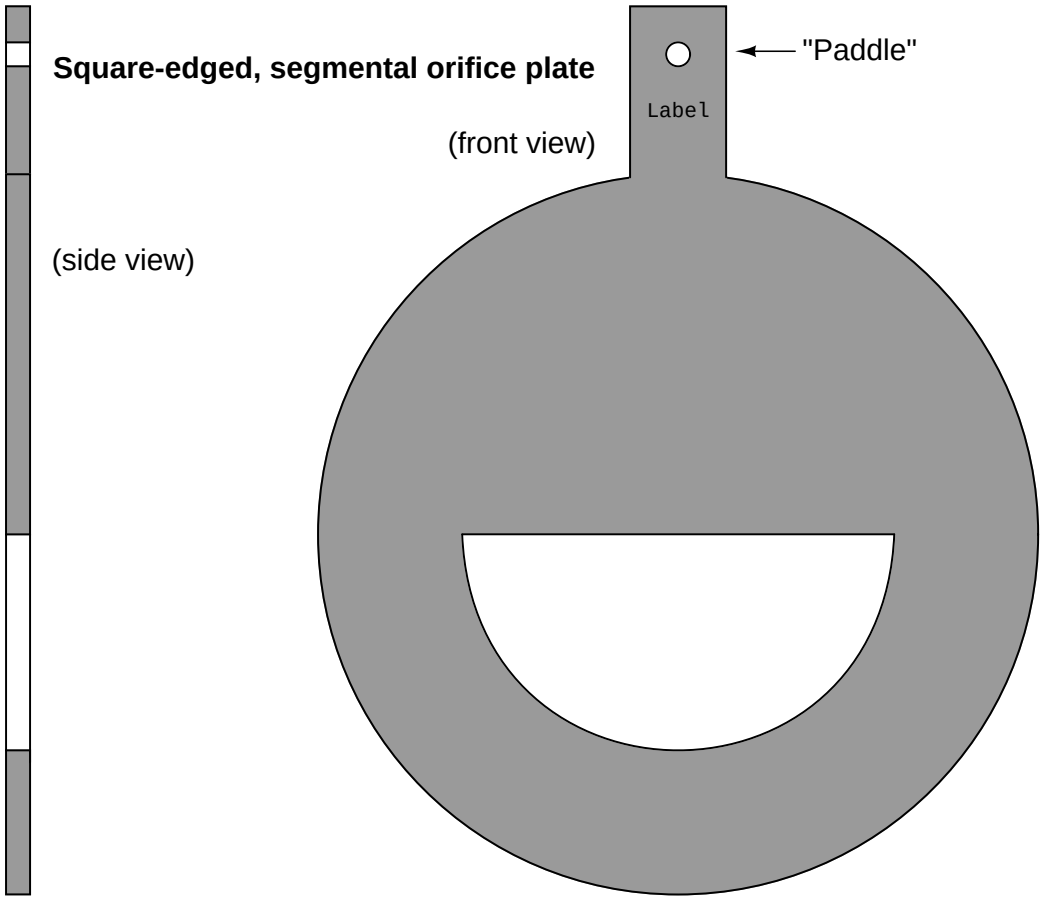
Looking at the side-view of this orifice plate, the intended direction of flow is left-to-right, with the sharp edge facing the incoming fluid stream and the bevel providing a non-contact outlet for the fluid. Beveled orifice plates are obviously uni-directional, and *must* be installed with the paddle text facing upstream.

Other square-edged orifice plates exist to address conditions where gas bubbles or solid particles may be present in liquid flows, or where liquid droplets or solid particles may be present in gas flows. The first of this type is called the *eccentric* orifice plate, where the hole is located off-center to allow the undesired portions of the fluid to pass through the orifice rather than build up on the upstream face:



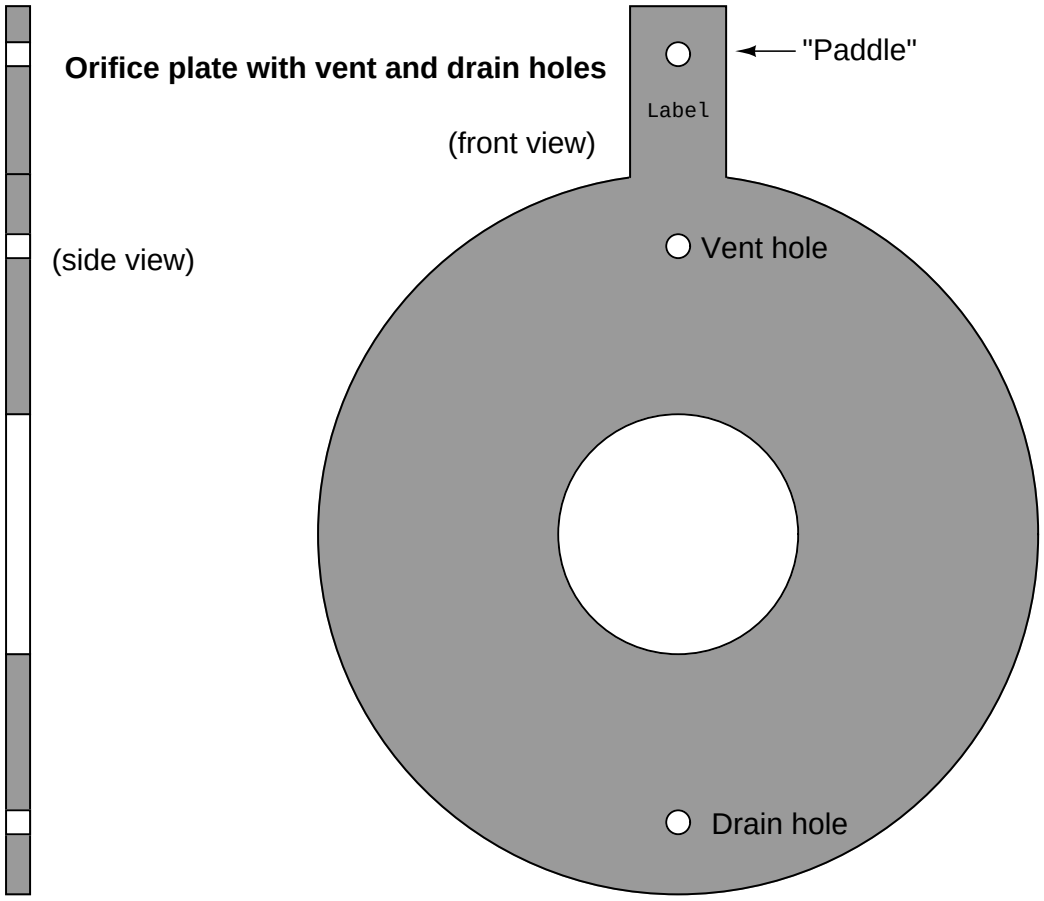
For gas flows, the hole should be offset downward, so any liquid droplets or solid particles may easily pass through. For liquid flows, the hole should be offset upward to allow gas bubbles to pass through and offset downward to allow heavy solids to pass through.

The second off-center orifice plate type is called the *segmental orifice plate*, where the hole is not circular but rather just a segment of a concentric circle:



As with the eccentric orifice plate design, the segmental hole should be offset downward in gas flow applications and either upward or downward in liquid flow applications depending on the type of undesired material(s) in the flowstream.

An alternative to offsetting or re-shaping the bore hole of an orifice plate is to simply drill a small hole near the edge of the plate, flush with the inside diameter of the pipe, allowing undesired substances to pass through the plate rather than collect on the upstream side. If such a hole is oriented upward to pass vapor bubbles, it is called a *vent hole*. If the hole is oriented downward to pass liquid droplets or solids, it is called a *drain hole*. Vent and drain holes are useful when the concentration of these undesirable substances is not significant enough to warrant either an eccentric or segmental orifice:



The addition of a vent or drain hole should have negligible impact on the performance of an orifice plate due to its small size relative to the main bore. If the quantity of undesirable material in the flowstream (bubbles, droplets, or solids) is excessive, an eccentric or segmental orifice plate might be a better choice[9].

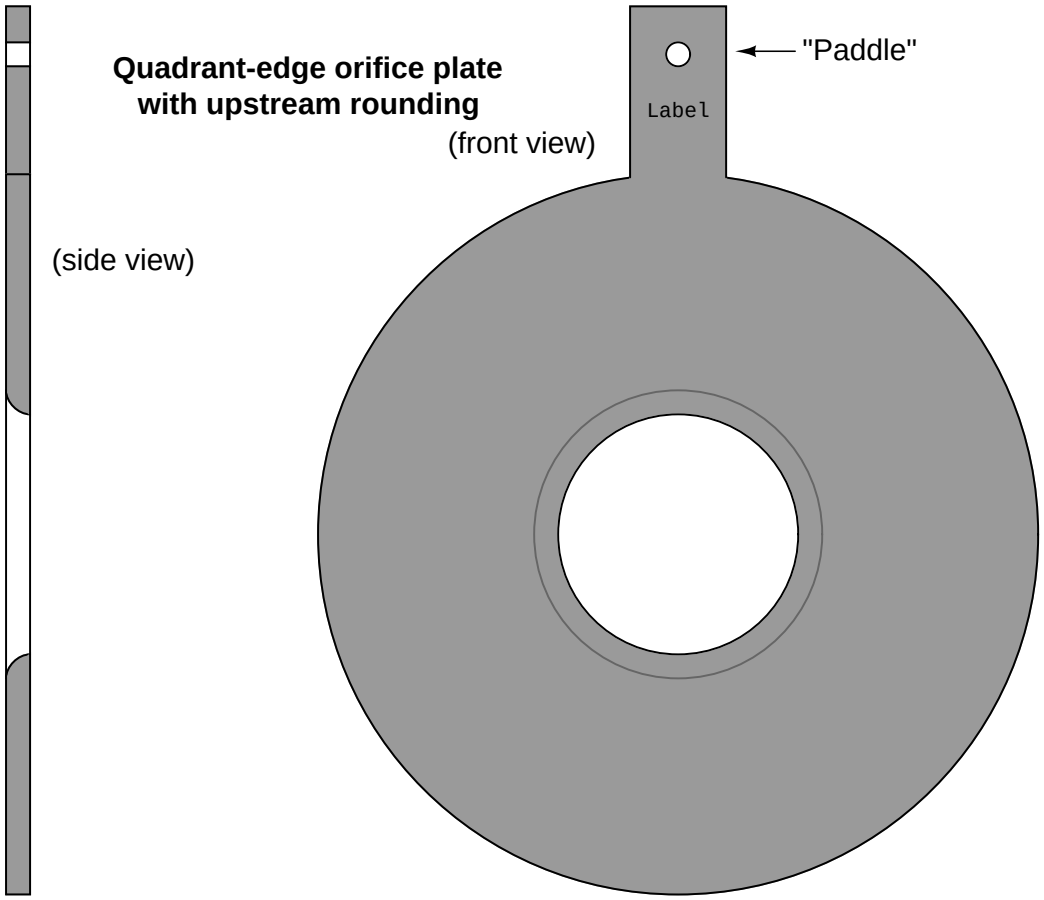
[9] L.K. Spink, in his book *Principles and Practice of Flow Meter Engineering*, notes that drain holes intended to pass solid objects may be useless in small pipe sizes, where the hole is so small it will probably become plugged with solid debris and cease to provide benefit. In such installations he recommends re-orienting the pipe vertically instead

Some orifice plates employ non-square-edged holes for the purpose of improving performance at low Reynolds number[10] values, where the effects of fluid viscosity are more apparent. These orifice plate types employ rounded- or conical-entrance holes in an effort to minimize the effects of fluid viscosity. Experiments have shown that decreased Reynolds number causes the flowstream to not contract as much when traveling through an orifice, thus limiting fluid acceleration and decreasing the amount of differential pressure produced by the orifice plate. However, experiments have also shown that decreased Reynolds number in a venturi-type flow element causes an *increase* in differential pressure due to the effects of friction against the entrance cone walls. By manufacturing an orifice plate in such a way that the hole exhibits "venturi-like" properties (i.e. a dull edge where the fast-moving fluid stream has more contact with the plate), these two effects tend to cancel each other, resulting in an orifice plate that maintains consistent accuracy at lower flow rates and/or higher viscosities than the simple square-edged orifice.

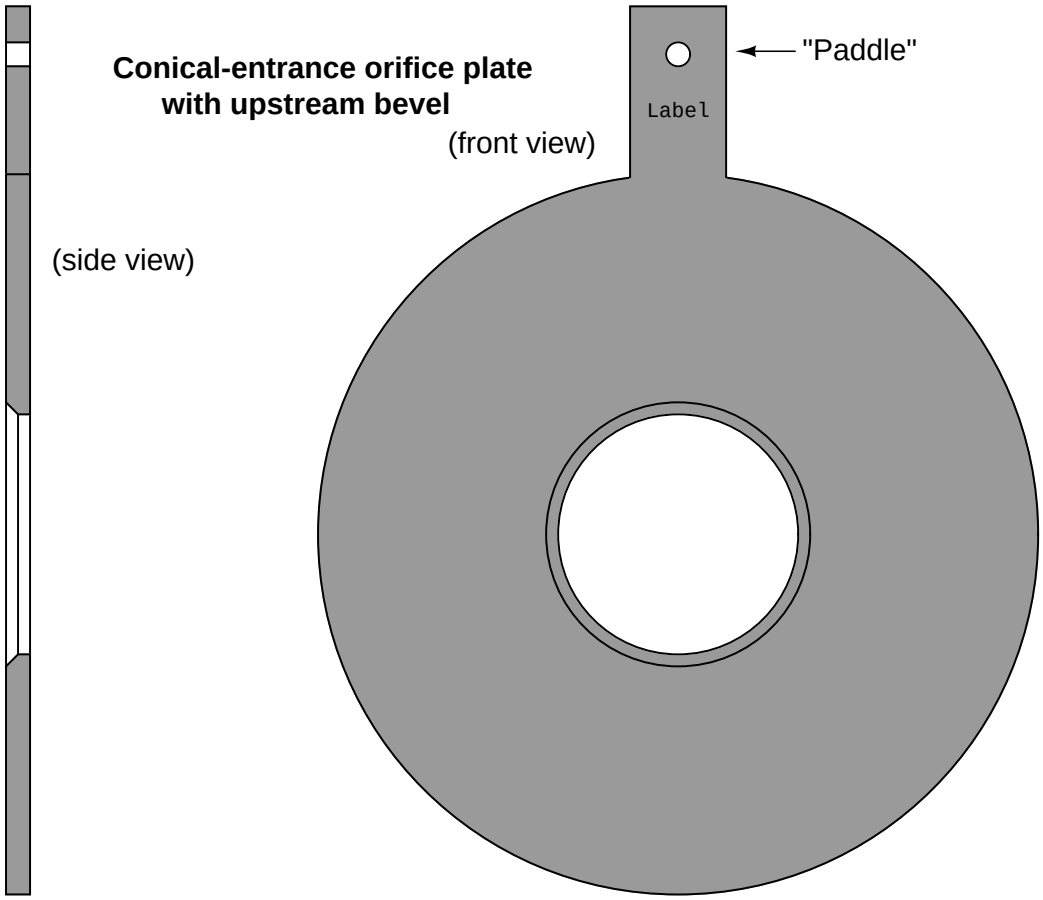
of horizontally. This allows solids to pass through the main bore of the orifice without "damming" on the upstream side of the orifice plate. I would add the suggestion to consider a different primary element entirely, such as a venturi tube. The small size of the line will limit the cost of such an element, and the performance is likely to be far better than an orifice plate anyway.

[10] To read more about the concept of Reynolds number, refer to section 2.9.9 beginning on page 115.

Two common non-square-edge orifice plate designs are the *quadrant-edge* and *conic-entrance* orifices. The quadrant-edge is shown first:

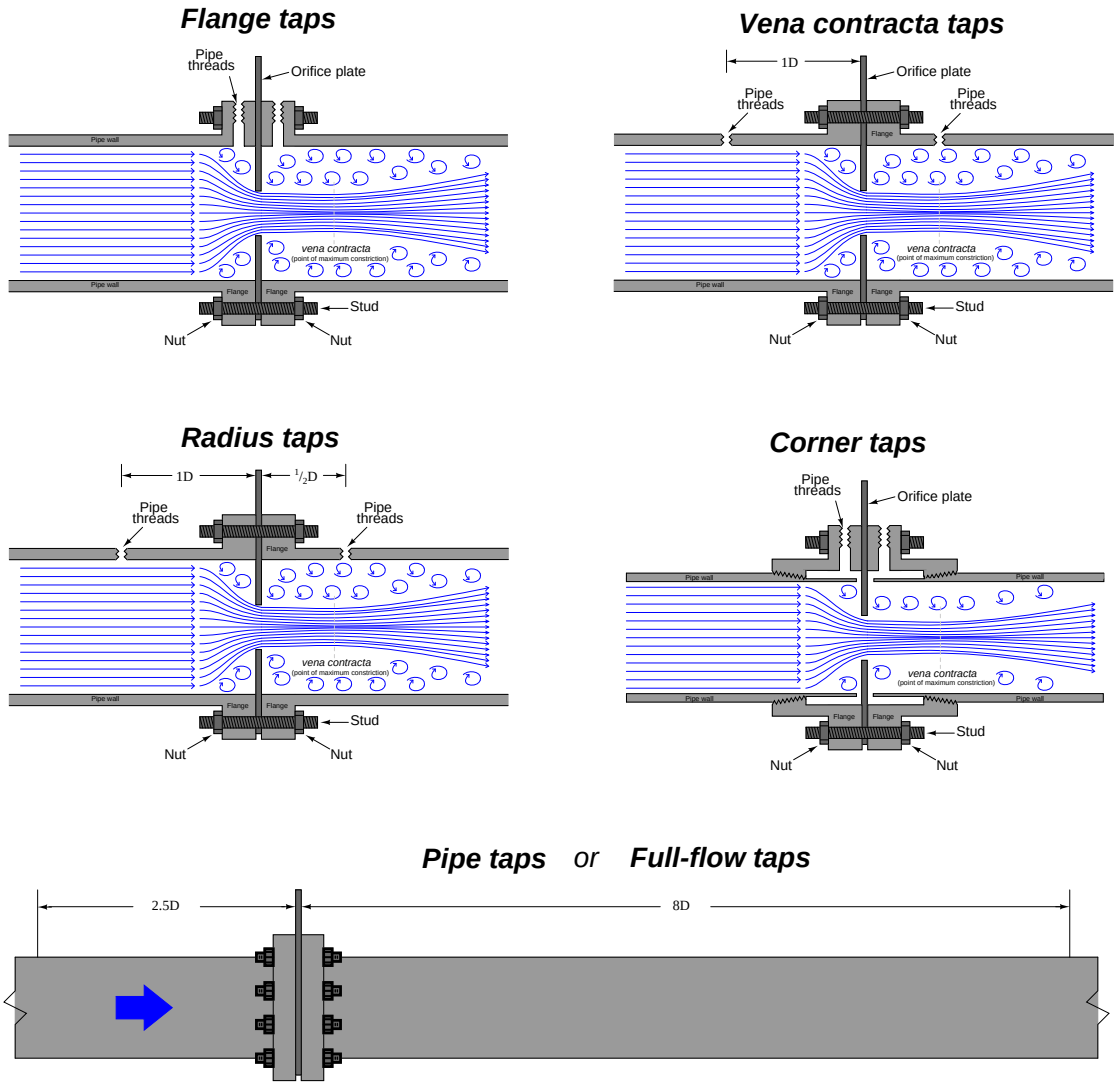


The conical-entrance orifice plate looks like a beveled square-edge orifice plate installed backwards, with flow entering the conical side and exiting the square-edged side:



Here, is it vitally important to pay attention to the paddle's text label. This is the only sure indication of which direction an orifice plate needs to be installed. One can easily imagine an instrument technician mistaking a conical-entrance orifice plate for a square-edged, beveled orifice plate and installing it backward!

Several standards exist for pressure tap locations. Ideally, the upstream pressure tap will detect fluid pressure at a point of minimum velocity, and the downstream tap will detect pressure at the vena contracta (maximum velocity). In reality, this ideal is never perfectly achieved. An overview of the most popular tap locations for orifice plates is shown in the following illustration:



Flange taps are the most popular tap location for orifice meter runs on large pipes in the United States. Flanges may be manufactured with tap holes pre-drilled and finished before the flange is even welded to the pipe, making this a very convenient pressure tap configuration. Most of the other tap configurations require drilling into the pipe after installation, which is not only labor-intensive, but may possibly weaken the pipe at the locations of the tap holes.

Vena contracta taps offer the greatest differential pressure for any given flow rate, but require precise calculations to properly locate the downstream tap position. *Radius taps* are an approximation of vena contracta taps for large pipe sizes (one-half pipe diameter downstream for the low-pressure tap location). An unfortunate characteristic of both these taps is the requirement of drilling through the pipe wall. Not only does this weaken the pipe, but the practical necessity of drilling the tap holes in the installed location rather than in a controlled manufacturing environment means there is considerable room for installation error[11].

Corner taps must be used on small pipe diameters where the vena contracta is so close to the downstream face of the orifice plate that a downstream flange tap would sense pressure in the highly turbulent region (too far downstream). Corner taps obviously require special (i.e. expensive) flange fittings, which is why they tend to be used only when necessary.

Care should be taken to avoid measuring downstream pressure in the highly turbulent region following the vena contracta. This is why the *pipe tap* (also known as *full-flow tap*) standard calls for a downstream tap location eight pipe diameters away from the orifice: to give the flow stream room to stabilize for more consistent pressure readings[12].

Wherever the taps are located, it is vitally important that the tap holes be completely flush with the inside wall of the pipe or flange. Even the smallest recess or burr left from drilling will cause measurement errors, which is why tap holes are best drilled in a controlled manufacturing environment rather that at the installation site where the task will likely be performed by non-experts.

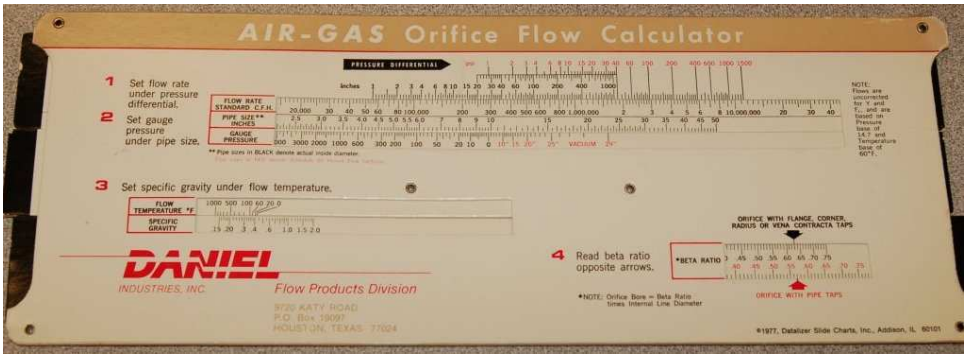
[11] One significant source of error for customer-drilled tap holes is the interior finish of the holes. Even a small "burr" of metal left where the hole penetrates the inner surface of the pipe wall will cause substantial flow measurement errors!

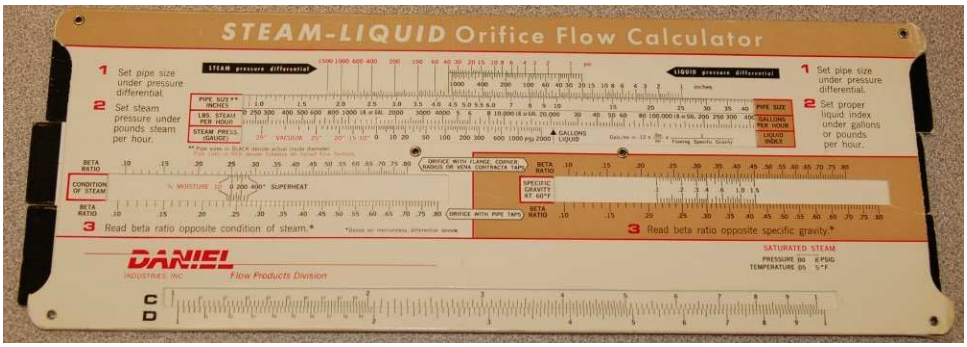
[12] What this means is that a "pipe tap" installation is actually measuring permanent pressure loss, which also happens to scale with the square of flow rate because the primary mechanism for energy loss in turbulent flow conditions is the translation of linear velocity to angular (swirling) velocity in the form of eddies. This kinetic energy is eventually dissipated in the form of heat as the eddies eventually succumb to viscosity.

For relatively low flow rates, an alternative arrangement is the *integral orifice plate*. This is where a small orifice plate directly attaches to the differential pressure-sensing element, eliminating the need for impulse lines. A photograph of an integral orifice plate and transmitter is shown here:

The task of properly sizing an orifice plate for any given application is complex enough to recommend the use of special orifice sizing computer software provided by orifice plate manufacturers. There are a number of factors to consider in orifice plate sizing, and these software packages account for all of them. Best of all, the software provided by manufacturers is often linked to data for that manufacturer's product line, helping to assure installed results in close agreement with predictions.

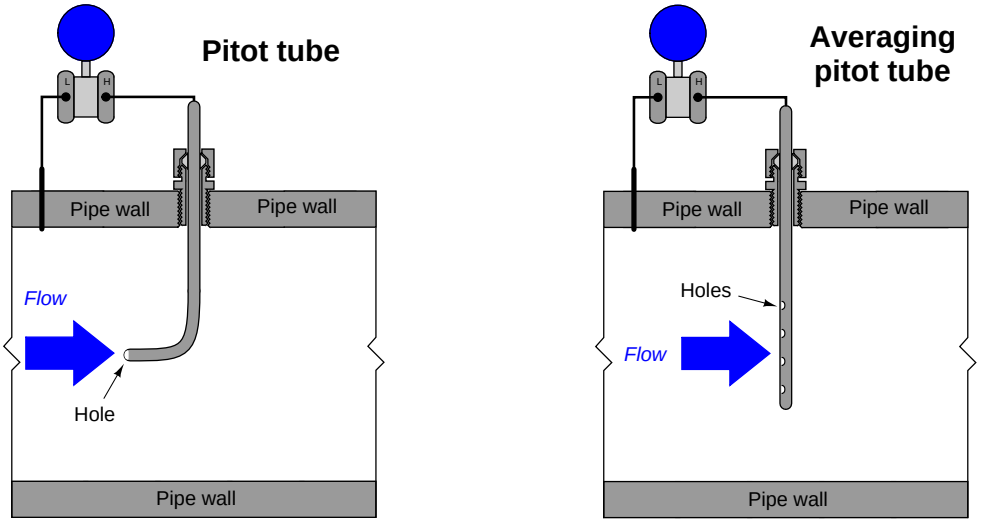
In the days before ubiquitous personal computers and the internet, some orifice plate manufacturers provided customers with paper "slide rule" calculators to help them select appropriate orifice plate sizes from known process parameters. The following photographs show the front and back sides of one such slide rule:



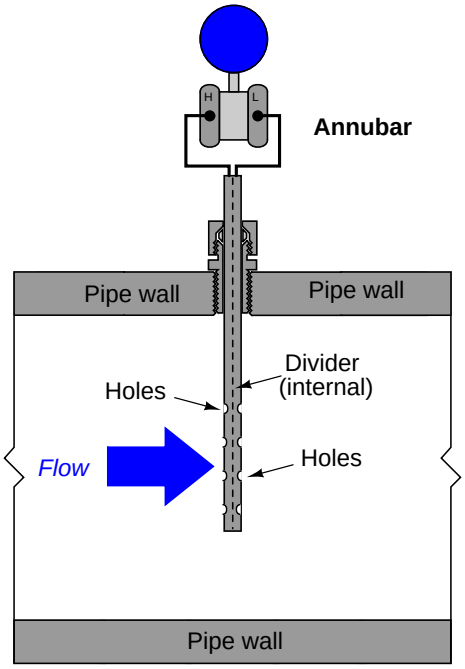


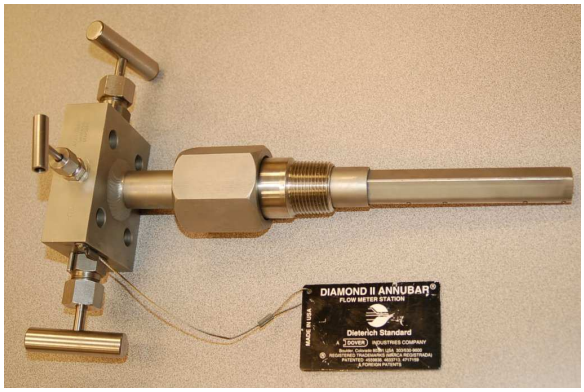
21.1.6 Other differential producers

Other pressure-based flow elements exist as alternatives to the orifice plate. The *Pitot tube*, for example, senses pressure as the fluid stagnates (comes to a complete stop) against the open end of a forward-facing tube. A shortcoming of the classic single-tube Pitot assembly is sensitivity to fluid velocity at just one point in the pipe, so a more common form of Pitot tube seen in industry is the *averaging* Pitot tube consisting of several stagnation holes sensing velocity at multiple points across the width of the flow:



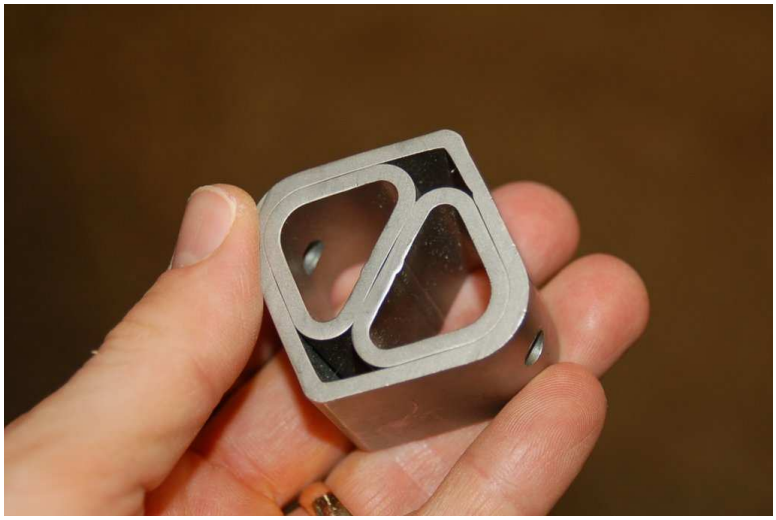
A variation on the latter theme is the *Annubar* flow element, a trade name of the Dieterich Standard corporation. An "Annubar" is an averaging pitot tube consolidating high and low pressure-sensing ports in a single probe assembly:



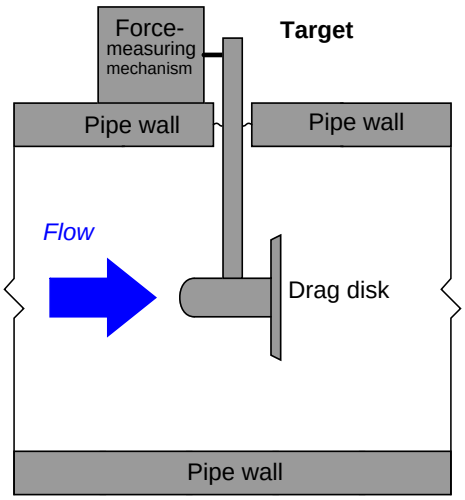


What appears at first glance to be a single, square-shaped tube inserted into the pipe is actually a double-ported tube with holes on both the upstream and downstream edges:

A section of Annubar tube clearly shows the porting and dual chambers, designed to bring upstream (stagnation) and downstream pressures out of the pipe to a differential pressure-sensing instrument:

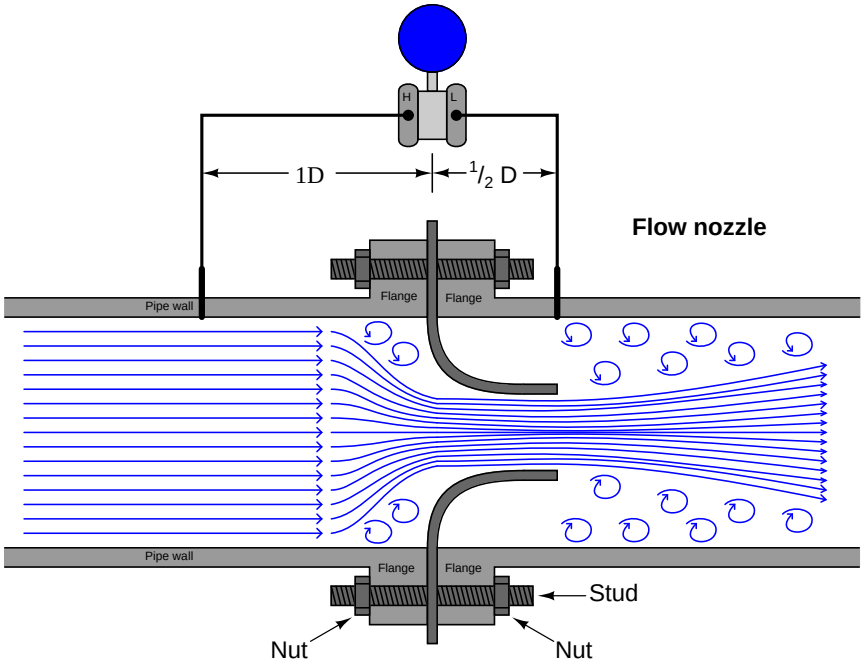


A less sophisticated realization of the stagnation principle is the *target* flow sensor, consisting of a blunt "paddle" (or "drag disk") inserted into the flowstream. The force exerted on this paddle by the moving fluid is sensed by a special transmitter mechanism, which then outputs a signal corresponding to flow rate (proportional to the square of fluid velocity, just like an orifice plate):

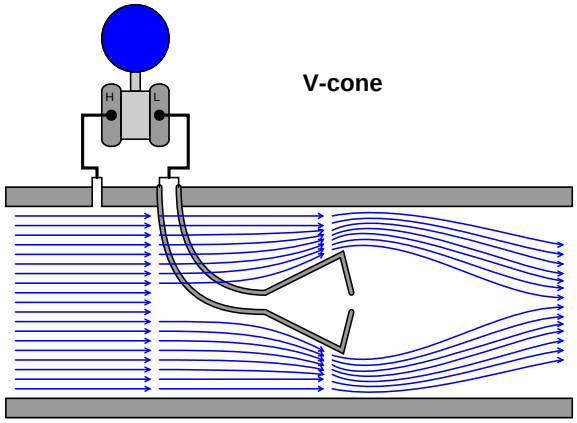


The classic venturi tube pioneered by Clemens Herschel in 1887 has been adapted in a variety of forms broadly classified as *flow tubes*. All flow tubes work on the same principle: developing a differential pressure by channeling fluid flow from a wide tube to a narrow tube. The differ from the classic venturi only in construction details, the most significant detail being a significantly shorter length than the classic venturi tube. Examples of flow tube designs include the *Dall* tube, *Lo-Loss* flow tube, *Gentile* or *Bethlehem* flow tube, and the *B.I.F. Universal Venturi*.

Another variation on the venturi theme is called a *flow nozzle*, designed to be clamped between the faces of two pipe flanges in a manner similar to an orifice plate. The goal here is to achieve simplicity of installation approximating that of an orifice plate while improving performance (less permanent pressure loss) over orifice plates:

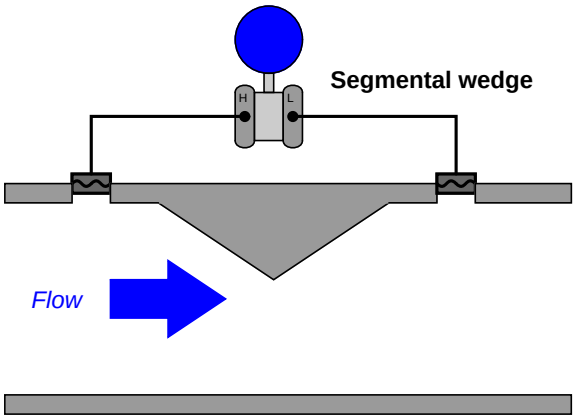


Two more variations on the venturi theme are the *V-cone* and *Segmental wedge* flow elements. The V-cone (or "venturi cone," a trade name of the McCrometer division of the Danaher corporation) may be thought of as a venturi tube in reverse: instead of narrowing the tube's diameter to cause fluid acceleration, fluid must flow around a cone-shaped obstruction placed in the middle of the tube. The tube's effective area will be reduced by the presence of this cone, causing fluid to accelerate through the restriction just as it would through the throat of a classic venturi tube:

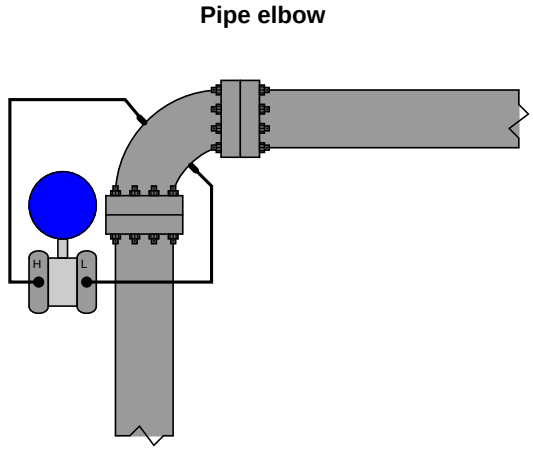


This cone is hollow, with a pressure-sensing port on the downstream side allowing for easy detection of fluid pressure near the vena contracta. Upstream pressure is sensed by another port in the pipe wall upstream of the cone. The following photograph shows a V-cone flow tube, cut away for demonstration purposes:

Segmental wedge elements are special pipe sections with wedge-shaped restrictions built in. These devices, albeit crude, are useful for measuring the flow rates of *slurries*, especially when pressure is sensed by the transmitter through remote-seal diaphragms (to eliminate the possibility of impulse tube plugging):



Finally, the lowly pipe elbow may be pressed into service as a flow-measuring element, since fluid turning a corner in the elbow experiences radial acceleration and therefore generates a differential pressure along the axis of acceleration:



Pipe elbows should be considered for flow measurement only as a last resort. Their inaccuracies tend to be extreme, owing to the non-precise construction of most pipe elbows and the relatively weak differential pressures generated[13].

[13]The fact that a pipe elbow generates small differential pressure is an accuracy concern because other sources of pressure become larger by comparison. Noise generated by fluid turbulence in the elbow, for example, becomes a significant portion of the pressure sensed by the transmitter when the differential pressure is so low (i.e. the signal-to-noise ratio becomes smaller). Errors caused by differences in elbow tap elevation and different impulse line fill fluids,

A final point should be mentioned on the subject of differential-producing elements, and that is their energy dissipation. Orifice plates are simple and relatively inexpensive to install, but their permanent pressure loss is high compared with other primary elements such as venturi tubes. Permanent pressure loss is permanent energy loss from the flowstream, which usually represents a loss in energy invested into the process by pumps, compressors, and/or blowers. Fluid energy dissipated by an orifice plate thus (usually) translates into a requirement of greater energy input to that process[14].

With the financial and ecological costs of energy being non-trivial in our modern world, it is important to consider energy loss as a significant factor in choosing the appropriate primary element for a pressure-based flowmeter. It might very well be that an "expensive" venturi tube saves more money in the long term than a "cheap" orifice plate, while delivering greater measurement accuracy as an added benefit.

for example, become more significant as well.

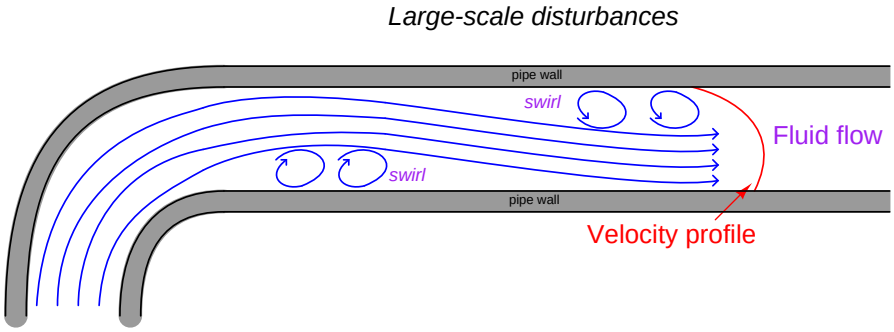
[14]This is not always the case, as primary elements are often found on throttled process lines. In such cases where a control valve normally throttles the flow rate, any energy dissipated by the orifice plate is simply less energy that the valve would otherwise be required to dissipate. Therefore, the presence or absence of an orifice plate has no net impact on energy dissipation when used on a process flow throttled by a control valve.

21.1.7 Proper installation

Perhaps the most common way in which the flow measurement accuracy of any flowmeter becomes compromised is incorrect installation, and pressure-based flowmeters are no exception to this rule. The following list shows some of the details one must consider in installing a pressure-based flowmeter element:

- Necessary upstream and downstream straight-pipe lengths
- Beta ratio (ratio of orifice bore diameter to pipe diameter: $\beta = \frac{d}{D}$)
- Impulse tube tap locations
- Tap finish
- Transmitter location in relation to the pipe

Sharp turns in piping networks introduce large-scale turbulence[15] into the flowstream. Elbows, tees, valves, fans, and pumps are some of the most common causes of large-scale turbulence in piping systems. Successive pipe elbows in different planes are some of the worst offenders in this regard. When the natural flow path of a fluid is disturbed by such piping arrangements, the velocity profile of that fluid will become asymmetrical; e.g. the velocity gradient from one wall boundary of the pipe to the other will not be orderly. Large eddies in the flowstream (called *swirl*) will be present. This may cause problems for pressure-based flow elements which rely on linear acceleration (change in velocity in one dimension) to measure fluid flow rate. If the flow profile is distorted enough, the acceleration detected at the element may be too great or too little, and therefore not properly represent the full fluid flowstream[16].

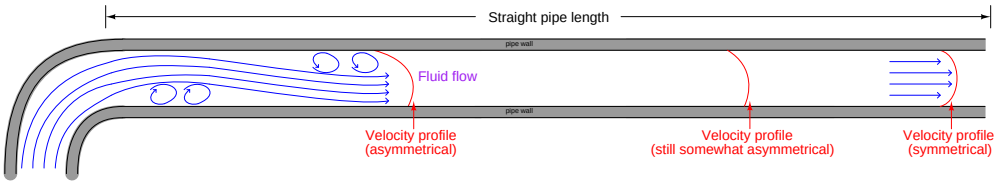


Even disturbances located *downstream* of the flow element impact measurement accuracy (albeit not as much as upstream disturbances). Unfortunately, both upstream and downstream flow

[15] This is not to be confused with micro-turbulence in the fluid, which cannot be eliminated at high Reynolds number values. In fact, "fully-developed turbulent flow" is desirable for head-based meter elements such as orifice plates because it means the flow profile will be relatively flat (even velocities across the pipe's diameter) and frictional forces (viscosity) will be negligible. The thing we are trying to avoid is *large-scale* turbulent effects such as eddies, swirl, and asymmetrical flow profiles, which compromise the ability of most flowmeters to accurate measure flow rate.

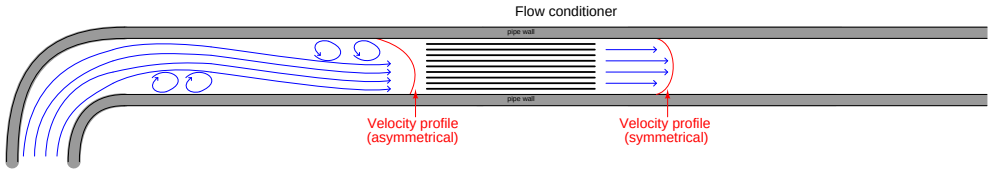
[16] L.K. Spink mentions in his book *Principles and Practice of Flow Meter Engineering* that certain tests have shown flow measurement errors induced from severe disturbances as far as 60 to 100 pipe diameters upstream of the primary flow element!

disturbances are unavoidable on all but the simplest fluid systems. This means we must devise ways to stabilize a flowstream's velocity profile near the flow element in order to achieve accurate measurements of flow rate. A very simple and effective way to stabilize a flow profile is to provide adequate lengths of straight pipe ahead of (and behind) the flow element. Given enough time, even the most chaotic flowstream will "settle down" to a symmetrical profile all on its own. The following illustration shows the effect of a pipe elbow on a flowstream, and how the velocity profile returns to a normal (symmetrical) form after traveling through a sufficient length of straight pipe:



Recommendations for minimum upstream and downstream straight-pipe lengths vary significantly with the nature of the turbulent disturbance, piping geometry, and flow element. As a general rule, elements having a smaller beta ratio (ratio of throat diameter d to pipe diameter D) are more tolerant of disturbances, with profiled flow devices (e.g. venturi tubes, flow tubes, V-cones) having the greatest tolerance[17]. Ultimately, you should consult the flow element manufacturer's documentation for a more detailed recommendation appropriate to any specific application.

In applications where sufficient straight-run pipe lengths are impractical, another option exists for "taming" turbulence generated by piping disturbances. Devices called *flow conditioners* may be installed upstream of the flow element to help the flow profile achieve symmetry in a far shorter distance than simple straight pipe could do alone. Flow conditioners take the form of a series of tubes or vanes installed inside the pipe, parallel to the direction of flow. These tubes or vanes force the fluid molecules to travel in straighter paths, thus stabilizing the flowstream prior to entering a flow element:

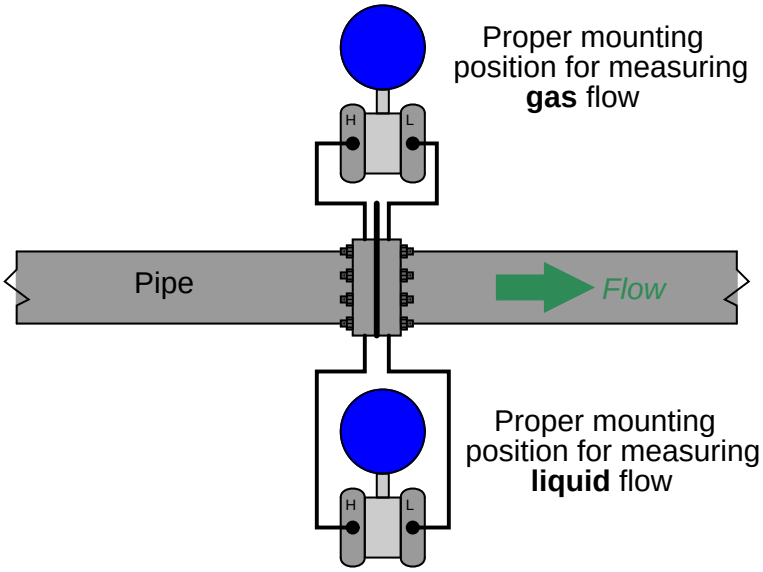


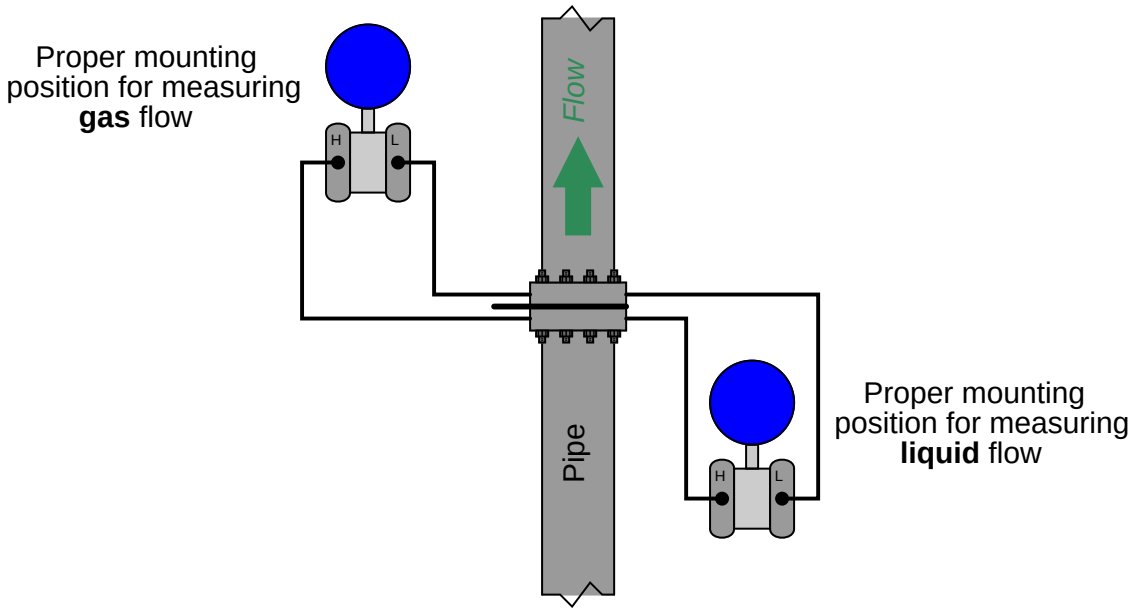
Another common source of trouble for pressure-based flowmeters is improper transmitter location. Here, the type of process fluid flow being measured dictates how the pressure-sensing

[17] Flow elements with low beta ratio values tolerate greater disturbance in the flow pattern because they accelerate the flowstream to a greater degree. This may be best visualized by a thought experiment where we imagine an orifice plate with a very large beta ratio (i.e. one where the bore size is nearly as large as the pipe diameter): such an orifice plate would hardly accelerate the fluid at all, which would mean a mis-shapen flow profile entering the bore would probably remain mis-shapen exiting it. The acceleration imparted to a flowstream by a low-beta element tends to overshadow any asymmetries in the flow profile. However, there are disadvantages to using low-beta elements, one of them being increased permanent pressure loss which may translate to increased operating costs due to energy loss.

instrument should be located in relation to the pipe. For gas and vapor flows, it is important that no stray liquid droplets collect in the impulse lines leading to the transmitter, lest a vertical liquid column begin to collect in those lines and generate an error-producing pressure. For liquid flows, it is important that no gas bubbles collect in the impulse lines, or else those bubbles may displace liquid from the lines and thereby cause unequal vertical liquid columns, which would (again) generate an error-producing differential pressure.

In order to let gravity do the work of preventing these problems, we must locate the transmitter *above* the pipe for gas flow applications and *below* the pipe for liquid flow applications:





Condensible vapor applications (such as steam flow measurement) should be treated the same as liquid measurement applications. Here, condensed liquid will collect in the transmitter's impulse lines so long as the impulse lines are cooler than the vapor flowing through the pipe (which is typically the case). Placing the transmitter below the pipe allows vapors to condense and fill the impulse lines with liquid (condensate), which then acts as a natural seal protecting the transmitter from exposure to hot process vapors.

In such applications it is important for the technician to pre-fill both impulse lines with condensed liquid prior to placing the flowmeter into service. "Tee" fittings with removable plugs or fill valves are provided to do this. Failure to pre-fill the impulse lines will likely result in measurement errors during initial operation, as condensed vapors will inevitably fill the impulse lines at slightly different rates and cause a difference in vertical liquid column heights within those lines.

If tap holes must be drilled into the pipe (or flanges) at the process site, great care must be taken to properly drill and de-burr the holes. A pressure-sensing tap hole should be flush with the inner pipe wall, with no rough edges or burrs to create turbulence. Also, there should be no reliefs or countersinking near the hole on the inside of the pipe. Even small irregularities at the tap holes may generate surprisingly large flow-measurement errors.

21.1.8 High-accuracy flow measurement

Many assumptions were made in formulating flow equations from physical conservation laws. Suffice it to say, the flow formulae you have seen so far in this chapter are only approximations of reality. Orifice plates are some of the worst offenders in this regard, since the fluid encounters such abrupt changes in geometry passing through the orifice. Venturi tubes are nearly ideal, since the machined contours of the tube ensure gradual changes in fluid pressure and minimize turbulence.

However, in the real world we must often do the best we can with imperfect technologies. Orifice plates, despite being less than perfect as flow-sensing elements, are convenient and economical to install in flanged pipes. Orifice plates are also the easiest type of flow element to replace in the event of damage or routine servicing. In applications such as custody transfer, where the flow of fluid represents product being bought and sold, flow measurement accuracy is paramount. It is therefore important to figure out how to coax the most accuracy from the common orifice plate in order that we may measure fluid flows both accurately and economically.

If we compare the true flow rate through a pressure-generating primary sensing element against the theoretical flow rate predicted by an idealized equation, we may notice a substantial discrepancy[18]. Causes of this discrepancy include, but are not limited to:

- Energy losses due to turbulence and viscosity
- Energy losses due to friction against the pipe and element surfaces
- Unstable location of *vena contracta* with changes in flow
- Uneven velocity profiles caused by irregularities in the pipe
- Fluid compressibility
- Thermal expansion (or contraction) of the element and piping
- Non-ideal pressure tap location(s)
- Excessive turbulence caused by rough internal pipe surfaces

The ratio between true flow rate and theoretical flow rate for any measured amount of differential pressure is known as the *discharge coefficient* of the flow-sensing element, symbolized by the variable C. Since a value of 1 represents a theoretical ideal, the actual value of C for any real pressure-generating flow element will be less than 1:

$$C = \frac{\text{True flow}}{\text{Theoretical flow}}$$

[18] Richard W. Miller, in his outstanding book *Flow Measurement Engineering Handbook*, states that venturi tubes may come within 1 to 3 percent of ideal, while a square-edged orifice plate may perform as poorly as only 60 percent of theoretical!

For gas and vapor flows, true flow rate deviates even more from the theoretical (ideal) flow value than liquids do, for reasons that have to do with the compressible nature of gases and vapors. A *gas expansion factor* (Y) may be calculated for any flow element by comparing its discharge coefficient for gases against its discharge coefficient for liquids. As with the discharge coefficient, values of Y for any real pressure-generating element will be less than 1:

$$Y = \frac{C_{gas}}{C_{liquid}}$$

$$Y = \frac{\left(\frac{\text{True gas flow}}{\text{Theoretical gas flow}}\right)}{\left(\frac{\text{True liquid flow}}{\text{Theoretical liquid flow}}\right)}$$

Incorporating these factors into the ideal volumetric flow equation developed on page 1002, we arrive at the following formulation:

$$Q = \sqrt{2}\frac{CYA_2}{\sqrt{1 - \left(\frac{A_2}{A_1}\right)^2}}\sqrt{\frac{P_1 - P_2}{\rho}}$$

If we wished, we could even add another factor to account for any necessary unit conversions (N), getting rid of the constant $\sqrt{2}$ in the process:

$$Q = N\frac{CYA_2}{\sqrt{1 - \left(\frac{A_2}{A_1}\right)^2}}\sqrt{\frac{P_1 - P_2}{\rho}}$$

Sadly, neither the discharge coefficient (C) nor the gas expansion factor (Y) will remain constant across the entire measurement range of any given flow element. These variables are subject to some change with flow rate, which further complicates the task of accurately inferring flow rate from differential pressure measurement. However, if we know the values of C and Y for typical flow conditions, we may achieve good accuracy most of the time.

Likewise, the fact that C and Y change with flow places limits on the accuracy obtainable with the "proportionality constant" formulae seen earlier. Whether we are measuring volumetric or mass flow rate, the k factor calculated at one particular flow condition will not hold constant for *all* flow conditions:

$$Q = k\sqrt{\frac{P_1 - P_2}{\rho}}$$

$$W = k\sqrt{\rho(P_1 - P_2)}$$

This means after we have calculated a value for k based on a particular flow condition, we can only trust the results of the equation for flow conditions not too different from the one we used to calculate k.

As you can see in both flow equations, the density of the fluid (ρ) is an important factor. If fluid density is relatively stable, we may treat ρ as a constant, incorporating its value into the proportionality factor (k) to make the two formulae even simpler:

$$Q = k_Q \sqrt{P_1 - P_2}$$

$$W = k_W \sqrt{P_1 - P_2}$$

However, if fluid density is subject to change over time, we will need some means to continually calculate ρ so our inferred flow measurement will remain accurate. Variable fluid density is a typical state of affairs in gas flow measurement, since all gases are compressible by definition. A simple change in static gas pressure within the pipe is all that is needed to make ρ change, which in turn affects the relationship between flow rate and differential pressure drop.

The American Gas Association (AGA) provides a formula for calculating volumetric flow of any gas using orifice plates in their #3 Report, compensating for changes in gas pressure and temperature. A variation of that formula is shown here (consistent with previous forms in this section):

$$Q = N \frac{CYA_2}{\sqrt{1 - \left(\frac{A_2}{A_1}\right)^2}} \sqrt{\frac{Z_s P_1 (P_1 - P_2)}{G_f Z_{f1} T}}$$

Where,

Q = Volumetric flow rate (SCFM = standard cubic feet per minute)

N = Unit conversion factor

C = Discharge coefficient (accounts for energy losses, Reynolds number corrections, pressure tap locations, etc.)

A_1 = Cross-sectional area of mouth

A_2 = Cross-sectional area of throat

Z_s = Compressibility factor of gas under standard conditions

Z_{f1} = Compressibility factor of gas under flowing conditions, upstream

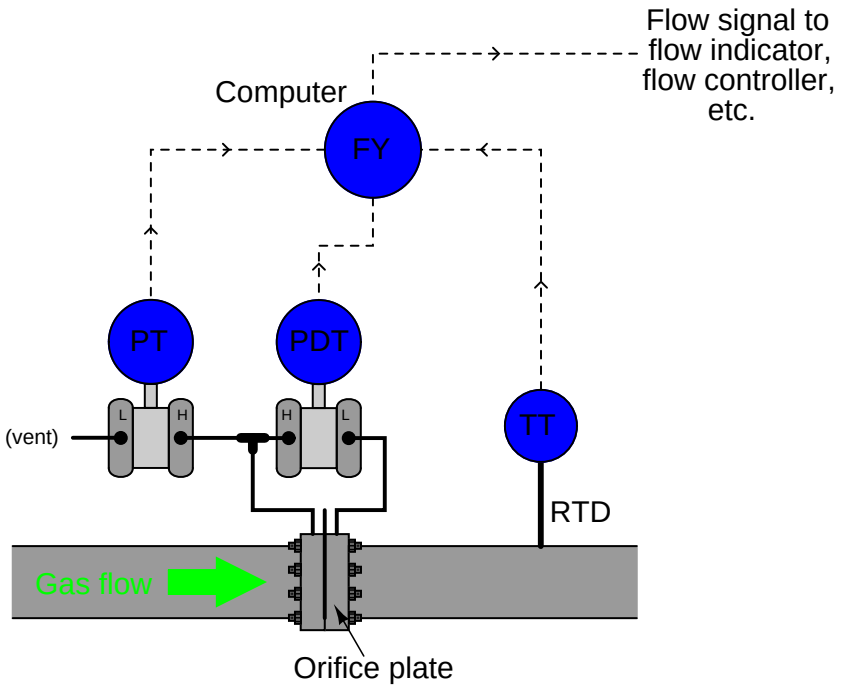
G_f = Specific gravity of gas (density compared to ambient air)

T = Absolute temperature of gas

P_1 = Upstream pressure (absolute)

P_2 = Downstream pressure (absolute)

This equation implies the continuous measurement of gas pressure (P_1) and temperature (T) inside the pipe, in addition to the differential pressure produced by the orifice plate ($P_1 - P_2$). These measurements may be taken by three separate devices, their signals routed to a gas flow computer:



Note the location of the RTD (thermowell), positioned downstream of the orifice plate so the turbulence it generates will have negligible impact on the fluid dynamics at the orifice plate. The American Gas Association (AGA) allows for upstream placement of the thermowell, but only if located at least three feet upstream of a flow conditioner[19].

In order to best control all the physical parameters necessary for good orifice metering accuracy, it is standard practice for custody transfer flowmeter installations to use *honed meter runs* rather than standard pipe and pipe fittings. A "honed run" is a complete piping assembly consisting of a manufactured fitting to hold the orifice plate and sufficient straight lengths of pipe upstream and downstream, the interior surfaces of that pipe machined ("honed") to have a glass-smooth surface with precise and symmetrical dimensions. Such piping "runs" are quite expensive, but necessary if flow measurement accuracy worthy of custody transfer is to be achieved.

[19]Specified in Part 2 of the AGA Report #3, section 2.6.5, page 22.

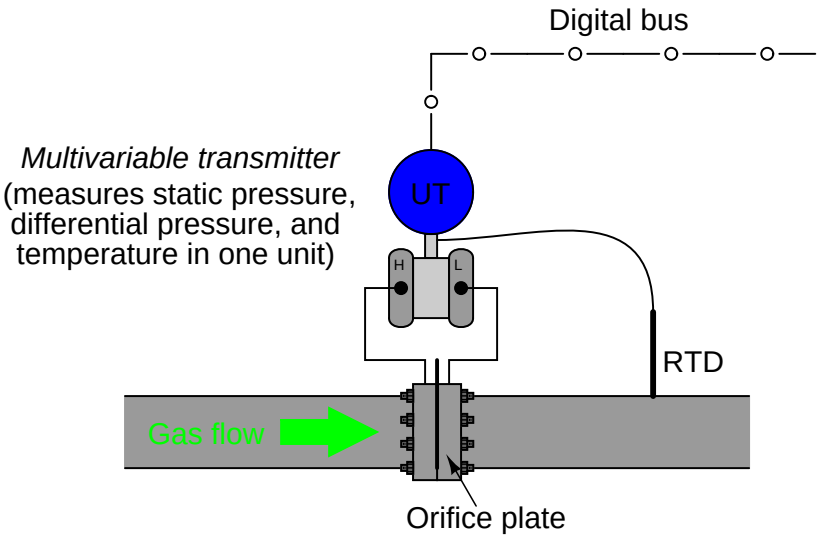
This photograph shows a set of AGA3-compliant orifice meter runs measuring the flow of natural gas:

Note the special transmitter manifolds, built to accept both the differential pressure and absolute pressure (Rosemount model 3051) transmitters. Also note the quick-change fittings (the ribbed cast-iron housings) holding the orifice plates, to facilitate convenient change-out of the orifice plates which is periodically necessary due to wear. It is not unheard of to replace orifice plates on a daily basis in some industries to ensure the sharp orifice edges necessary for accurate measurement[20].

Although not visible in this photograph, these meter runs are connected together by a network of shut-off valves directing the flow of natural gas through as few meter runs as desired. When the total gas flow rate is great, all meter runs are placed into service and their respective flow rates summed to yield a total flow measurement. When the total flow rate decreases, individual meter runs are shut off, resulting in increased flow rates through the remaining meter runs. This "staging" of meter runs expands the effective *rangeability* of the orifice plate as a flow-sensing element, resulting in much more accurate flow measurement over a wide range of flow rates than if a single (large) orifice meter run were used.

[20] This is especially true in the gas exploration industry, where natural gas coming out of the Earth is laden with a substantial amount of sand, rocks, and grit.

An alternative to multiple instruments (differential pressure, absolute pressure, and temperature) installed on each meter run is to use a single *multi-variable* transmitter capable of measuring gas temperature as well as both static and differential pressures. This approach enjoys the advantage of simpler installation over the multi-instrument approach:



The Rosemount model 3095MV and Yokogawa model EJX910 are examples of multi-variable transmitters designed to perform compensated gas flow measurement, equipped with multiple pressure sensors, a connection port for an RTD temperature sensor, and sufficient digital computing power to continuously calculate flow rate based on the AGA equation. Such multi-variable transmitters may provide an analog output for computed flow rate, or a digital output where all three primary variables *and* the computed flow rate may be transmitted to a host system (as shown in the previous illustration). The Yokogawa EJX910A provides an interesting signal output option: a digital *pulse* signal, where each pulse represents a specific quantity (either volume or mass) of fluid. The frequency of this pulse train represents flow rate, while the total number of pulses counted over a period of time represents the total amount of fluid that has passed through the orifice plate over that amount of time.

This photograph shows a Rosemount 3095MV transmitter used to measure mass flow on an oxygen line. The orifice plate is an "integral" unit immediately below the transmitter body, sandwiched between two flange plates on the copper line. A three-valve manifold interfaces the model 3095MV transmitter to the integral orifice plate structure:

The temperature-compensation RTD may be clearly seen on the left-hand side of the photograph, installed at the elbow fitting in the copper pipe.

Liquid flow measurement applications may also benefit from compensation, because liquid density changes with temperature. Static pressure is not a concern here, because liquids are considered incompressible for all practical purposes[21]. Thus, the formula for compensated liquid flow measurement does not include any terms for static pressure, just differential pressure and temperature:

$$Q = N \frac{CYA_2}{\sqrt{1 - \left(\frac{A_2}{A_1}\right)^2}} \sqrt{(P_1 - P_2)[1 + k_T(T - T_{ref})]}$$

The constant k_T shown in the above equation is the proportionality factor for liquid expansion with increasing temperature. The difference in temperature between the measured condition (T) and the reference condition (T_{ref}) multiplied by this factor determines how much less dense the liquid is compared to its density at the reference temperature. It should be noted that some liquids – notably hydrocarbons – have thermal expansion factors significantly greater than water. This makes temperature compensation for hydrocarbon liquid flow measurement very important if the measurement principle is volumetric rather than mass-based.

[21] Liquids can and do compress, the measurement of their "compressibility" being what is called the *bulk modulus*. However, this compressibility is too slight to be of any consequence in most flow measurement applications.

21.1.9 Equation summary

Volumetric flow rate (Q) full equation:

$$Q = N \frac{CYA_2}{\sqrt{1 - \left(\frac{A_2}{A_1}\right)^2}} \sqrt{\frac{P_1 - P_2}{\rho_f}}$$

Volumetric flow rate (Q) simplified equation:

$$Q = k\sqrt{\frac{P_1 - P_2}{\rho_f}}$$

Mass flow rate (W):

$$W = N \frac{CYA_2}{\sqrt{1 - \left(\frac{A_2}{A_1}\right)^2}} \sqrt{\rho_f (P_1 - P_2)}$$

Mass flow rate (W) simplified equation:

$$W = k\sqrt{\rho_f (P_1 - P_2)}$$

Where,

Q = Volumetric flow rate (e.g. gallons per minute, flowing cubic feet per second)

W = Mass flow rate (e.g. kilograms per second, slugs per minute)

N = Unit conversion factor

C = Discharge coefficient (accounts for energy losses, Reynolds number corrections, pressure tap locations, etc.)

Y = Gas expansion factor ($Y = 1$ for liquids)

A_1 = Cross-sectional area of mouth

A_2 = Cross-sectional area of throat

ρ_f = Fluid density at flowing conditions (actual temperature and pressure at the element)

k = Constant of proportionality (determined by experimental measurements of flow rate, pressure, and density)

The beta ratio (β) of a differential-producing element is the ratio of throat diameter to mouth diameter ($\beta = \frac{d}{D}$). This is the primary factor determining acceleration as the fluid increases velocity entering the constricted throat of a flow element (venturi tube, orifice plate, wedge, etc.). The following expression is often called the *velocity of approach factor* (commonly symbolized as E_v), because it relates the velocity of the fluid through the constriction to the velocity of the fluid as it approaches the flow element:

$$E_v = \frac{1}{\sqrt{1 - \beta^4}} = \text{ Velocity of approach factor}$$

This same velocity approach factor may be expressed in terms of mouth and throat areas (A_1 and A_2, respectively):

$$E_v = \frac{1}{\sqrt{1 - \left(\frac{A_2}{A_1}\right)^2}} = \text{ Velocity of approach factor}$$

Beta ratio has a significant impact on the number of straight-run pipe lengths needed to condition the flow profile upstream and downstream of the flow element. Large beta ratios (where the bore diameter approaches the flowtube's inside diameter) are more sensitive to piping disturbances, since there is less acceleration of the flowstream through the element, and therefore flow profile asymmetries caused by piping disturbances are significant in comparison to the fluid's through-bore velocity. Small beta ratio values correspond to larger acceleration factors, where disturbances in the flow profile become "swamped[22]" by the high throat velocities created by the element's constriction. A disadvantage of small beta ratio values is that the flow element exhibits a greater permanent pressure loss, which is an operational cost if the flow is provided by a machine such as an engine- or motor-driven pump (more energy required to turn the pump, equating to a greater operating cost to run the process).

[22] "Swamping" is a term commonly used in electrical engineering, where a bad effect is overshadowed by some other effect much larger in magnitude, to the point where the undesirable effect is negligible in comparison.

When computing the volumetric flow of a gas in *standard* volume units (e.g. SCFM), the equation becomes much more complex than the simple (flowing) volumetric rate equation. Any equation computing flow in standard units must predict the effective expansion of the gas if it were to transition from flowing conditions (the actual pressure and temperature it experiences flowing through the pipe) to standard conditions (one atmosphere pressure at 60 degrees Fahrenheit). The compensated gas flow measurement equation published by the American Gas Association (AGA Report #3) in 1992 for orifice plates with flange taps calculates this expansion to standard conditions with a series of factors accounting for flowing and standard ("base") conditions, in addition to the more common factors such as velocity of approach and gas expansion. Most of these factors are represented in the AGA3 equation by different variables beginning with the letter F:

$$Q = F_n(F_c + F_{sl})Y F_{pb} F_{tb} F_{tf} F_{gr} F_{pv} \sqrt{h_W P_{f1}}$$

Where,

Q = Volumetric flow rate (standard cubic feet per hour – SCFH)

F_n = Numeric conversion factor (accounts for certain numeric constants, unit-conversion coefficients, and the velocity of approach factor E_v)

F_c = Orifice calculation factor (a polynomial function of the orifice plate's β ratio and Reynolds number), appropriate for flange taps

F_{sl} = Slope factor (another polynomial function of the orifice plate's β ratio and Reynolds number), appropriate for flange taps

$F_c + F_{sl} = C_d$ = Discharge coefficient, appropriate for flange taps

Y = Gas expansion factor (a function of β, differential pressure, static pressure, and specific heats)

F_{pb} = Base pressure factor = $\frac{14.73\ \text{PSI}}{P_b}$, with pressure in PSIA (absolute)

F_{tb} = Base temperature factor = $\frac{T_b}{519.67}$, with temperature in degrees Rankine

F_{tf} = Flowing temperature factor = $\sqrt{\frac{519.67}{T_f}}$, with temperature in degrees Rankine

F_{gr} = Real gas relative density factor = $\sqrt{\frac{1}{G_r}}$

F_{pv} = Supercompressibility factor = $\sqrt{\frac{Z_b}{Z_{f1}}}$

h_W = Differential pressure produced by orifice plate (inches water column)

P_{f1} = Flowing pressure of gas at the upstream tap (PSI absolute)

21.2 Laminar flowmeters

A unique form of differential pressure-based flow measurement deserves its own section in this flow measurement chapter, and that is the *laminar* flowmeter.

Laminar flow is a condition of fluid motion where viscous (internal fluid friction) forces greatly overshadow inertial (kinetic) forces. A flowstream in a state of laminar flow exhibits no turbulence, with each fluid molecule traveling in its own path, with limited mixing and collisions with adjacent molecules. The dominant mechanism for resistance to fluid motion in a laminar flow regime is friction with the pipe or tube walls. Laminar flow is qualitatively predicted by low values of Reynolds number.

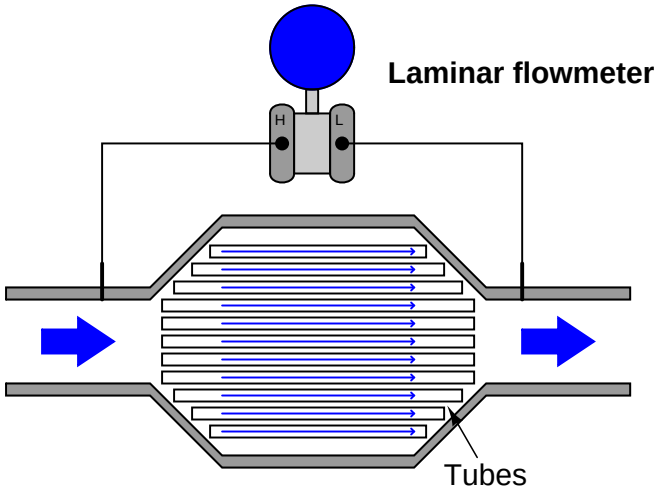
This pressure drop created by fluid friction in a laminar flowstream is quantifiable, and is expressed in the Hagen-Poiseuille equation:

$$Q = k\left(\frac{\Delta P D^4}{\mu L}\right)$$

Where,

Q = Flow rate
ΔP = Pressure dropped across a length of pipe
D = Pipe diameter
μ = Fluid viscosity
L = Pipe length
k = Coefficient accounting for units of measurement

Laminar flowmeter elements generally consist of one or more tubes whose length greatly exceeds the inside diameter, arranged in such a way as to produce a slow-moving flow velocity. An example is shown here:



The expanded diameter of the flow element ensures a lower fluid velocity than in the pipes entering and exiting the element. This decreases the Reynolds number to the point where the flow

regime exhibits laminar behavior. The large number of small-diameter tubes packed in the wide area of the element provide adequate wall surface area for the fluid's viscosity to act upon, creating an overall pressure drop from inlet to outlet which is measured by the differential pressure transmitter. This pressure drop is permanent (no recovery of pressure downstream) because the mechanism of pressure drop is friction: total dissipation (loss) of energy in the form of heat.

Another common form of laminar flow element is simply a coiled *capillary tube*: a long tube with a very small inside diameter. The small inside diameter of such a tube makes wall-boundary effects dominant, such that the flow regime will remain laminar over a wide range of flow rates. The extremely restrictive nature of a capillary tube, of course, limits the use of such flow elements to very low flow rates such as those encountered in the sampling networks of certain analytical instruments.

A unique advantage of the laminar flowmeter is its linear relationship between flow rate and developed pressure drop. It is the only pressure-based flow measurement device for filled pipes that exhibits a linear pressure/flow relationship. This means no "square-root" characterization is necessary to obtain linear flow measurements with a laminar flowmeter. The big disadvantage of this meter type is its dependence on fluid viscosity, which in turn is strongly influenced by fluid temperature. Thus, all laminar flowmeters require temperature compensation in order to derive accurate measurements, and some even use temperature *control* systems to force the fluid's temperature to be constant as it moves through the element[23].

Laminar flow elements find their widest application inside pneumatic instruments, where a linear pressure/flow relationship is highly advantageous (behaving like a "resistor" for instrument air flow) and the viscosity of the fluid (instrument air) is relatively constant. Pneumatic controllers, for instance, use laminar restrictors as part of the derivative and integral calculation modules, the combination of "resistance" from the restrictor and "capacitance" from volume chambers forming a sort of pneumatic time-constant (τ) network.

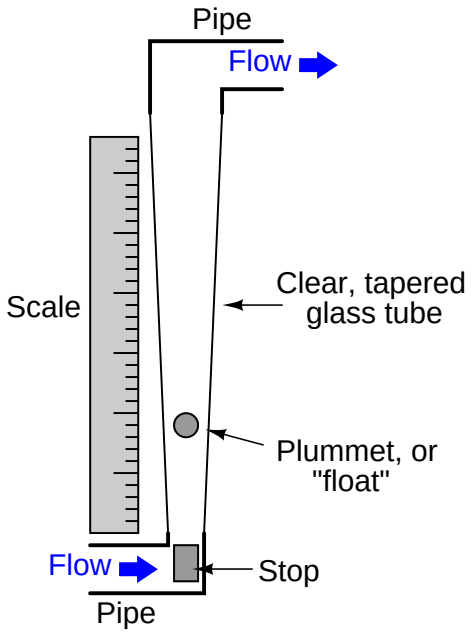
21.3 Variable-area flowmeters

An *Variable-area* flowmeter is one where the fluid must pass through a restriction whose area increases with flow rate. This stands in contrast to flowmeters such as orifice plates and venturi tubes where the cross-sectional area of the flow element remains fixed.

[23] This includes elaborate oil-bath systems where the laminar flow element is submerged in a temperature-controlled oil bath, the purpose of which is to hold temperature inside the laminar element constant despite sudden changes in the measured fluid's temperature.

21.3.1 Rotameters

The simplest example of a variable-area flowmeter is the *rotameter*, which uses a solid object (called a *plummet* or *float*) as a flow indicator, suspended in the midst of a tapered tube:



As fluid flows upward through the tube, a pressure differential develops across the plummet. This pressure differential, acting on the effective area of the plummet body, develops an upward force ($F = \frac{P}{A}$). If this force exceeds the weight of the plummet, the plummet moves up. As the plummet moves further up in the tapered tube, the area between the plummet and the tube walls (through which the fluid must travel) grows larger. This increased flowing area allows the fluid to make it past the plummet without having to accelerate as much, thereby developing less pressure drop across the plummet's body. At some point, the flowing area reaches a point where the pressure-induced force on the plummet body exactly matches the weight of the plummet. This is the point in the tube where the plummet stops moving, indicating flow rate by it position relative to a scale mounted (or etched) on the outside of the tube.

The following rotameter uses a spherical plummet, suspended in a flow tube machined from a solid block of clear plastic. An adjustable valve at the bottom of the rotameter provides a means for adjusting gas flow:

The same basic flow equation used for pressure-based flow elements holds true for rotameters as well:

$$Q = k\sqrt{\frac{P_1 - P_2}{\rho}}$$

However, the difference in this application is that the value inside the radicand is constant, since the pressure difference will remain constant[24] and the fluid density will likely remain constant as well. Thus, k will change in proportion to Q. The only variable within k relevant to plummet position is the flowing area between the plummet and the tube walls.

[24] If we know that the plummet's weight will remain constant, its area will remain constant, and that the force generated by the pressure drop will always be in equilibrium with the plummet's weight for any steady flow rate, then the relationship $F = \frac{P}{A}$ dictates a constant pressure. Thus, we may classify the rotameter as a *constant-pressure*, *variable-area* flowmeter. This stands in contrast to devices such as orifice plates, which are *variable-pressure*, *constant-area*.

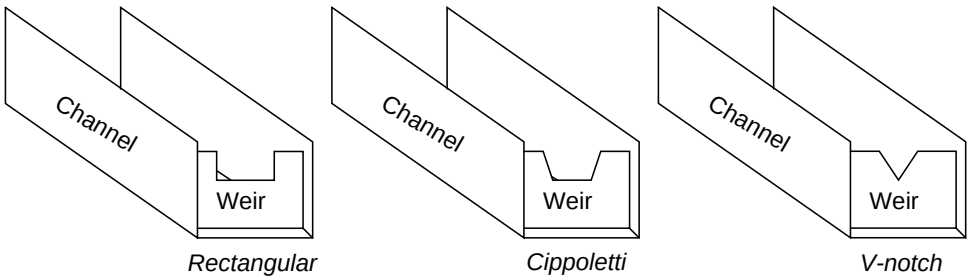
Most rotameters are indicating devices only. They may be equipped to transmit flow information electronically by adding sensors to detect the plummet's position in the tube, but this is not common practice.

Rotameters are very commonly used as purge flow indicators for pressure and level measurement systems requiring a constant flow of purge fluid (see pages 832 and 861 for examples). Such rotameters are usually equipped with hand-adjustable needle valves for manual regulation of purge fluid flow rate.

21.3.2 Weirs and flumes

A very different style of variable-area flowmeter is used extensively to measure flow rate through open channels, such as irrigation ditches. If an obstruction is placed within a channel, any liquid flowing through the channel must rise on the upstream side of the obstruction. By measuring this liquid level rise, it is possible to infer the rate of liquid flow past the obstruction.

The first form of open-channel flowmeter is the *weir*, which is nothing more than a dam obstructing passage of liquid through the channel. Three styles of weir are shown in the following illustration; the *rectangular*, *Cippoletti*, and *V-notch*:

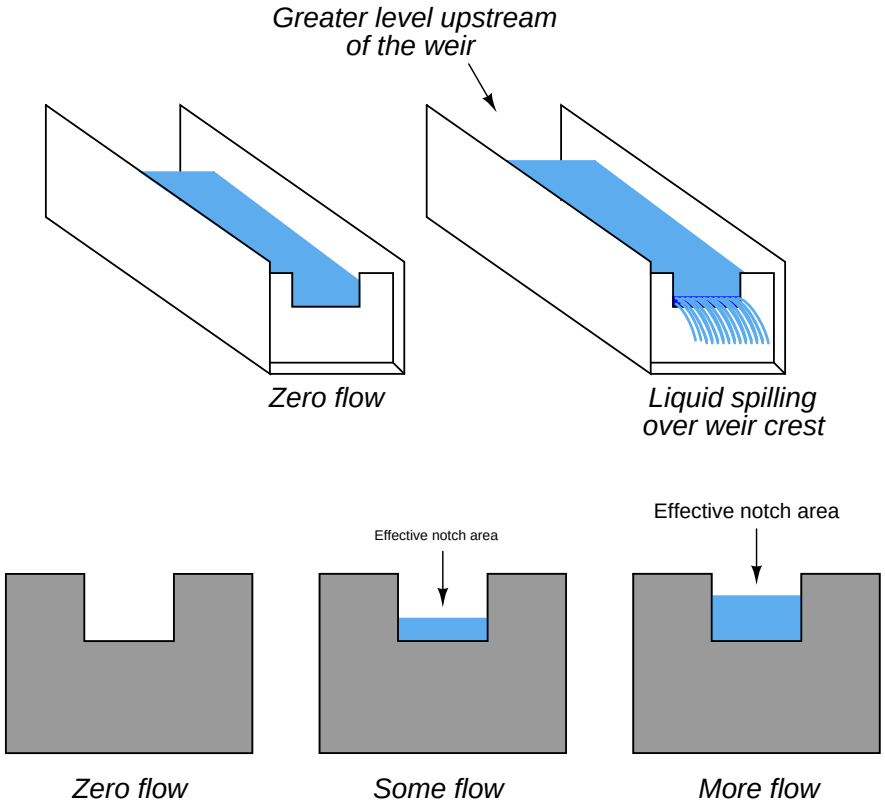


A rectangular weir has a notch of simple rectangular shape, as the name implies. A Cippoletti weir is much like a rectangular weir, except that the vertical sides of the notch have a 4:1 slope (rise of 4, run of 1; approximately a 14 degree angle from vertical). A V-notch weir has a triangular notch, customarily measuring either 60 or 90 degrees.

The following photograph shows water flowing through a Cippoletti weir made of 1/4 inch steel plate:

At a condition of zero flow through the channel, the liquid level will be at or below the crest (lowest point on the opening) of the weir. As liquid begins to flow through the channel, it must spill over the crest of the weir in order to get past the weir and continue downstream in the channel. In order for this to happen, the level of the liquid upstream of the weir must rise above the weir's crest height. This height of liquid upstream of the weir represents a hydrostatic pressure, much the same as liquid heights in piezometer tubes represent pressures in a liquid flowstream through an enclosed pipe (see page 129 for examples of this). The height of liquid above the crest of a weir is analogous

to the pressure differential generated by an orifice plate. As liquid flow is increased even more, a greater pressure (head) will be generated upstream of the weir, forcing the liquid level to rise. This effectively increases the cross-sectional area of the weir's "throat" as a taller stream of liquid exits the notch of the weir[25].



[25] Orifice plates are *variable-pressure*, *constant-area* flowmeters. Rotameters are *constant-pressure*, *variable-area* flowmeters. Weirs are *variable-pressure*, *variable-area* flowmeters. As one might expect, the mathematical functions describing each of these flowmeter types is unique!

This dependence of notch area on flow rate creates a very different relationship between flow rate and liquid height (measured above the crest) than the relationship between flow rate and differential pressure in an orifice plate:

$$Q = 3.33(L - 0.2H)H^{1.5} \qquad \text{Rectangular weir}$$

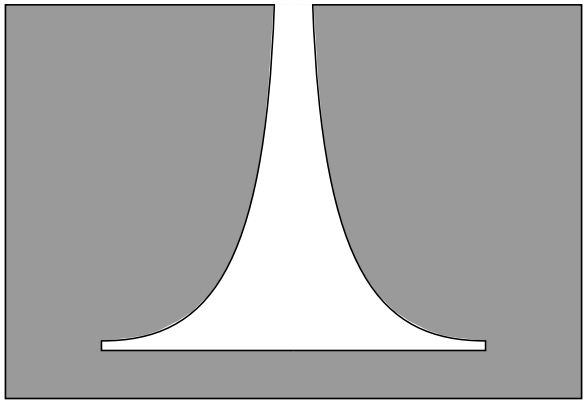
$$Q = 3.367LH^{1.5} \qquad \text{Cippoletti weir}$$

$$Q = 2.48\left(\tan\frac{\theta}{2}\right)H^{2.5} \qquad \text{V-notch weir}$$

Where,

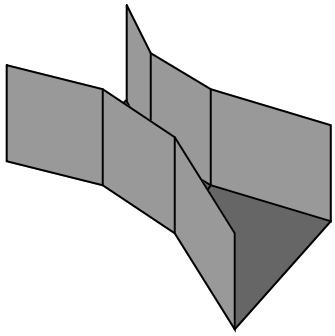
Q = Volumetric flow rate (cubic feet per second – CFS)
L = Width of crest (feet)
θ = V-notch angle (degrees)
H = Head (feet)

As you can see from a comparison of characteristic flow equations between these three types of weirs, the shape of the weir's notch has a dramatic effect on the mathematical relationship between flow rate and head (liquid level upstream of the weir, measured above the crest height). This implies that it is possible to create almost any characteristic equation we might like just by carefully shaping the weir's notch in some custom form. A good example of this is the so-called *proportional* or *Sutro* weir:



This weir design is not used very often, due to its inherently weak structure and tendency to clog with debris.

A variation on the theme of a weir is another open-channel device called a *flume*. If weirs may be thought of as open-channel orifice plates, then flumes may be thought of as open-channel venturi tubes:



Like weirs, flumes generate upstream liquid level height changes indicative of flow rate. One of the most common flume design is the *Parshall flume*, named after its inventor R.L. Parshall when it was developed in the year 1920.

The following formulae relate head (upstream liquid height) to flow rate for free-flowing Parshall flumes[26]:

$$Q = 0.992H^{1.547} \qquad \text{3-inch wide throat Parshall flume}$$

$$Q = 2.06H^{1.58} \qquad \text{6-inch wide throat Parshall flume}$$

$$Q = 3.07H^{1.53} \qquad \text{9-inch wide throat Parshall flume}$$

$$Q = 4LH^{1.53} \qquad \text{1-foot to 8-foot wide throat Parshall flume}$$

$$Q = (3.6875L + 2.5)H^{1.53} \qquad \text{10-foot to 50-foot wide throat Parshall flume}$$

Where,

Q = Volumetric flow rate (cubic feet per second – CFS)

L = Width of flume throat (feet)

H = Head (feet)

Flumes are generally less accurate than weirs, but they do enjoy the advantage of being inherently self-cleaning. If the liquid stream being measured is drainage- or waste-water, a substantial amount of solid debris may be present in the flow that could cause repeated clogging problems for weirs. In such applications, flumes are often the more practical flow element for the task (and more accurate

[26] It is also possible to operate a Parshall flume in fully *submerged* mode, where liquid level must be measured at both the upstream and throat sections of the flume. Correction factors must be applied to these equations if the flume is submerged.

over the long term as well, since even the finest weir will not register accurately once fouled by debris).

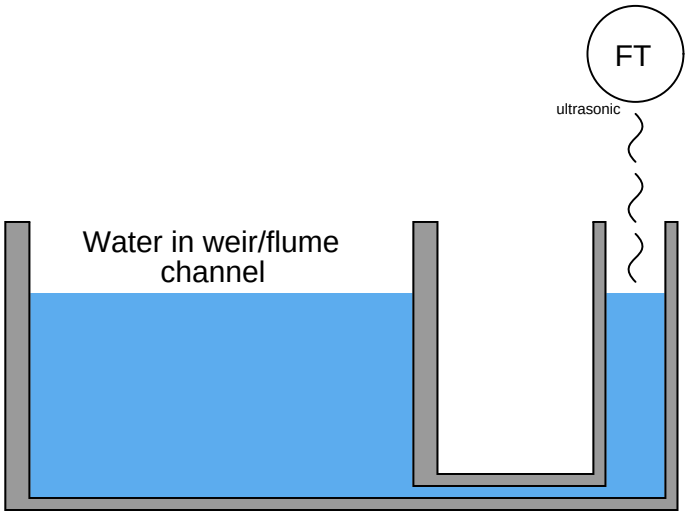
Once a weir or flume has been installed in an open channel to measure the flow of liquid, some method must be employed to sense upstream liquid level and translate this level measurement into a flow measurement. Perhaps the most common technology for weir/flume level sensing is *ultrasonic* (see section 19.5.1, page number 899, for more information on how this technology works). Ultrasonic level sensors are completely non-contact, which means they cannot become fouled by the process liquid (or debris in the process liquid). However, they may be "fooled" by foam or debris floating on top of the liquid, as well as waves on the liquid surface.

The following photograph shows a Parshall flume measuring effluent flow from a municipal sewage treatment plant, with an ultrasonic transducer mounted above the middle of the flume to detect water level flowing through:

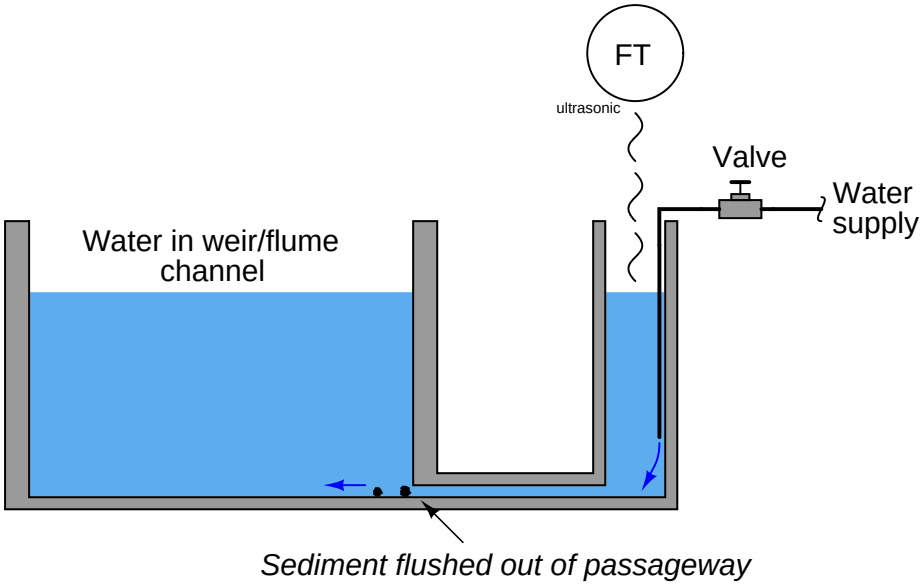
Once the liquid level is successfully measured, a computing device is used to translate that level measurement into a suitable flow measurement (and in some cases even integrate that flow measurement with respect to time to arrive at a value for total liquid volume passed through the element, in accordance with the calculus relationship $V = \int Q\,dt + C$).

A technique for providing a clean and "quiet" (still) liquid surface to measure the level of is called a *stilling well*. This is an open-top chamber connected to the weir/flume channel by a pipe, so the liquid level in the stilling well matches the liquid level in the channel. The following illustration shows a stilling well connected to a weir/flume channel, with the direction of liquid flow in the

channel being perpendicular to the page (i.e. either coming toward your eyes or going away from your eyes):

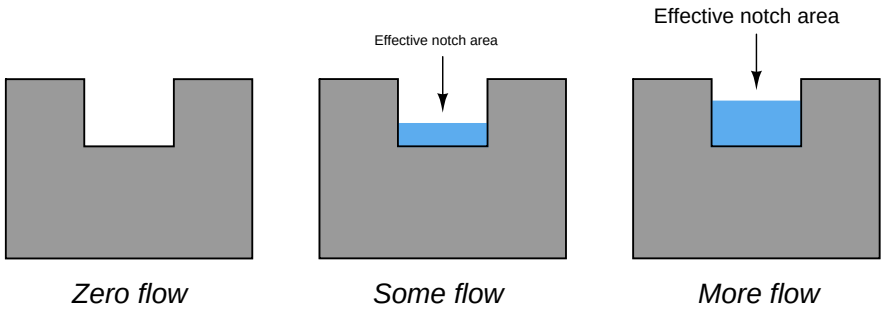


To discourage plugging of the passageway connecting the stilling well to the channel, a small flow rate of clean water may be introduced into the well. This forms a constant *purge flow* into the channel, flushing out debris that might otherwise find its way into the connecting passageway to plug it up. Note how the purge water enters the stilling well through a submerged tube, so it does not cause splashing on the water's surface inside the well which could cause measurement problems for the ultrasonic sensor:



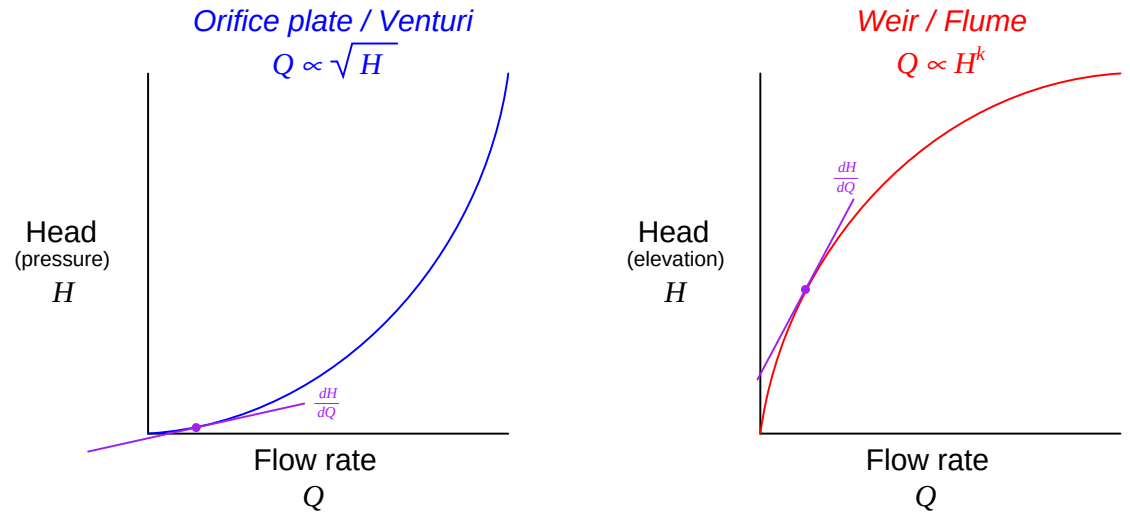
A significant advantage that weirs and flumes have over other forms of flow measurement is exceptionally high *rangeability*: the ability to measure very wide ranges of flow with a modest pressure (height) span. Another way to state this is to say that the accuracy of a weir or flume is quite high even at low flow rates.

Earlier in this section you saw a three-image representation of liquid flow through a rectangular weir. As fluid flow rate increased, so did the height (head) of the liquid upstream of the weir:



The height of liquid upstream of the weir depends on the flow rate (volumetric Q or mass W) as well as the effective area of the notch through which the fluid must pass. Unlike an orifice plate, this area changes with flow rate in both weirs and flumes. One way to envision this by comparison is to imagine a weir as acting like an elastic orifice plate, whose bore area increases with flow rate. This flow-dependent notch area exhibited by both weirs and flumes means that these devices become *more sensitive* to changes in flow as the flow rate becomes smaller.

A comparison of transfer function graphs for closed-pipe head elements such as orifice plates and venturi tubes versus weirs and flumes shows this striking difference in characteristics:



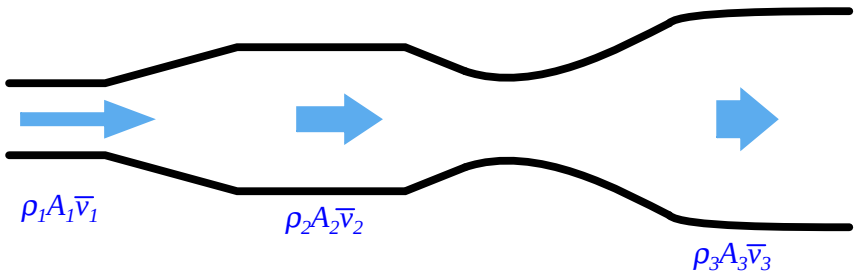
Looking at the orifice plate / venturi tube graph near the lower-left corner, you can see how small changes in flow result in extremely small changes in head (differential pressure), because the

function has a very low slope (small $\frac{dH}{dQ}$) at that end. By comparison, a weir or flume produces relatively large changes in head (liquid elevation) for small changes in flow near the bottom end of the range, because the function has a very steep slope (large $\frac{dH}{dQ}$) at that end.

The practical advantage this gives weirs and flumes is the ability to maintain high accuracy of flow measurement at very low flow rates – something a fixed-orifice element simply cannot do. It is commonly understood in industry that traditional orifice plate flowmeters cannot maintain good measurement accuracy much below a third of their full-range flow (a rangeability of 3:1), whereas weirs (especially the V-notch design) can achieve far greater rangeability (up to 500:1 according to some sources[27]).

21.4 Velocity-based flowmeters

The Law of Continuity for fluids states that the product of mass density (ρ), cross-sectional pipe area (A) and average velocity ($\overline{v}$) must remain constant through any continuous length of pipe:



If the density of the fluid is not subject to change as it travels through the pipe (a very good assumption for liquids), we may simplify the Law of Continuity by eliminating the density terms from the equation:

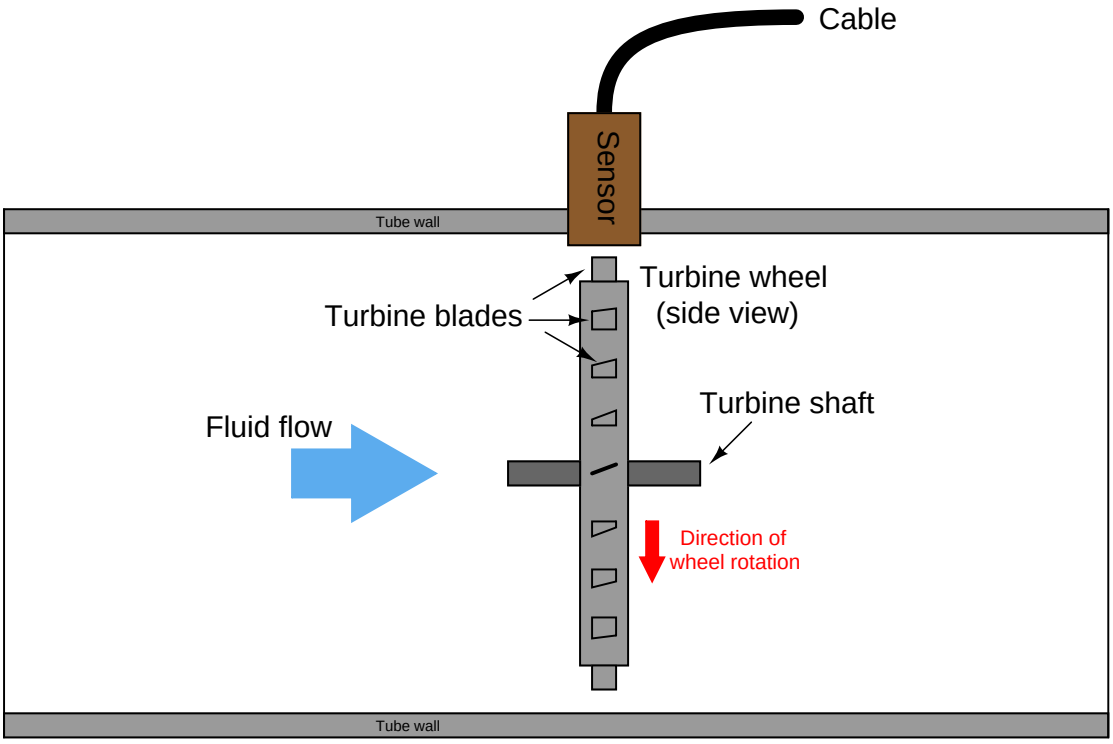
$$A_1 \overline{v_1} = A_2 \overline{v_2}$$

The product of cross-sectional pipe area and average fluid velocity is the volumetric flow rate of the fluid through the pipe ($Q = A\overline{v}$). This tells us that fluid velocity will be directly proportional to volumetric flow rate given a known cross-sectional area and a constant density for the fluid flowstream. Any device able to directly measure fluid velocity is therefore capable of inferring volumetric flow rate of fluid in a pipe. This is the basis for *velocity-based* flowmeter designs.

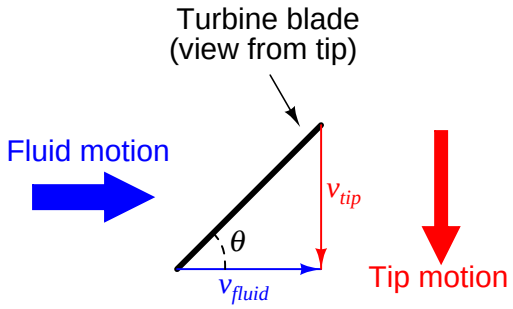
[27] These figures are reported in Béla Lipták's excellent reference book *Instrument Engineers' Handbook – Process Measurement and Analysis Volume I* (Fourth Edition). To be fair to closed-pipe elements such as orifice plates and venturi tubes, much improvement in the classic 3:1 rangeability limitation has been achieved through the use of microprocessor-based differential pressure sensors. Lipták reports rangeabilities for orifice plates as great as 10:1 through the use of such modern differential pressure instruments. However, even this pales in comparison to the rangeability of a typical weir or flume, which Lipták reports to be 75:1 for "most devices" in this category.

21.4.1 Turbine flowmeters

Turbine flowmeters use a free-spinning turbine wheel to measure fluid velocity, much like a miniature windmill installed in the flow stream. The fundamental design goal of a turbine flowmeter is to make the turbine element as free-spinning as possible, so no torque will be required to sustain the turbine's rotation. If this goal is achieved, the turbine blades will achieve a rotating (tip) velocity directly proportional to the linear velocity of the fluid:



The mathematical relationship between fluid velocity and turbine tip velocity – assuming frictionless conditions – is a ratio defined by the *tangent* of the turbine blade angle:

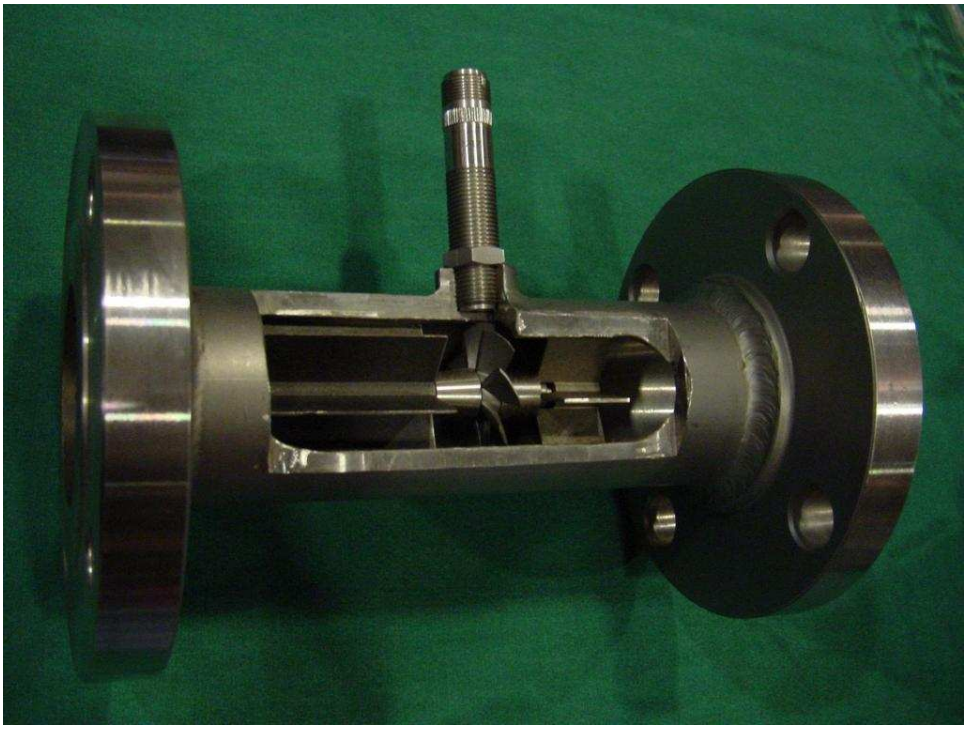


$$\tan \theta = \frac{v_{tip}}{v_{fluid}}$$

For a 45^o blade angle, the relationship is 1:1, with tip velocity equaling fluid velocity. Smaller blade angles (each blade closer to parallel with the fluid velocity vector) results in the tip velocity being a fractional proportion of fluid velocity.

Turbine tip velocity is quite easy to sense using a magnetic sensor, generating a voltage pulse each time one of the ferromagnetic turbine blades passes by. Traditionally, this sensor is nothing more than a coil of wire in proximity to a stationary magnet, called a *pickup coil* or *pickoff coil* because it "picks" (senses) the passing of the turbine blades. Magnetic flux through the coil's center increases and decreases as the passing of the steel turbine blades presents a varying reluctance ("resistance" to magnetic flux), causing voltage pulses equal in frequency to the number of blades passing by each second. It is the *frequency* of this signal that represents fluid velocity, and therefore volumetric flow rate.

A cut-away demonstration model of a turbine flowmeter is shown in the following photograph. The blade sensor may be seen protruding from the top of the flowtube, just above the turbine wheel:



Note the sets of "flow conditioner" vanes immediately before and after the turbine wheel in the photograph. As one might expect, turbine flowmeters are very sensitive to *swirl* in the process fluid flowstream. In order to achieve high accuracy, the flow profile must not be swirling in the vicinity of the turbine, lest the turbine wheel spin faster or slower than it should to represent the velocity of a straight-flowing fluid.

Mechanical gears and rotating cables have also been historically used to link a turbine flowmeter's turbine wheel to indicators. These designs suffer from greater friction than electronic ("pickup coil") designs, potentially resulting in more measurement error (less flow indicated than there actually is, because the turbine wheel is slowed by friction). One advantage of mechanical turbine flowmeters, though, is the ability to maintain a running total of gas usage by turning a simple odometer-style totalizer. This design is often used when the purpose of the flowmeter is to track total fuel gas consumption (e.g. natural gas used by a commercial or industrial facility) for billing.

In an electronic turbine flowmeter, volumetric flow is directly proportional to pickup coil output frequency. We may express this relationship in the form of an equation:

$$f = kQ$$

Where,
f = Frequency of output signal (Hz, equivalent to pulses per second)
Q = Volumetric flow rate (e.g. gallons per second)
k = "K" factor of the turbine element (e.g. pulses per gallon)

Dimensional analysis confirms the validity of this equation. Using units of GPS (gallons per second) and pulses per gallon, we see that the product of these two quantities is indeed pulses per second (equivalent to cycles per second, or Hz):

$$\left[\frac{\text{Pulses}}{\text{s}}\right] = \left[\frac{\text{Pulses}}{\text{gal}}\right]\left[\frac{\text{gal}}{\text{s}}\right]$$

Using algebra to solve for flow (Q), we see that it is the quotient of frequency and k factor that yields a volumetric flow rate for a turbine flowmeter:

$$Q = \frac{f}{k}$$

If pickup signal frequency directly represents volumetric flow rate, then the total number of pulses accumulated in any given time span will represent the amount of fluid volume passed through the turbine meter over that same time span. We may express this algebraically as the product of average flow rate ($\overline{Q}$), average frequency ($\overline{f}$), k factor, and time:

$$V = \overline{Q}t = \frac{\overline{f}t}{k}$$

A more sophisticated way of calculating total volume passed through a turbine meter requires calculus, representing total volume as the time-integral of instantaneous signal frequency and k factor over a period of time from $t = 0$ to $t = T$:

$$V = \int_0^T Q\,dt \qquad \text{or} \qquad V = \int_0^T \frac{f}{k}\,dt$$

We may achieve approximately the same result simply by using a digital counter circuit to totalize pulses output by the pickup coil and a microprocessor to calculate volume in whatever unit of measurement we deem appropriate.

As with the orifice plate flow element, standards have been drafted for the use of turbine flowmeters as precision measuring instruments in gas flow applications, particularly the custody transfer[28] of natural gas. The American Gas Association has published a standard called the Report

[28] "Custody transfer" refers to measurement applications where a product is exchanging ownership. In other words, someone is selling, and someone else is buying, quantities of fluid as part of a business transaction. It is not difficult to understand why accuracy is important in such applications, as both parties have a vested interest in a fair exchange. Government institutions also have a stake in accurate metering, as taxes are typically levied on the sale of commodity fluids such as natural gas.

#7 specifying the installation of turbine flowmeters for high-accuracy gas flow measurement, along with the associated mathematics for precisely calculating flow rate based on turbine speed, gas pressure, and gas temperature.

Pressure and temperature compensation is relevant to turbine flowmeters in gas flow applications because the density of the gas is a strong function of both pressure and temperature. The turbine wheel itself only senses gas *velocity*, and so these other factors must be taken into consideration to accurately calculate mass flow (or *standard* volumetric flow; e.g. SCFM).

In high-accuracy applications, it is important to individually determine the k factor for a turbine flowmeter's calibration. Manufacturing variations from flowmeter to flowmeter make precise duplication of k factor challenging, and so a flowmeter destined for high-accuracy measurement should be tested against a "flow prover" in a calibration laboratory to empirically determine its k factor. If possible, the best way to test the flowmeter's k factor is to connect the prover to the meter on site where it will be used. This way, the any effects due to the piping before and after the flowmeter will be incorporated in the measured k factor.

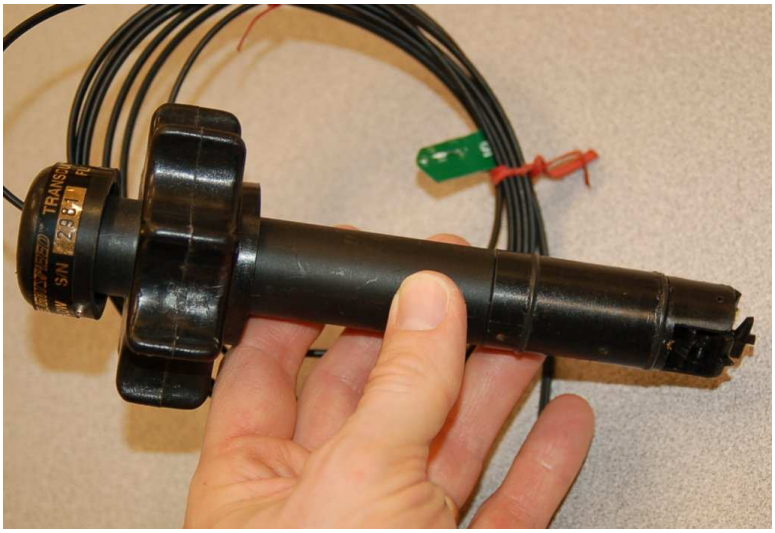
The following photograph shows three AGA7-compliant installations of turbine flowmeters for measuring the flow rate of natural gas:

Note the pressure-sensing and temperature-sensing instrumentation installed in the pipe, reporting gas pressure and gas temperature to a flow-calculating computer (along with turbine pulse frequency) for the calculation of natural gas flow rate.

Less-critical gas flow measurement applications may use a "compensated" turbine flowmeter that mechanically performs the same pressure- and temperature-compensation functions on turbine speed to achieve true gas flow measurement, as shown in the following photograph:

The particular flowmeter shown in the above photograph uses a filled-bulb temperature sensor (note the coiled, armored capillary tube connecting the flowmeter to the bulb) and shows total gas flow by a series of pointers, rather than gas flow *rate*.

A variation on the theme of turbine flow measurement is the *paddlewheel* flowmeter, a very inexpensive technology usually implemented in the form of an insertion-type sensor. In this instrument, a small wheel equipped with "paddles" parallel to the shaft is inserted in the flowstream, with half the wheel shrouded from the flow. A photograph of a plastic paddlewheel flowmeter appears here:



A surprisingly sophisticated method of "pickup" for the plastic paddlewheel shown in the photograph uses *fiber-optic cables* to send and receive light. One cable sends a beam of light to the edge of the paddlewheel, and the other cable receives light on the other side of the paddlewheel. As the paddlewheel turns, the paddles alternately block and pass the light beam, resulting in a pulsed light beam at the receiving cable. The frequency of this pulsing is, of course, directly proportional to volumetric flow rate.

The external ends of the two fiber optic cables appear in this next photograph, ready to connect to a light source and light pulse sensor to convert the paddlewheel's motion into an electronic signal:

A problem common to all turbine flowmeters is that of the turbine "coasting" when the fluid flow suddenly stops. This is more often a problem in batch processes than continuous processes, where the fluid flow is regularly turned on and shut off. This problem may be minimized by configuring the measurement system to ignore turbine flowmeter signals any time the automatic shutoff valve reaches the "shut" position. This way, when the shutoff valve closes and fluid flow immediately halts, any coasting of the turbine wheel will be irrelevant. In processes where the fluid flow happens to pulse for reasons other than the control system opening and shutting automatic valves, this problem is more severe.

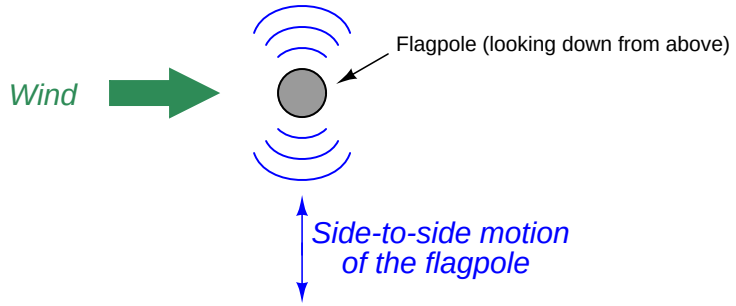
Another problem common to all turbine flowmeters is lubrication of the turbine bearings. Frictionless motion of the turbine wheel is essential for accurate flow measurement, which is a daunting design goal for the flowmeter manufacturing engineers. The problem is not as severe in applications where the process fluid is naturally lubricating (e.g. diesel fuel), but in applications such as natural gas flow where the fluid provides no lubrication to the turbine bearings, external lubrication must be supplied. This is often a regular maintenance task for instrument technicians: using a hand pump to inject light-weight "turbine oil" into the bearing assemblies of turbine flowmeters used in gas service.

Process fluid viscosity is another source of friction for the turbine wheel. Fluids with high viscosity (e.g. heavy oils) will tend to slow down the turbine's rotation even if the turbine rotates on frictionless bearings. This effect is especially pronounced at low flow rates, which leads to a *minimum linear flow* rating for the flowmeter: a flowrate below which it refuses to register proportionately to fluid flow rate.

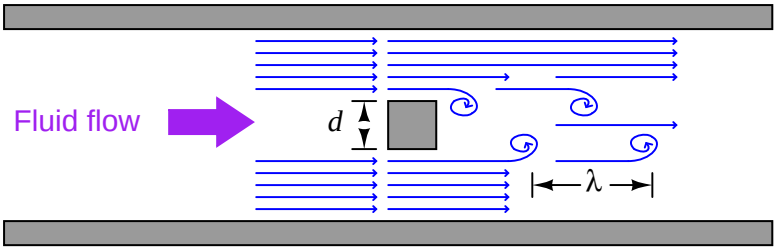
21.4.2 Vortex flowmeters

When a fluid moves with high Reynolds number past a stationary object (a "bluff body"), there is a tendency for the fluid to form *vortices* on either side of the object. Each vortex will form, then detach from the object and continue to move with the flowing gas or liquid, one side at a time in alternating fashion. This phenomenon is known as *vortex shedding*, and the pattern of moving vortices carried downstream of the stationary object is known as a *vortex street*.

It is commonplace to see the effects of vortex shedding on a windy day by observing the motion of flagpoles, light poles, and tall smokestacks. Each of these objects has a tendency to oscillate perpendicular to the direction of the wind, owing to the pressure variations caused by the vortices as they alternately form and break away from the object:



This alternating series of vortices was studied by Vincenc Strouhal in the late nineteenth century and later by Theodore von Kármán in the early twentieth century. It was determined that the distance between successive vortices downstream of the stationary object is relatively constant, and directly proportional to the width of the object, for a wide range of Reynolds number values[29]. If we view these vortices as crests of a continuous wave, the distance between vortices may be represented by the symbol customarily reserved for wavelength: the Greek letter "lambda" (λ).

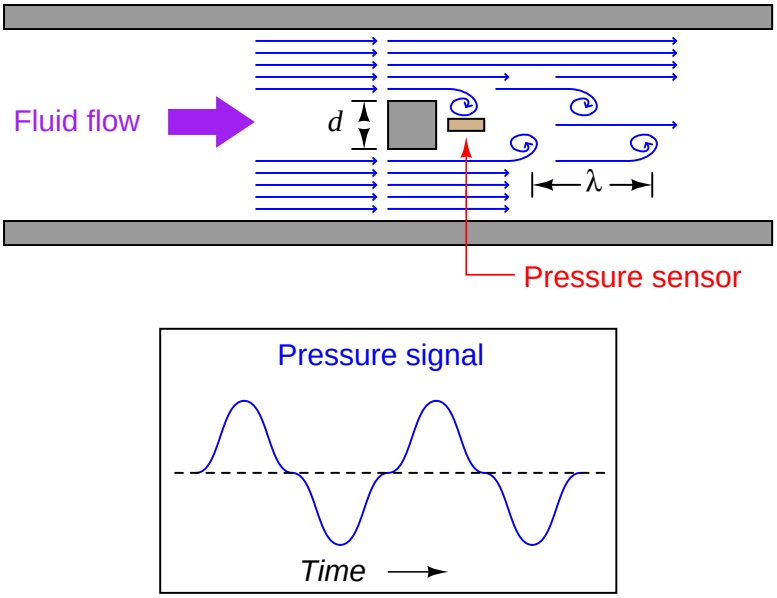


The proportionality between object width (d) and vortex street wavelength (λ) is called the *Strouhal number* (S), approximately equal to 0.17:

$$\lambda S = d \qquad\qquad \lambda \approx \frac{d}{0.17}$$

[29] It is important to note that the vortex-shedding phenomenon ceases altogether if the Reynolds number is too low. Laminar flow produces no vortices, but rather stream-line flow around any object placed in its way.

If a differential pressure sensor is installed immediately downstream of the stationary object in such an orientation that it detects the passing vortices as pressure variations, an alternating signal will be detected:



The *frequency* of this alternating pressure signal is directly proportional to fluid velocity past the object, since the wavelength is constant. This follows the classic frequency-velocity-wavelength formula common to all traveling waves ($\lambda f = v$). Since we know the wavelength will be equal to the bluff body's width divided by the Strouhal number (approximately 0.17), we may substitute this into the frequency-velocity-wavelength formula to solve for fluid velocity (v) in terms of signal frequency (f) and bluff body width (d).

$$v = \lambda f$$

$$v = \frac{d}{0.17} f$$

$$v = \frac{df}{0.17}$$

Thus, a stationary object and pressure sensor installed in the middle of a pipe section constitute a form of flowmeter called a *vortex flowmeter*. Like a turbine flowmeter with an electronic "pickup" sensor to detect the passage of rotating turbine blades, the output frequency of a vortex flowmeter is linearly proportional to volumetric flow rate.

The pressure sensors used in vortex flowmeters are not standard differential pressure transmitters, since the vortex frequency is too high to be successfully detected by such bulky instruments. Instead, the sensors are typically piezoelectric crystals. These pressure sensors need not be calibrated, since the amplitude of the pressure waves detected is irrelevant. Only the frequency of the waves matter

for measuring flow rate, and so nearly any pressure sensor with a fast enough response time will suffice.

Like turbine meters, the relationship between sensor frequency (f) and volumetric flow rate (Q) may be expressed as a proportionality, with the letter k used to represent the constant of proportionality for any particular flowmeter:

$$f = kQ$$

Where,

f = Frequency of output signal (Hz)

Q = Volumetric flow rate (e.g. gallons per second)[30]

k = "K" factor of the vortex shedding flowtube (e.g. pulses per gallon)

This means vortex flowmeters, like electronic turbine meters, each have a particular "k factor" relating the number of pulses generated per unit volume passed through the meter[31]. Counting the total number of pulses over a certain time span yields total fluid volume passed through the meter over that same time span, making the vortex flowmeter readily adaptable for "totalizing" fluid volume just like turbine meters.

Since vortex flowmeters have no moving parts, they do not suffer the problems of wear and lubrication facing turbine meters. There is no moving element to "coast" as in a turbine flowmeter if fluid flow suddenly stops, which means vortex flowmeters are better suited to measuring erratic flows.

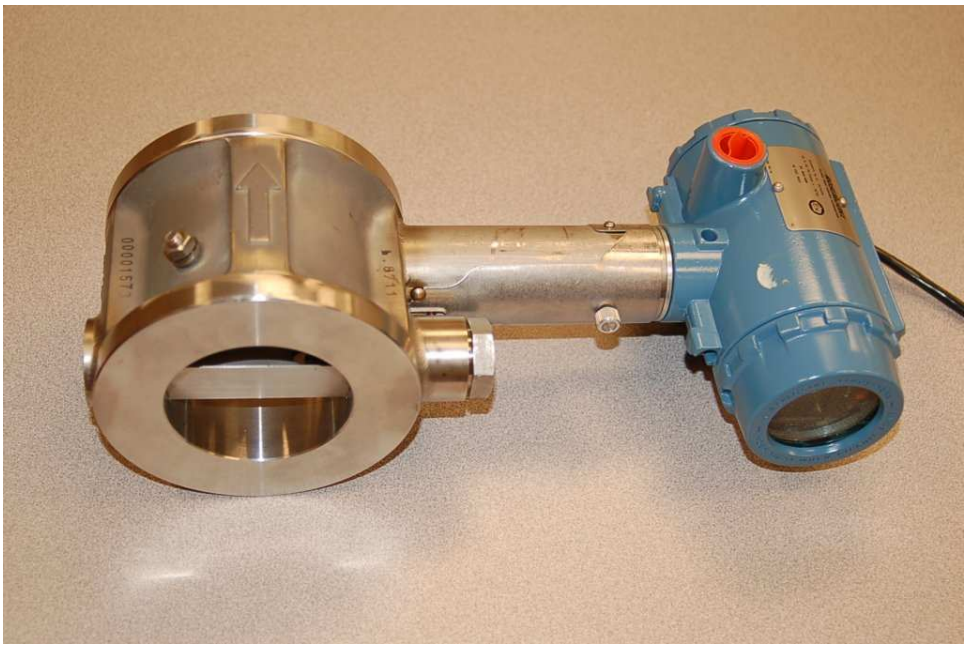
A significant disadvantage of vortex meters is a behavior known as *low flow cutoff*, where the flowmeter simply stops working below a certain flow rate. The reason for this is the cessation of vortices when the fluid's Reynolds number drops below a critical value and the flow regime passes from turbulent to laminar. When the flow is laminar, fluid viscosity is sufficient to prevent vortices from forming, causing the vortex flowmeter to register zero flow even when there may be some (laminar) flow through the pipe.

The phenomenon of low-flow cutoff for a vortex flowmeter at first seems analogous to the *minimum linear flow* limitation of a turbine flowmeter. However, vortex flowmeter low-flow cutoff is actually a far more severe problem. If the volumetric flow rate through a turbine flowmeter falls below the minimum linear value, the turbine continues to spin, albeit slower than it should. If the volumetric flow rate through a vortex flowmeter falls below the low-flow cutoff value, the flowmeter's signal *goes completely to zero*, indicating no flow at all.

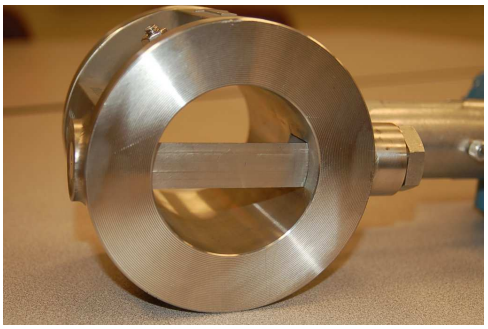
[30] Note that if flow rate is to be expressed in units of gallons per *minute* as is customary, the equation must contain a factor for minutes-to-seconds conversion: $f = \frac{kQ}{60}$

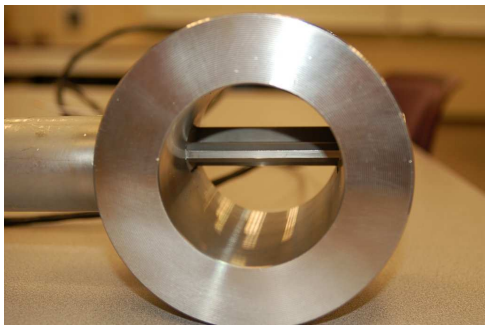
[31] This k factor is empirically determined for each flowmeter by the manufacturer using water as the test fluid (a factory "wet-calibration"), to ensure optimum accuracy.

The following photograph shows a Rosemount model 8800C vortex flow transmitter:



The next two photographs show close-up views of the flowtube assembly, front (left) and rear (right):



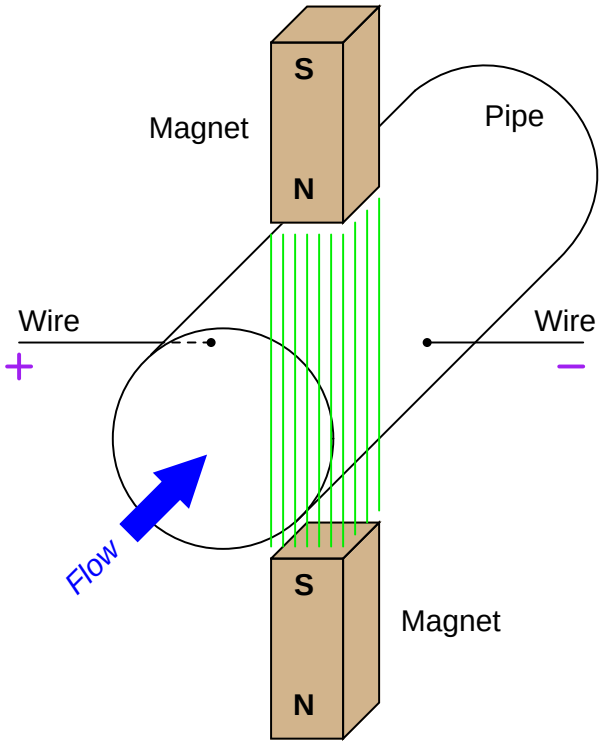


21.4.3 Magnetic flowmeters

When an electrical conductor moves perpendicular to a magnetic field, a voltage is induced in that conductor perpendicular to both the magnetic flux lines and the direction of motion. This phenomenon is known as *electromagnetic induction*, and it is the basic principle upon which all electro-mechanical generators operate.

In a generator mechanism, the conductor in question is typically a coil (or set of coils) made of copper wire. However, there is no reason the conductor must be made of copper wire. *Any* electrically conductive substance in motion is sufficient to electromagnetically induce a voltage, even if that substance is a liquid (or a gas[32]).

Consider water flowing through a pipe, with a magnetic field passing perpendicularly through the pipe:



The direction of liquid flow cuts perpendicularly through the lines of magnetic flux, generating a voltage along an axis perpendicular to both. Metal electrodes opposite each other in the pipe wall intercept this voltage, making it readable to an electronic circuit.

[32] Technically, a gas must be super-heated into a *plasma* state before it is able to conduct electricity.

A voltage induced by the linear motion of a conductor through a magnetic field is called *motional EMF*, the magnitude of which is predicted by the following formula (assuming perfect perpendicularity between the direction of velocity, the orientation of the magnetic flux lines, and the axis of voltage measurement):

$$\mathcal{E} = Blv$$

Where,
$\mathcal{E}$ = Motional EMF (volts)
B = Magnetic flux density (Tesla)
l = Length of conductor passing through the magnetic field (meters)
v = Velocity of conductor (meters per second)

Assuming a fixed magnetic field strength (constant B) and an electrode spacing equal to the fixed diameter of the pipe (constant $l = d$), the only variable capable of influencing the magnitude of induced voltage is velocity (v). In our example, v is not the velocity of a wire segment, but rather the average velocity of the liquid flowstream ($\overline{v}$). Since we see that this voltage will be proportional to average fluid velocity, it must also be proportional to volumetric flow rate, since volumetric flow rate is also proportional to average fluid velocity[33]. Thus, what we have here is a type of flowmeter based on electromagnetic induction. These flowmeters are commonly known as *magnetic flowmeters* or simply *magflow meters*.

We may state the relationship between volumetric flow rate (Q) and motional EMF ($\mathcal{E}$) more precisely by algebraic substitution. First, we will write the formula relating volumetric flow to average velocity, and then manipulate it to solve for average velocity:

$$Q = A\overline{v}$$

$$\frac{Q}{A} = \overline{v}$$

Next, we re-state the motional EMF equation, and then substitute $\frac{Q}{A}$ for $\overline{v}$ to arrive at an equation relating motional EMF to volumetric flow rate (Q), magnetic flux density (B), pipe diameter (d), and pipe area (A):

$$\mathcal{E} = Bd\overline{v}$$

$$\mathcal{E} = Bd\frac{Q}{A}$$

$$\mathcal{E} = \frac{BdQ}{A}$$

[33] This is an application of the transitive property in mathematics: if two quantities are both equal to a common third quantity, they must also be equal to each other. This property applies to proportionalities as well as equalities: if two quantities are proportional to a common third quantity, they must also be proportional to each other.

Since we know this is a circular pipe, we know that area and diameter are directly related to each other by the formula $A = \frac{\pi d^2}{4}$. Thus, we may substitute this definition for area into the last equation, to arrive at a formula with one less variable (only d, instead of both d and A):

$$\mathcal{E} = \frac{BdQ}{\frac{\pi d^2}{4}}$$

$$\mathcal{E} = \frac{BdQ}{1}\frac{4}{\pi d^2}$$

$$\mathcal{E} = \frac{4BQ}{\pi d}$$

If we wish to have a formula defining flow rate Q in terms of motional EMF ($\mathcal{E}$), we may simply manipulate the last equation to solve for Q:

$$Q = \frac{\pi d\mathcal{E}}{4B}$$

This formula will successfully predict flow rate only for absolutely perfect circumstances. In order to compensate for inevitable imperfections, a "proportionality constant" (k) is usually included in the formula[34]:

$$Q = k\frac{\pi d\mathcal{E}}{4B}$$

Where,

Q = Volumetric flow rate (cubic meters per second)
$\mathcal{E}$ = Motional EMF (volts)
B = Magnetic flux density (Tesla)
d = Diameter of flowtube (meters)

Note the linearity of this equation. Nowhere do we encounter a power, root, or other non-linear mathematical function in the equation for a magnetic flowmeter. This means no special characterization is required to calculate volumetric flow rate.

A few conditions must be met for this formula to successfully infer volumetric flow rate from induced voltage:

- The liquid must be a reasonably good conductor of electricity
- Both electrodes must contact the liquid
- The pipe must be completely filled with liquid
- The flowtube must be properly grounded to avoid errors caused by stray electric currents in the liquid

[34] The colloquial term in the United States for this sort of thing is *fudge factor*.

The first condition is met by careful consideration of the process liquid prior to installation. Magnetic flowmeter manufacturers will specify the minimum conductivity value of the liquid to be measured. The second and third conditions are met by correct installation of the magnetic flowtube in the pipe. The installation must be done in such a way as to guarantee full flooding of the flowtube (no gas pockets). The flowtube is usually installed with electrodes across from each other horizontally (never vertically!) so even a momentary gas bubble will not break electrical contact between an electrode tip and the liquid flowstream.

Electrical conductivity of the process liquid must meet a certain minimum value, but that is all. It is surprising to some technicians that changes in liquid conductivity have little to no effect on flow measurement accuracy. It is not as though a doubling of liquid conductivity will result in a doubling of induced voltage! Motional EMF is strictly a function of physical dimensions, magnetic field strength, and fluid velocity. Liquids with poor conductivity simply present a greater electrical resistance in the voltage-measuring circuit, but this is of little consequence because the input impedance of the detection circuitry is phenomenally high. Common fluid types that will *not* work with magnetic flowmeters include deionized water (e.g. steam boiler feedwater, ultrapure water for pharmaceutical and semiconductor manufacturing) and oils.

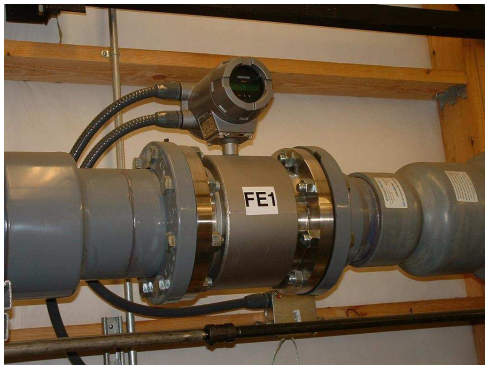
Proper grounding of the flowtube is very important for magnetic flowmeters. The motional EMF generated by most liquid flowstreams is very weak (1 millivolt or less!), and therefore may be easily overshadowed by noise voltage present as a result of stray electric currents in the piping and/or liquid. To combat this problem, magnetic flowmeters are usually equipped to shunt stray electric currents around the flowtube so the only voltage intercepted by the electrodes will be the motional EMF produced by liquid flow. The following photograph shows a Rosemount model 8700 magnetic flowtube, with braided-wire grounding straps clearly visible:

Note how both grounding straps attach to a common junction point on the flowtube housing. This common junction point should also be bonded to a functional earth ground when the flowtube is installed in the process line. On this particular flowtube you can see a stainless steel *grounding ring* on the face of the near flange, connected to one of the braided grounding straps. An identical grounding ring lays on the other flange, but it is not clearly visible in this photograph. These rings provide points of electrical contact with the liquid in installations where the pipe is made of plastic, or where the pipe is metal but lined with a plastic material for corrosion resistance.

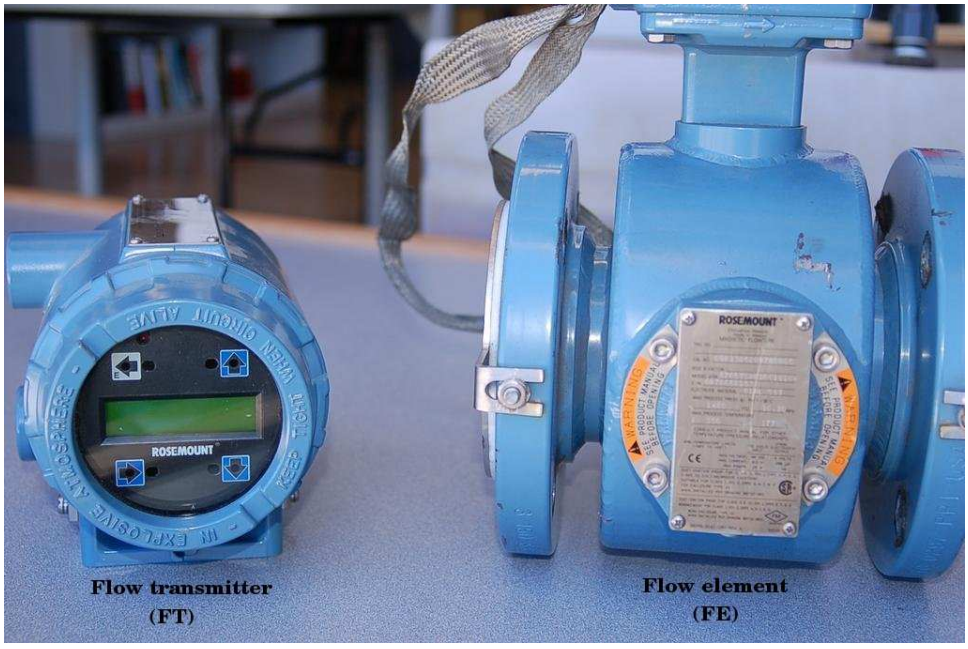
Magnetic flowmeters are fairly tolerant of swirl and other large-scale turbulent fluid behavior.

They do not require the long straight-runs of pipe upstream and downstream that orifice plates do, which is a great advantage in many piping systems.

Some magnetic flowmeters have their signal conditioning electronics located integral to the flowtube assembly. A couple of examples are shown here (a pair of small Endress+Hauser flowmeters on the left and a large Toshiba flowmeter on the right):



Other magnetic flowmeters have separate electronics and flowtube assemblies, connected together by shielded cable. In these installations, the electronics assembly is referred to as the flow transmitter (FT) and the flowtube as the flow element (FE):



This next photograph shows an enormous (36 inch diameter!) magnetic flow element (black) and flow transmitter (blue, behind the person's hand shown for scale) used to measure wastewater flow at a municipal sewage treatment plant:

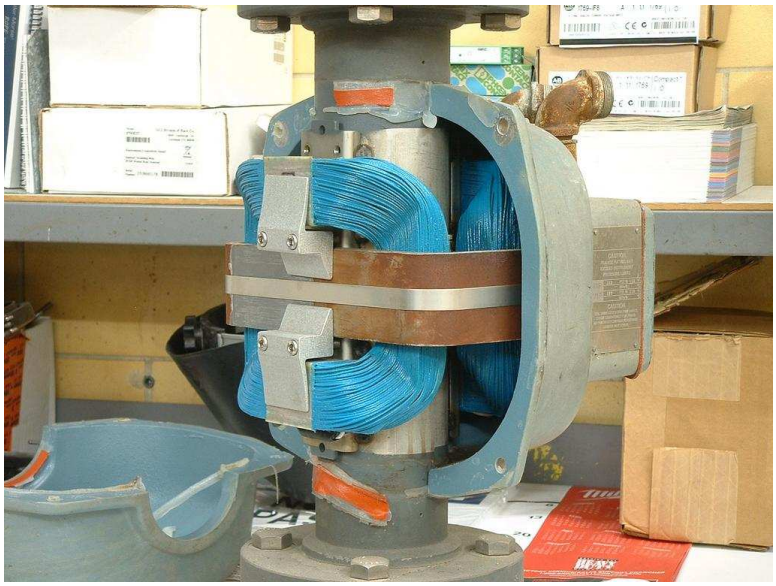
Note the vertical pipe orientation, ensuring constant contact between the electrodes and the water during flowing conditions.

While in theory a permanent magnet should be able to provide the necessary magnetic flux for a magnetic flowmeter to function, this is never done in industrial practice. The reason for this has to do with a phenomenon called *polarization* which occurs when a DC voltage is impressed across a liquid containing ions (electrically charged molecules). Electrically-charged molecules (ions) tend to collect near poles of opposite charge, which in this case would be the flowmeter electrodes. This "polarization" would soon interfere with detection of the motional EMF if a magnetic flowmeter were to use a constant magnetic flux such as that produced by a permanent magnet. A simple solution to this problem is to alternate the polarity of the magnetic field, so the motional EMF polarity also

alternates and never gives the fluid ions enough time to polarize.

This is why magnetic flowmeter tubes always employ electromagnet *coils* to generate the magnetic flux instead of permanent magnets. The electronics package of the flowmeter energizes these coils with currents of alternating polarity, so as to alternate the polarity of the induced voltage across the moving fluid. Permanent magnets, with their unchanging magnetic polarities, would only be able to create an induced voltage with constant polarity, leading to ionic polarization and subsequent flow measurement errors.

A photograph of a Foxboro magnetic flowtube with one of the protective covers removed shows these wire coils clearly (in blue):



Perhaps the simplest form of coil excitation is when the coil is energized by 60 Hz AC power taken from the line power source, such as the case with this Foxboro flowtube. Since motional EMF is proportional to fluid velocity and to the flux density of the magnetic field, the induced voltage for such a coil will be a sine wave whose amplitude varies with volumetric flow rate.

Unfortunately, if there is any stray electric current traveling through the liquid to produce erroneous voltage drops between the electrodes, chances are it will be 60 Hz AC as well[35]. With the coil energized by 60 Hz AC, any such noise voltage may be falsely interpreted as fluid flow because the sensor electronics has no way to distinguish between 60 Hz noise in the fluid and a 60 Hz motional EMF caused by fluid flow.

A more sophisticated solution to this problem uses a *pulsed* excitation power source for the flowtube coils. This is called *DC* excitation by magnetic flowmeter manufacturers, which is a bit misleading because these "DC" excitation signals often reverse polarity, appearing more like an AC square wave on an oscilloscope display. The motional EMF for one of these flowmeters will exhibit the same waveshape, with amplitude once again being the indicator of volumetric flow rate. The sensor electronics can more easily reject any AC noise voltage because the frequency and waveshape of the noise (60 Hz, sinusoidal) will not match that of the flow-induced motional EMF signal.

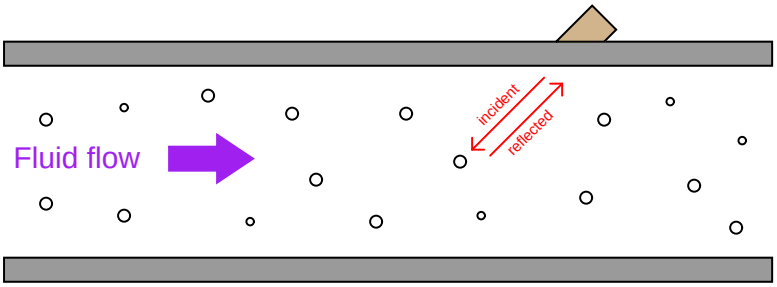
The most significant disadvantage of pulsed-DC magnetic flowmeters is slower response time to changing flow rates. In an effort to achieve a "best-of-both-worlds" result, some magnetic flowmeter manufacturers produce *dual-frequency* flowmeters which energize their flowtube coils with two mixed frequencies: one below 60 Hz and one above 60 Hz. The resulting voltage signal intercepted by the electrodes is demodulated and interpreted as a flow rate.

[35] We know this because the largest electrical noise sources in industry are electric motors, transformers, and other power devices operating on the exact same frequency (60 Hz in the United States, 50 Hz in Europe) as the flowtube coils.

21.4.4 Ultrasonic flowmeters

Ultrasonic flowmeters measure fluid velocity by passing high-frequency sound waves along the fluid flow path. Fluid motion influences the propagation of these sound waves, which may then be measured to infer fluid velocity. Two major sub-types of ultrasonic flowmeters exist: *Doppler* and *transit-time*. Both types of ultrasonic flowmeter work by transmitting a high-frequency sound wave into the fluid stream (the *incident* pulse) and analyzing the received pulse.

Doppler flowmeters exploit the *Doppler effect*, which is the shifting of frequency resulting from waves emitted by or reflected by a moving object. Doppler flowmeters bounce sound waves off of bubbles or particulate material in the flow stream, measure the frequency shift, and infer fluid velocity from the magnitude of that shift.



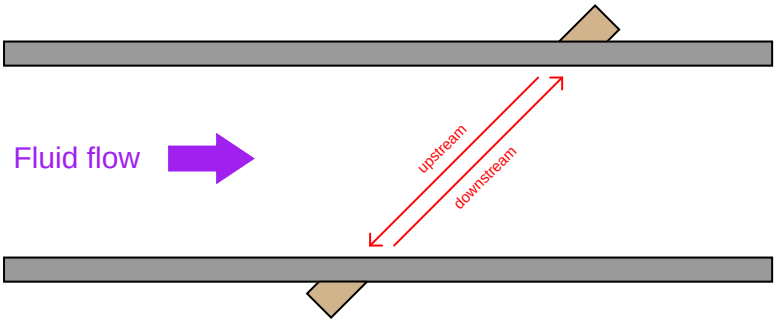
If the reflected wave returns from a bubble advancing toward the ultrasonic transducer[36], the reflected frequency will be greater than the incident frequency. If the flow reverses direction and the reflected wave returns from a bubble traveling away from the transducer, the reflected frequency will be less than the incident frequency.

Note that the Doppler effect yields a direct measurement of fluid *velocity* from each echo received by the transducer. This stands in marked contrast to measurements of *distance* based on time-of-flight (time domain reflectometry – where the amount of *time* between the incident pulse and the returned echo is proportional to distance between the transducer and the reflecting surface). In a Doppler flowmeter, the time delay between the incident and reflected pulses is irrelevant. Only the *frequency shift* between the incident and reflected signals matters.

Doppler-effect ultrasonic flowmeters obviously require flowstream containing bubbles or particulate matter. In many applications this is a normal state of affairs (municipal wastewater, for example). However, some process fluids are simply too clean and too homogeneous to reflect sound waves. In such applications, a different sort of ultrasonic velocity detection technique must be applied.

[36] In the industrial instrumentation world, the word "transducer" usually has a very specific meaning: a device used to process or convert standardized instrumentation signals, such as 4-20 mA converted into 3-15 PSI, etc. In the general scientific world, however, the word "transducer" describes any device converting one form of energy into another. It is this latter definition of the word that I am using when I describe an ultrasonic "transducer" – a device used to convert electrical energy into ultrasonic sound waves, and visa-versa.

Transit-time flowmeters, sometimes called *counterpropagation* flowmeters, use a pair of opposed sensors to measure the time difference between a sound pulse traveling with the fluid flow versus a sound pulse traveling against the fluid flow. Since the motion of fluid tends to carry a sound wave along, the sound pulse transmitted downstream will make the journey faster than a sound pulse transmitted upstream:



The rate of volumetric flow through a transit-time flowmeter is a simple function of the upstream and downstream propagation times:

$$Q = k\frac{t_{up} - t_{down}}{(t_{up})(t_{down})}$$

Where,

Q = Volumetric flow rate

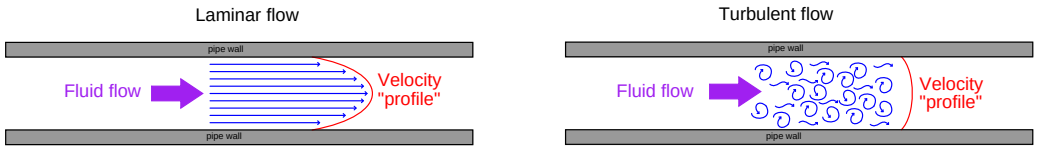
k = Constant of proportionality

t_{up} = Time for sound pulse to travel from downstream location to upstream location (upstream, against the flow)

t_{down} = Time for sound pulse to travel from upstream location to downstream location (downstream, with the flow)

A requirement for reliable operation of a transit-time ultrasonic flowmeter is that the process fluid be free from gas bubbles or solid particles which might scatter or obstruct the sound waves. Note that this is precisely the opposite requirement of Doppler ultrasonic flowmeters, which *require* bubbles or particles to reflect sound waves. These opposing requirements neatly distinguish applications suitable for transit-time flowmeters from applications suitable for Doppler flowmeters.

One potential problem with the transit-time flowmeter is being able to measure the true average fluid velocity when the flow profile changes with Reynolds number. If just one ultrasonic "beam" is used to probe the fluid velocity, the path this beam takes will likely see a different velocity profile as the flow rate changes (and the Reynolds number changes along with it). Recall the difference in fluid velocity profiles between low Reynolds number flows (left) and high Reynolds number flows (right):



A popular way to mitigate this problem is to use multiple sensor pairs, sending acoustic signals along multiple paths through the fluid (i.e. a *multipath* ultrasonic flowmeter), and to average the resulting velocity measurements. Dual-beam flowmeters have been in use for well over a decade, and one manufacturer even has a *five beam* ultrasonic flowmeter model which they claim maintains an accuracy of +/- 0.15% through the laminar-to-turbulent flow regime transition[37].

Some modern ultrasonic flowmeters have the ability to switch back and forth between Doppler and transit-time (counterpropagation) modes, automatically adapting to the fluid being sensed. This capability enhances the suitability of ultrasonic flowmeters to a wider range of process applications.

Ultrasonic flowmeters are adversely affected by swirl and other large-scale fluid disturbances, and as such may require substantial lengths of straight pipe upstream and downstream of the measurement flowtube to stabilize the flow profile.

Advances in ultrasonic flow measurement technology have reached a point where it is now feasible to consider ultrasonic flowmeters for custody transfer measurement of natural gas. The American Gas Association has released a report specifying the use of multipath ultrasonic flowmeters in this capacity (Report #9). Just like the AGA's #3 (orifice plate) and #7 (turbine) high-accuracy gas flow measurement standards, the AGA9 standard requires the addition of pressure and temperature instruments on the gas line to compensate for changes in gas pressure and temperature, so that a flow computer may calculate either true mass flow or volumetric flow in standardized units (e.g. SCFM).

A unique advantage of ultrasonic flow measurement is the ability to measure flow through the use of temporary *clamp-on* sensors rather than a specialized flowtube with built-in ultrasonic transducers. While clamp-on sensors are not without their share of problems[38], they constitute an excellent solution for certain flow measurement applications.

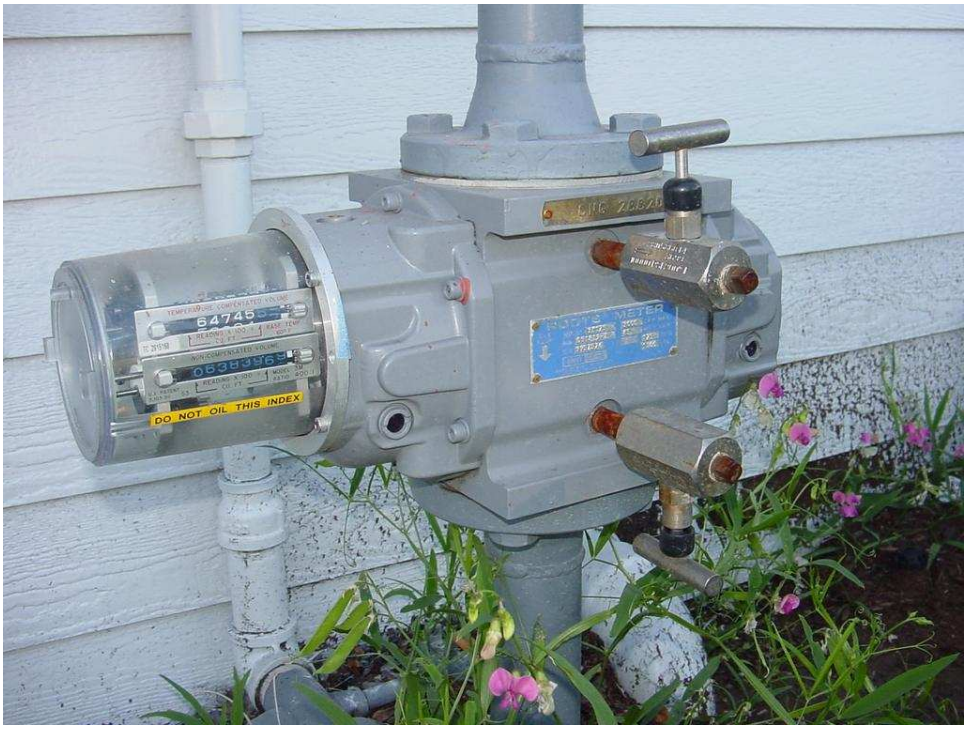
[37] See page 10 of Friedrich Hofmann's *Fundamentals of Ultrasonic Flow Measurement for industrial applications* paper.

[38] Most notably, the problem of achieving good acoustic coupling with the pipe wall so signal transmission to the fluid and signal reception back to the sensor may be optimized. Also, there is the potential for sound waves to "ring around the pipe" instead of travel through the fluid with clamp-on ultrasonic flowmeters because the sound waves must travel through the full thickness of the pipe walls in order to enter and exit the fluid stream.

21.5 Positive displacement flowmeters

A *positive displacement* flowmeter is a cyclic mechanism built to pass a fixed volume of fluid through with every cycle. Every cycle of the meter's mechanism *displaces* a precisely defined ("positive") quantity of fluid, so that a count of the number of mechanism cycles yields a precise quantity for the total fluid volume passed through the flowmeter. Many positive displacement flowmeters are rotary in nature, meaning each shaft revolution represents a certain volume of fluid has passed through the meter. Some positive displacement flowmeters use pistons, bellows, or expandable bags working on an alternating fill/dump cycle to measure off fluid quantities.

Positive displacement flowmeters have been the traditional choice for residential and commercial natural gas flow and water flow measurement in the United States (a simple application of *custody transfer* flow measurement, where the fluid being measured is a commodity bought and sold). The cyclic nature of a positive displacement meter lends itself well to total gas quantity measurement (and not just flow *rate*), as the mechanism may be coupled to a mechanical counter which is read by utility personnel on a monthly basis. A rotary gas flowmeter is shown in the following photograph. Note the odometer-style numerical display on the left-hand end of the meter, totalizing gas usage over time:



Positive displacement flowmeters rely on moving parts to shuttle quantities of fluid through them, and these moving parts must effectively seal against each other to prevent leakage past the mechanism (which will result in the instrument indicating less fluid passing through than there actually is). In fact, the defining characteristic of any positive displacement device is that fluid *cannot* move through without actuating the mechanism, and that the mechanism *cannot* move

without fluid passing through. This stands in contrast to machines such as centrifugal pumps and turbines, where it is possible for the moving part (the impeller or turbine wheel) to jam in place and still have fluid pass through the mechanism. If a positive displacement mechanism jams, fluid flow absolutely halts.

The finely-machined construction of a positive displacement flowmeter will suffer damage from grit or other abrasive materials present in the fluid, which means these flowmeters are applicable only to clean fluid flowstreams. Even with clean fluid flowing through, the sealing surfaces of the mechanisms are subject to wear and accumulating inaccuracies over time. However, there is really nothing more definitive for measuring volumetric flow rate than an instrument built to measure individual volumes of fluid with each mechanical cycle. As one might guess, these instruments are completely immune to swirl and other large-scale fluid turbulence, and may be installed nearly anywhere in a piping system (no need for long sections of straight-length pipe upstream or downstream). Positive displacement flowmeters are also very linear, since mechanism cycles are directly proportional to fluid volume.

A large positive displacement flowmeter used to measure the flow of liquid (registering total accumulated volume in units of gallons) is shown here, having been cut away for use as an instructional display:

The left-hand photograph shows the gear mechanism used to convert rotor motion into a visible total readout. The right-hand photograph shows a close-up of the interlocking rotors (one with three lobes, the other with four slots which those lobes mesh with). Both the lobes and slots are spiral-shaped, such that fluid passing along the spiral pathways must "push" the lobes out of the slots and cause the rotors to rotate. So long as there is no leakage between rotor lobes and slots, rotor turns will have a precise relationship to fluid volume passed through the flowmeter.

21.6 Standardized volumetric flow

The majority of flowmeter technologies operate on the principle of interpreting fluid flow based on the *velocity* of the fluid. Magnetic, ultrasonic, turbine, and vortex flowmeters are prime examples, where the sensing elements (of each meter type) respond directly to fluid velocity. Translating fluid velocity into volumetric flow is quite simple, following this equation:

$$Q = A\overline{v}$$

Where,

Q = Volumetric flow rate (e.g. cubic feet per minute)

A = Cross-sectional area of flowmeter throat (e.g. square feet)

$\overline{v}$ = Average fluid velocity at throat section (e.g. feet per minute)

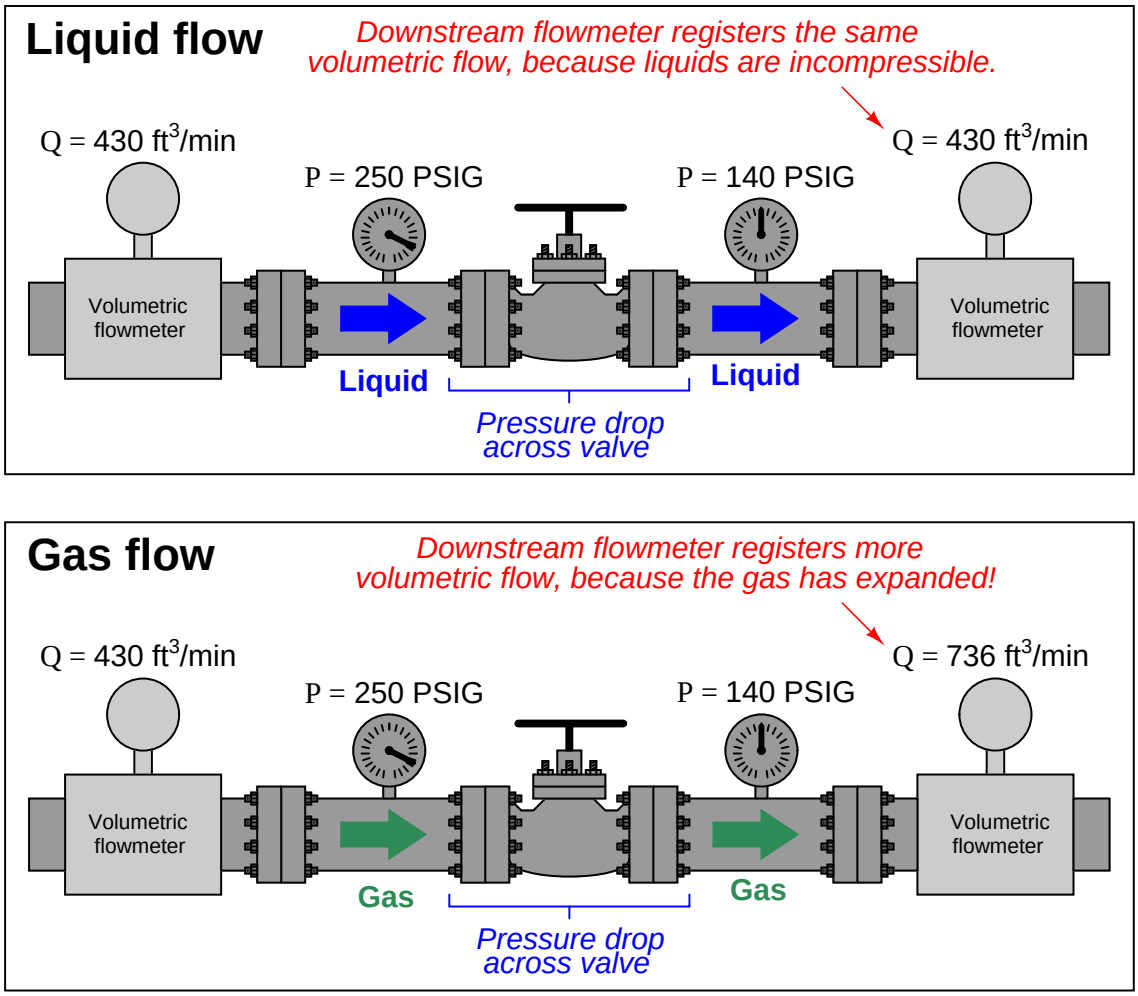
Positive displacement flowmeters are even more direct than velocity-sensing flowmeters. A positive displacement flowmeter directly measures volumetric flow, counting discrete volumes of fluid as it passes through the meter.

Even pressure-based flowmeters such as orifice plates and venturi tubes are usually calibrated to measure in units of volume over time (e.g. gallons per minute, barrels per hour, cubic feet per second, etc.). For a great many industrial fluid flow applications, measurement in volumetric units makes sense.

This is especially true if the fluid in question is a liquid. Liquids are essentially incompressible: that is, they do not easily yield in volume to applied pressure. This makes volumetric flow measurement relatively simple for liquids: one cubic foot of a liquid at high pressure and temperature inside a process vessel will occupy approximately the same volume (≈ 1 ft^3) when stored in a barrel at ambient pressure and temperature.

Gases and vapors, however, easily change volume under the influences of pressure and temperature. In other words, a gas will yield to an increasing pressure by decreasing in volume as the gas molecules are forced closer together, and it will yield to a decreasing temperature by decreasing in volume as the kinetic energy of the individual molecules is reduced. This makes volumetric flow measurement more complex for gases than for liquids. One cubic foot of gas at high pressure and temperature inside a process vessel will *not* occupy one cubic foot under different pressure and temperature conditions.

The practical difference between volumetric flow measurement for liquids versus gases is easily seen through an example where we measure the volumetric flow rate before and after a pressure-reduction valve:

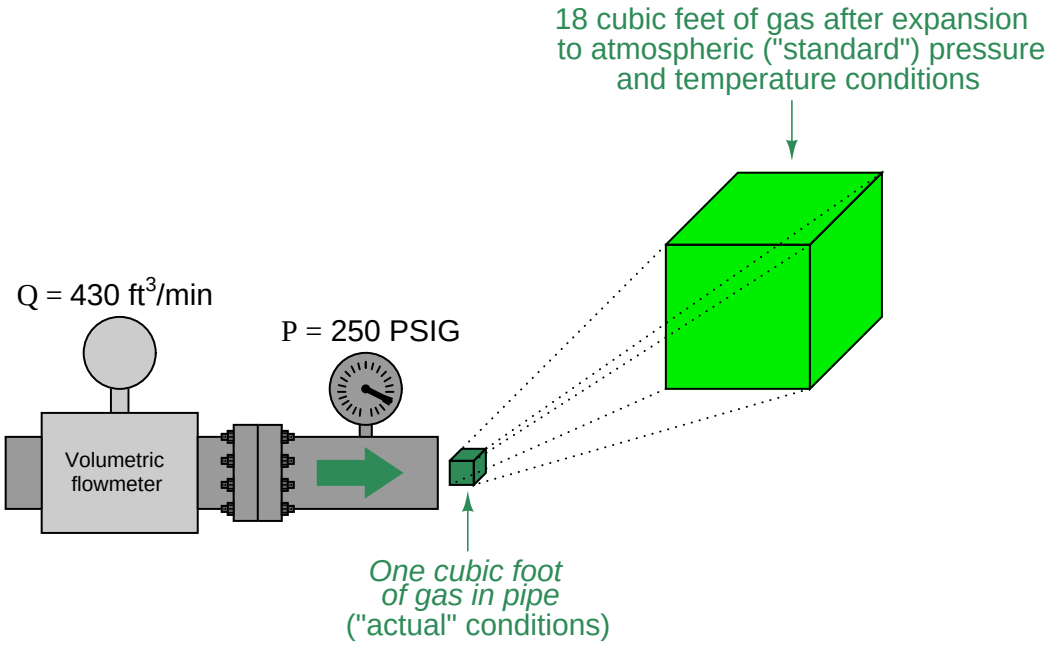


The volumetric flow rate of liquid before and after the pressure-reducing valve is the same, since the volume of a liquid does not depend on the pressure applied to it (i.e. liquids are *incompressible*). The volumetric flow rate of the gas, however, is significantly greater following the pressure-reducing valve than before, since the reduction in pressure allows the gas to expand (i.e. less pressure means the gas occupies a greater volume).

What this tells us is that volumetric flow measurement for gas is virtually meaningless without accompanying data on pressure and temperature. A flow rate of "430 ft^3/min" reported by a flowmeter measuring gas at 250 PSIG means something completely different than the same volumetric flow rate (430 ft^3/min) reported at a different line pressure.

One solution to this problem is to dispense with volumetric flow measurement altogether in favor of *mass flow measurement*, constructing the flowmeter in such a way that the actual *mass* of the gas molecules is measured as they pass through the instrument. This approach is explored in more detail in section 21.7, beginning on page 1097. A more traditional approach to this problem is to specify gas flow in volume units per time, at some agreed-upon (standardized) set of pressure and temperature conditions. This is known as *standardized* volumetric flow measurement.

Referring to our pictorial example previously shown, imagine if we took a sample of the gas flowing at a line pressure of 250 PSIG and let that sample expand to atmospheric pressure (0 PSIG) and ambient temperature (60 degrees Fahrenheit), measuring its new volume under those new conditions. Obviously, one cubic foot of gas at 250 PSIG would expand to a far greater volume than 1 ft^3 at atmospheric pressure. This ratio of "standard volume" to "actual volume" (in the pressurized pipe) could then be used to *scale* the flowmeter's measurement, so that the flowmeter registers in *standard cubic feet per minute*, or *SCFM*:



Assuming the same temperature inside the pipe as outside, the expansion ratio for this gas will be about 18:1, meaning the "actual" flowing rate of 430 ft^3/min is equivalent to approximately 7740 ft^3/min of gas flow at "standard" atmospheric conditions. In order to unambiguously distinguish "actual" volumetric flow rate from "standardized" volumetric flow rate, we commonly preface each unit with a letter "A" or letter "S" (e.g. ACFM and SCFM).

If a volumetric-registering flowmeter is equipped with pressure and temperature sensors, it may automatically scale its own output signal to measure gas flow rate in standard volumetric units. All we need to determine is the mathematical procedure to scale actual conditions to standard conditions for a gas.

To do this, we will refer to the Ideal Gas Law ($PV = nRT$), which is a fair approximation for most real gases at conditions far from their critical phase-change points. First, we shall write two versions of the Ideal Gas Law, one for the gas under "standard" atmospheric conditions, and one for the gas under "actual" flowing conditions (using the subscripts "S" and "A" to distinguish one equation from the other):

$$P_S V_S = nRT_S$$

$$P_A V_A = nRT_A$$

What we need to do is determine the *ratio* of standard volume to actual volume ($\frac{V_S}{V_A}$). To do this, we may divide one equation by the other[39]:

$$\frac{P_S V_S}{P_A V_A} = \frac{nRT_S}{nRT_A}$$

Seeing that both the n variables are identical and R is a constant, we may cancel them both from the right-hand side of the equation:

$$\frac{P_S V_S}{P_A V_A} = \frac{T_S}{T_A}$$

Solving for the ratio $\frac{V_S}{V_A}$:

$$\frac{V_S}{V_A} = \frac{P_A T_S}{P_S T_A}$$

Since we know the definition of volumetric flow (Q) is volume over time ($\frac{V}{t}$), we may divide each V variable by t to convert this into a volumetric flow *rate* correction ratio[40]:

$$\frac{\frac{V_S}{t}}{\frac{V_A}{t}} = \frac{P_A T_S}{P_S T_A}$$

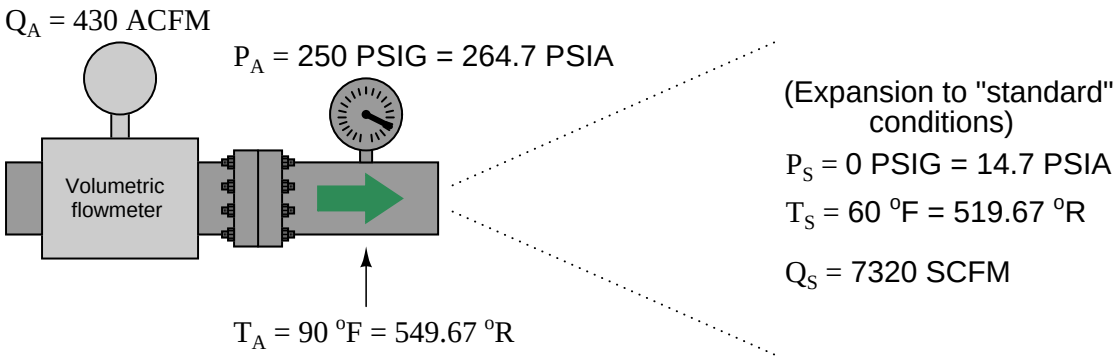
This leaves us with a ratio of "standardized" volumetric flow (Q_S) to "actual" volumetric flow (Q_A), for any known pressures and temperatures, standard to actual:

$$\frac{Q_S}{Q_A} = \frac{P_A T_S}{P_S T_A}$$

[39] Recall from algebra that we may perform any arithmetic operation we wish to any equation, so long as we apply that operation equally to both sides of the equation. Dividing one equation by another equation obeys this principle, because both sides of the second equation are equal. In other words, we could divide both sides of the first equation by $P_A V_A$ (although that would not give us the solution we are looking for), but dividing the left side by $P_A V_A$ and the right side by nRT_A is really doing the same thing, since nRT_A is identical in value to $P_A V_A$.

[40] Division by t does not alter the equation at all, since we are essentially multiplying the left-hand side by $\frac{t}{t}$ which is multiplication by 1. This is why we did not have to apply t to the right-hand side of the equation.

We may apply this to a practical example, assuming flowing conditions of 250 PSIG and 90 degrees Fahrenheit, and "standard" conditions of 0 PSIG and 60 degrees Fahrenheit. It is very important to ensure all values for pressure and temperature are expressed in *absolute* units (e.g. PSIA and degrees Rankine), which is what the Ideal Gas Law assumes:



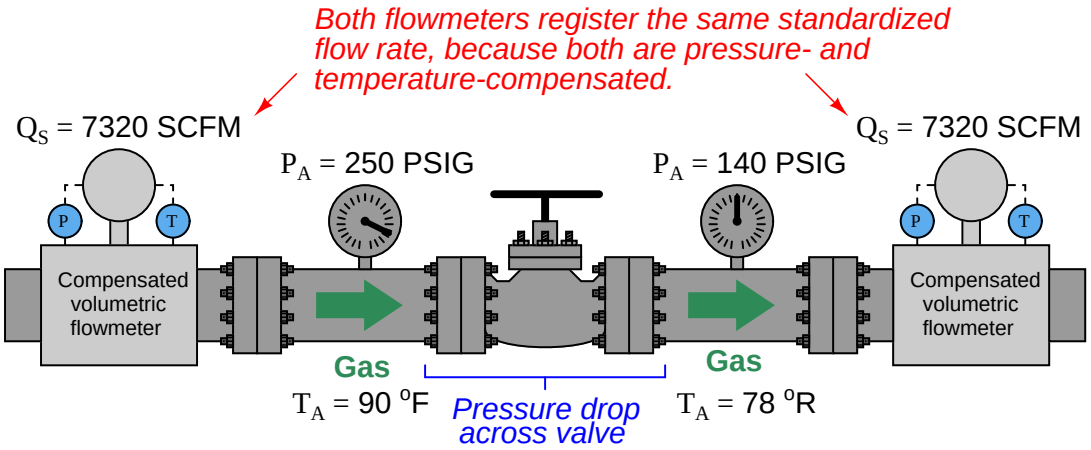
$$\frac{Q_S}{Q_A} = \frac{P_A T_S}{P_S T_A}$$

$$Q_S = Q_A \left(\frac{P_A T_S}{P_S T_A} \right)$$

$$7320 \text{ SCFM} = (430 \text{ ACFM}) \left(\frac{(264.7 \text{ PSIA})(519.67^o\text{R})}{(14.7 \text{ PSIA})(549.67^o\text{R})} \right)$$

This figure of 7320 SCFM indicates the volumetric flow rate of the gas through the pipe *had it been allowed to expand to atmospheric pressure and cool to ambient temperature*. Although we know these are definitely not the same conditions inside the gas pipe, the correction of actual volumetric flow measurement to these imagined conditions allows us to express gas flow rates in an equitable fashion regardless of the process line pressure or temperature. To phrase this in colloquial terms, standardized volumetric flow figures allow us to compare different process gas flow rates on an "apples to apples" basis, instead of the "apples to oranges" problem we faced earlier where flowmeters would register different volumetric flow values because their line pressures and/or flowing temperatures differed.

With pressure and temperature compensation integrated into volumetric flowmeters, we should be able to measure the exact same *standardized* flow rate at any point in a series gas piping system regardless of pressure or temperature changes:



An unfortunate state of affairs is the existence of multiple "standard" conditions of pressure and temperature defined by different organizations. In the previous example, 14.7 PSIA was assumed to be the "standard" atmospheric pressure, and 60 degrees Fahrenheit (519.67 degrees Rankine) was assumed to be the "standard" ambient temperature. This conforms to the API (American Petroleum Institute) standards in the United States of America, but it does *not* conform to other standards in America or in Europe. The ASME (American Society of Mechanical Engineers) uses 14.7 PSIA and 68 degrees Fahrenheit (527.67 degrees Rankine) as their "standard" conditions for calculating SCFM. In Europe, the PNEUROP agency has standardized with the American CAGI (Compressed Air and Gas Institute) organization on 14.5 PSIA and 68 degrees Fahrenheit being the "standard" conditions.

For gas flows containing condensible vapors, the *partial pressures* of the vapors must be subtracted from the absolute pressures (P_A and P_S) in order that the correction factor accurately reflects gas behavior alone. A common application of standardized gas flow involving partial pressure correction is found in compressed air systems, where water vapor (relative humidity) is a factor. Here too, standards differ as to the humidity conditions of "standard" cubic feet. The API and PNEUROP/CAGI standards call for 0% humidity (perfectly dry air) as the "standard," while the ASME defines 35% relative humidity as "standard" for compressed air calculations.

21.7 True mass flowmeters

Many traditional flowmeter technologies respond to the *volumetric flow rate* of the moving fluid. Velocity-based flowmeters such as magnetic, vortex, turbine, and ultrasonic generate output signals proportional to fluid *velocity* and nothing else. This means that if the fluid flowing through one of these flowmeter types were to suddenly become denser (while still flowing by at the same number of volumetric units per minute), the flowmeter's response would not change at all.

The information provided by a volumetric flowmeter may not be what is actually best for the process being measured, however. If the flowmeter in question happens to be measuring the flow rate of feed into a chemical reactor vessel, for example, what we're really concerned with is *how many molecules per unit time* of feed is entering that reactor, not how many cubic meters or how many gallons. We know that changes in temperature will cause gases and liquids alike to change density, which means each volumetric unit will contain a different number of molecules after a temperature change than before. Pressure has a similar influence on gases: increased pressure means more gas molecules occupying each cubic foot (or other volumetric unit), all other factors being equal. If a process requires an accounting of molecular flow rate, a volumetric flowmeter will not provide relevant information.

In steam boiler control systems, the flow rate of water into the boiler and the flow rate of steam coming out of the boiler must be matched in order to maintain a constant quantity of water within the boiler tubes and drums. However, water is a liquid and steam is a vapor, so flow measurements based on volume are meaningless. The only reasonable way for the control system to balance both flow rates is to measure them as *mass* flows rather than volumetric flows. No matter what form (phase) the H_2O molecules take, every kilogram going into the boiler must be matched by a kilogram coming out of the boiler in accordance with the Law of Mass Conservation: every H_2O molecule entering the boiler must be matched by one H_2O molecule exiting the boiler in order to maintain an unchanging quantity of H_2O molecules within the boiler. This is why boiler feedwater and steam flowmeters alike are typically calibrated to measure in units of lbm (pounds mass) per unit time.

A similar problem arises in instances where the flowmeter is used for *custody transfer*. This term denotes scenarios where a particular material is being bought and sold, and where accuracy of flow measurement is a matter of monetary importance. Again, in such instances, it is usually the *number of molecules* being bought and sold that really matters, not how many cubic meters or gallons those molecules occupy[41]. Here, as with the chemical reactor feed flow application, a volumetric flowmeter does not provide the most relevant information.

We know from the study of chemistry that all elements have fixed mass values: one *mole*[42] of any element in monatomic form (single, unbound atoms) will have a mass equal to the atomic mass of that element. For example, one mole of carbon (C) atoms has a mass of 12 grams because the element carbon has an atomic mass of 12. Similarly, one mole of oxygen (O) atoms is guaranteed to have a mass of 16 grams[43] because 16 is the atomic mass for the element oxygen. Consequently, one mole of carbon monoxide (CO) molecules will have a mass of 28 grams (12 + 16), and one mole

[41] In some applications, such as the custody transfer of natural gas, we are interested in something even more abstract: *heating value*. However, in order to calculate the gross heating value of a fuel gas stream, we must begin with an accurate mass flow measurement – volumetric flow is not really helpful.

[42] A "mole" is equal to a value of 6.022×10^{23} entities. Therefore, one mole of carbon atoms is 602,200,000,000,000,000,000,000 carbon atoms. For a more detailed examination of this subject, refer to section 3.7 beginning on page 162.

[43] I am purposely ignoring the fact that naturally occurring carbon has an average atomic mass of 12.011, and naturally occurring oxygen has an atomic mass of 15.9994.

of carbon dioxide (CO_2) molecules will have a mass of 44 grams (12 + 16×2). The relationship between molecule count and mass for any given chemical compound is fixed, because mass is an intrinsic property of matter.

If our desire is to account for the number of molecules passed through a pipe, and we happen to know the chemical composition of those molecules, measuring the *mass* of the fluid passing through is the most practical way to do it.

The mathematical relationship between volumetric flow (Q) and mass flow (W) is one of proportionality with mass density (ρ):

$$W = \rho Q$$

Dimensional analysis confirms this relationship. Volumetric flow is always measured in volume units (m^3, ft^3, cc, in^3, gallons, etc.) over time, whereas mass flow is always measured in mass units (g, kg, lbm[44], or slugs) over time. To use a specific example, a mass flow rate in pounds (mass) per minute will be obtained by multiplying a mass density in pounds per cubic foot by a volumetric flow rate in cubic feet per minute:

$$\left[\frac{\text{lbm}}{\text{min}}\right] = \left[\frac{\text{lbm}}{\text{ft}^3}\right] \left[\frac{\text{ft}^3}{\text{min}}\right]$$

With modern sensing and computational technology, it is possible to combine pressure, temperature, and volumetric flow measurements in such a way to *derive* a measurement of mass flow. This is precisely the goal with AGA3 flow measurement (orifice plates), AGA7 flow measurements (turbines), and AGA9 flow measurement (ultrasonic): "compensating" the fundamentally volumetric nature[45] of these flow-measuring elements with pressure and temperature data to calculate the flow rate in mass units over time.

However, compensated flowmeter systems require much more calibration effort to maintain their long-term accuracy, not to mention a significant capital investment in the multiple transmitters and flow computer required to gather all the necessary data and perform the mass flow calculations. It would be much simpler if there existed flowmeter technologies that naturally responded to the mass flow rate of a fluid! Fortunately, such flowmeter technologies *do* indeed exist, which is the subject of this section.

[44] The British unit of the "pound" is technically a measure of *force* or *weight* and not *mass*. The proper unit of mass measurement in the British system is the "slug." However, for better or worse, the "slug" is rarely used, and so engineers have gotten into the habit of using "pound" as a mass measurement. In order to distinguish the use of "pound" to represent mass (an intrinsic property of matter) as opposed to the use of "pound" to represent weight (an incidental property of matter), the former is abbreviated *lbm* (literally, "pounds mass"). In Earth gravity, "lbm" and "lb" are synonymous. However, the standard Newtonian equation relating force, mass, and acceleration ($F = ma$) does not work when "lbm" is the unit used for mass and "lb" is used for force (it does when "slug" is used for mass and "lb" is used for force, though!). A weird unit of force invented to legitimize "pound" as an expression of mass is the *poundal* ("pdl"): one "poundal" of force is the reaction of one "pound" of mass (lbm) accelerated one foot per second squared. By this definition, a one-pound mass (1 lbm) in Earth gravity weighs 32 poundals!

[45] One could argue that orifice plates and other pressure-based flowmeters respond primarily to mass flow rather than volumetric flow, since their operation is based on the pressure created by *accelerating a mass*. However, fluid density does affect the relationship between mass flow rate and differential pressure (note how the density term ρ appears in the mass flow equation $W = k\sqrt{\rho(P_1 - P_2)}$, where it would not if differential pressure were a strict function of mass flow rate and nothing else), and so the raw output of these instruments must still be "compensated" by pressure and temperature measurements.

For each of the following mass flowmeter technologies, it should be clearly understood that the instrument in question *naturally* responds to mass flow rate. To use our hypothetical example of a fluid stream whose density suddenly increases while the volumetric rate remains constant, a true mass flowmeter will immediately recognize the increase in mass flow (same volume rate, but more mass per unit volume) without the need for additional compensating measurements or computer calculations. True mass flowmeters operate on principles directly related to the mass of the fluid molecules passing through the meter, making them fundamentally different from other flowmeter types.

In the case of the Coriolis flowmeter, the instrument works on the principle of *inertia*: the force generated by an object when it is accelerated or decelerated. This basic property of mass (opposition to change in velocity) forms the basis of the Coriolis flowmeter's function. The inertial force generated inside a Coriolis flowmeter will thus double if the volumetric flowrate of a constant-mass fluid doubles; the inertial force will likewise double if the density of a constant volumetric flow of fluid doubles. Either way, the inertial force becomes a representation of *how fast mass is moving through the flowmeter*, and so the Coriolis flowmeter is a true mass flow instrument.

In the case of the thermal flowmeter, the instrument works on the principle of *convective heat transfer*: heat energy extracted from a hot object as cooler molecules pass by. The ability for fluid molecules to transport heat is a function of the *specific heat* of each molecule and the number of molecules moving past the warmer object. So long as the chemical composition of the fluid remains unchanged, the convective transfer of heat is a function of how many fluid molecules pass by in a given time. The heat transfer rate inside a thermal flowmeter will thus double if the volumetric flowrate of a given fluid doubles; the heat transfer rate will likewise double if the density of a given fluid doubles (i.e. twice the number of molecules passing by with each time interval). Either way, the convective heat transfer rate becomes a representation of *how many molecules of fluid are moving through the flowmeter*, which for any given fluid type is proportional to the fluid's mass flow rate. This makes the thermal flowmeter a true mass flow instrument for any (calibrated) fluid composition.

Some older, mechanical technologies[46] exist for measuring true mass flow, but these are being supplanted by Coriolis and thermal mass flowmeter technologies. Coriolis and thermal mass flowmeters are also fast becoming the technology of choice for applications formerly the domain of compensated orifice plate (e.g. AGA3) and turbine (e.g. AGA7) flowmeters.

[46] The impeller-turbine and twin-turbine mass flowmeter types are examples of mechanical true-mass flow technologies. Both work on the principle of fluid inertia. In the case of the impeller-turbine flowmeter, an impeller driven by a constant-speed electric motor imparts a "spin" to a moving fluid, which then impinges on a stationary turbine wheel to generate a measurable torque. The greater the mass flow rate, the greater the impulse force imparted to the turbine wheel. In the twin-turbine mass flowmeter, two rotating turbine wheels with different blade pitches are coupled together by a flexible coupling. As each turbine wheel attempts to spin at its own speed, the inertia of the fluid causes a differential torque to develop between the two wheels. The more mass flow rate, the greater the angular displacement (offset) between the two wheels.

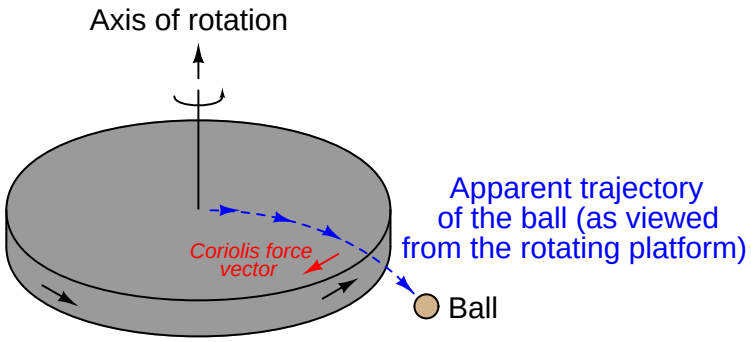
21.7.1 Coriolis flowmeters

In physics, certain types of forces are classified as *fictitious* or *pseudoforces* because they only appear to exist when viewed from an accelerating perspective (called a *non-inertial reference frame*). The feeling you get in your stomach when you accelerate either up or down in an elevator, or when riding a roller-coaster at an amusement park, feels like a force acting against your body when it is really nothing more than the reaction of your body's inertia to being accelerated by the vehicle you are in. The real force is the force of the vehicle against your body, causing it to accelerate. What you perceive is merely a reaction to that force, and not the primary cause of your discomfort as it might appear to be.

Centrifugal force is another example of a "pseudoforce" because although it may appear to be a real force acting on any rotating object, it is in fact nothing more than an inertial reaction. Centrifugal force is a common experience to any child who has ever played on a "merry-go-round:" that perception of a force drawing you away from the center of rotation, toward the rim. The real force acting on any rotating object is toward the center of rotation (a *centripetal* force) which is necessary to make the object radially accelerate toward a center point rather than travel in a straight line as it normally would without any forces acting upon it. When viewed from the perspective of the spinning object, however, it would seem there is a force drawing the object away from the center (a *centrifugal* force).

Yet another example of a "pseudoforce" is the *Coriolis force*, more complicated than centrifugal force, arising from motion perpendicular to the axis of rotation in a non-inertial reference frame. The example of a merry-go-round works to illustrate Coriolis force as well: imagine sitting at the center of a spinning merry-go-round, holding a ball. If you gently toss the ball away from you and watch the trajectory of the ball, you will notice it curve rather than travel away in a straight line. In reality, the ball *is* traveling in a straight line (as viewed from an observer standing on the ground), but from your perspective on the merry-go-round, it appears to be deflected by an invisible force which we call the Coriolis force.

In order to generate a Coriolis force, we must have a mass moving at a velocity perpendicular to an axis of rotation:



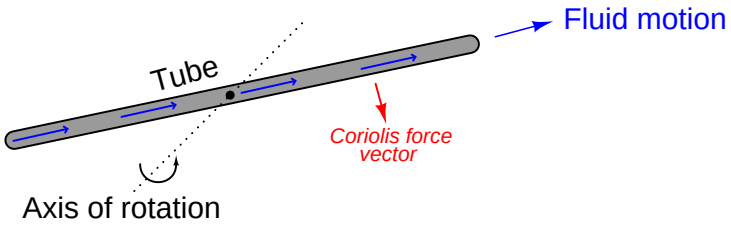
The magnitude of this force is predicted by the following vector equation[47]:

$$\vec{F}_c = -2\vec{\omega} \times \vec{v}' m$$

Where,

$\vec{F}_c$ = Coriolis force vector
$\vec{\omega}$ = Angular velocity (rotation) vector
$\vec{v}'$ = Velocity vector as viewed from the rotating reference frame
m = Mass of the object

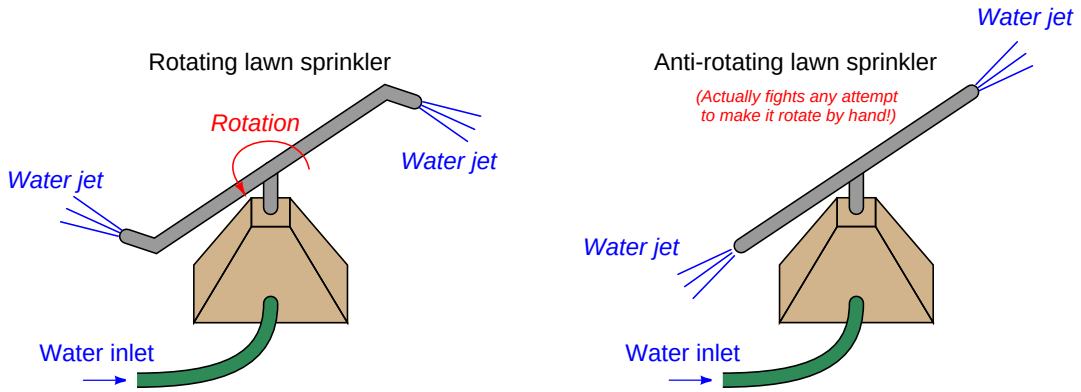
If we replace the ball with a fluid moving through a tube, and we introduce a rotation vector by tilting that tube around a stationary axis (a fulcrum), a Coriolis force develops on the tube in such a way as to oppose the direction of rotation just like the Coriolis force opposed the direction of rotation of the rotating platform in the previous illustration:



To phrase this in anthropomorphic terms, the fluid "fights" against this rotation because it "wants" to keep traveling in a straight line. For any given rotational velocity, the amount of "fight" will be directly proportional to the product of fluid velocity and fluid mass. In other words, the magnitude of the Coriolis force will be in direct proportion to the fluid's mass flow rate. This is the basis of a *Coriolis mass flowmeter*.

[47] This is an example of a vector *cross-product* where all three vectors are perpendicular to each other, and the directions follow the right-hand rule.

A demonstration of this Coriolis force may be made by modifying the nozzles on a rotary lawn sprinkler so they point straight out from the center rather than angle in one direction. As water squirts through the now-straight nozzles, they no longer generate a rotational reaction force to spin the nozzle assembly, and so the nozzles remain in place (this much should be obvious). However, if someone were to try rotating the nozzle assembly by hand, they would discover the Coriolis force *opposes* the rotation, acting to keep the nozzle assembly from rotating. The greater the mass flow rate of water through the nozzles, the stronger the inhibiting Coriolis force. Instead of a rotating lawn sprinkler, you are now the proud owner of an *anti*-rotating lawn sprinkler that actually *fights* any attempt to rotate it:

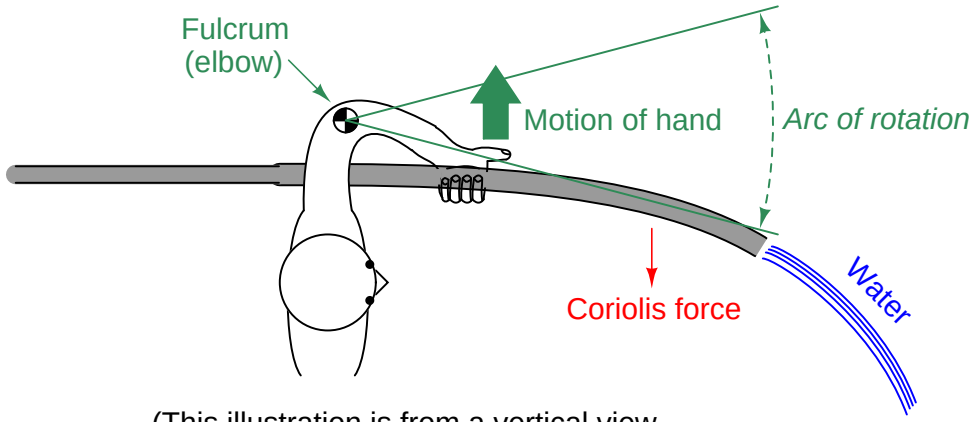


This is a very non-intuitive concept, so it deserves further explanation. The "anti-rotating" sprinkler doesn't just fail to rotate on its own – it actually *opposes* any attempt to rotate from an external force (e.g. a person trying to push the tubes by hand).

This opposition would not occur if the tubes were merely capped off at the ends and filled with stagnant water. If this were the case, the tubes would simply be heavy with the water's weight, and they would rotate freely about the axis just like any pair of heavy metal tubes would (whether hollow and filled with water, or solid metal). The tubes would have inertia, but they would not *actively* oppose any external effort to rotate.

Having liquid water *move* through the tubes is what makes the difference, and the reason becomes clear once we imagine what each water molecule experiences as it flows from the center (axis of rotation) to the nozzle at the tube tip. Each water molecule originating from the center begins with no lateral velocity, but must *accelerate* as it travels further along the tube toward the circumference of the tips' rotation where the lateral velocity is at a maximum. The fact that new water molecules are continually making this journey from center to tip means there will always be a *new* set of water molecules in need of acceleration from center velocity (zero) to tip velocity (maximum). In capped tubes filled with stagnant water, the acceleration would only occur in getting the tubes' rotation up to speed – once there, the lateral velocity of each water molecule sitting stagnant inside the tubes would remain the same. However, with water *flowing* from center to tip, this process of acceleration from zero velocity to tip velocity must occur over and over again (continually) for each new water molecule flowing through. This continual acceleration of *new mass* is what generates the Coriolis force, and what actively opposes any external force trying to rotate the "anti-rotating" sprinkler.

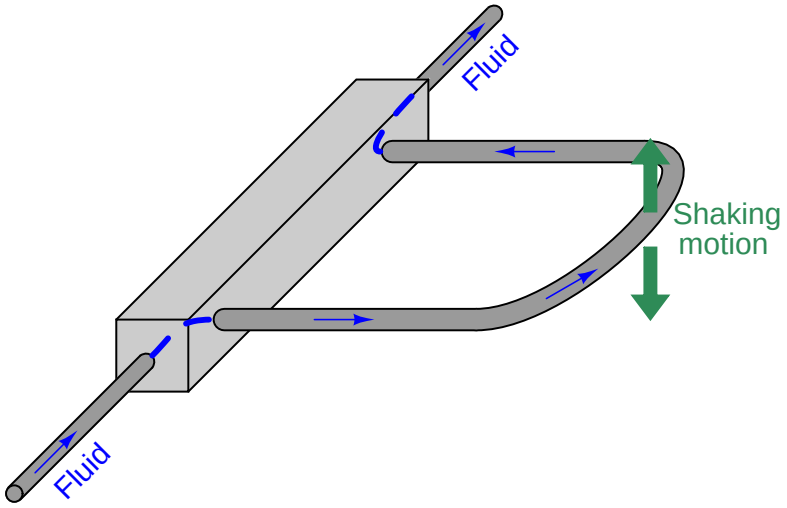
As you might guess, it can be difficult to engineer a tubing system capable of spinning in circles while carrying a flowstream of pressurized fluid. To bypass the practical difficulties of building a spinning tube system, Coriolis flowmeters are instead built on the principle of a flexible tube that *oscillates* back and forth, producing the same effect in a cyclic rather than continuous fashion. The effect is not unlike shaking a hose side to side as it carries a stream of water:



(This illustration is from a vertical view, looking down. The Coriolis force acts laterally, bending the hose to the side.)

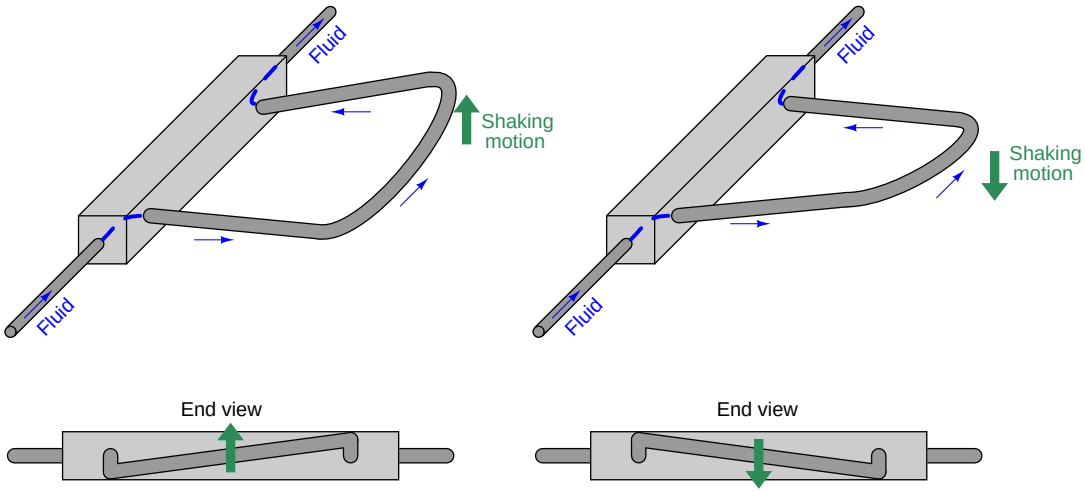
The Coriolis force acts to oppose the direction of rotation. The greater the mass flow rate of water through the hose, the stronger the Coriolis force. If we had a way to precisely measure the Coriolis force imparted to the hose by the water stream, and to precisely wave the hose so its rotational velocity held constant for every wave, we could directly infer the water's mass flow rate.

We cannot build a Coriolis flowmeter exactly like the water hose or lawn sprinkler unless we are willing to let the process fluid exit the tubing, so a common Coriolis flowmeter design uses a U-shaped tube that redirects the fluid flow back to the center of rotation. The curved end of the flexible U-tube is forced to shake back and forth by an electromagnetic force coil (like the force coil on an audio speaker) while the tube ends anchor to a stationary manifold:

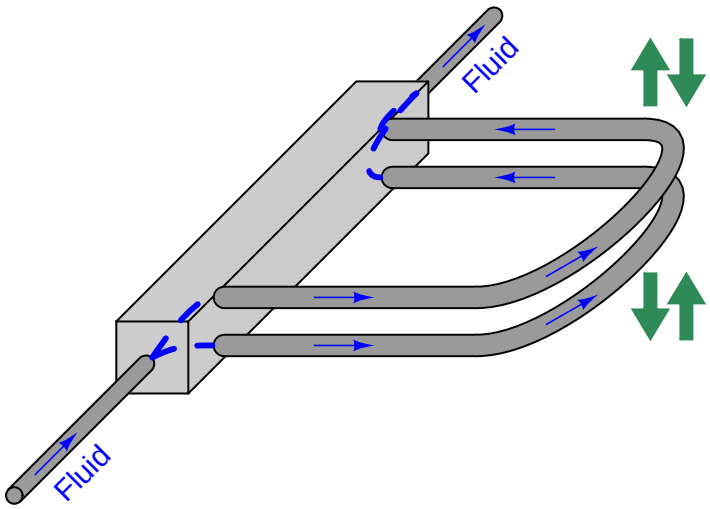


If fluid inside the tube is stagnant (no flow), the tube will simply vibrate back and forth with the applied force. However, if fluid *flows* through the tube, the moving fluid molecules will experience acceleration as they travel from the anchored base to the tube's rounded end, then experience *deceleration* as they travel back to the anchored base. This continual acceleration and subsequent deceleration of new mass generates a Coriolis force that alters the tube's motion.

This Coriolis force causes the U-tube assembly to *twist*. The tube portion carrying fluid from the anchored base to the end tends to *lag* in motion because the fluid molecules in that section of the tube are being accelerated to a greater lateral velocity. The tube portion carrying fluid from the end back to the anchored base tends to *lead* in motion because those molecules are being decelerated back to zero lateral velocity. As mass flow rate through the tube increases, so does the degree of twisting. By monitoring the amplitude of this twisting motion, we may infer the mass flow rate of the fluid passing through the tube:



In order to reduce the amount of vibration generated by a Coriolis flowmeter, and more importantly to reduce the effect any external vibrations may have on the flowmeter, two identical U-tubes are built next to each other and shaken in complementary fashion (always moving in opposite directions)[48]. Tube twist is measured as *relative* motion from one tube to the next, not as motion between the tube and the stationary housing of the flowmeter. This (ideally) eliminates the effect of any common-mode vibrations on the inferred flow measurement:



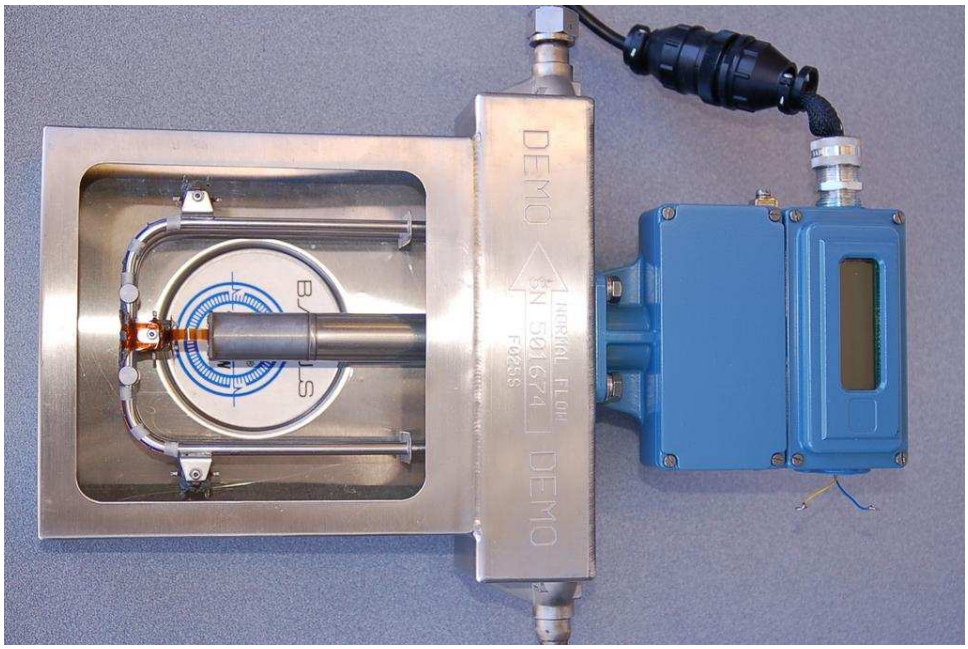
Viewed from the end, the complimentary shaking and twisting of the tubes looks like this:

Great care is taken by the manufacturer to ensure the two tubes are as close to identical as possible: not only are their physical characteristics precisely matched, but the fluid flow is split very evenly between the tubes[49] so their respective Coriolis forces should be identical in magnitude.

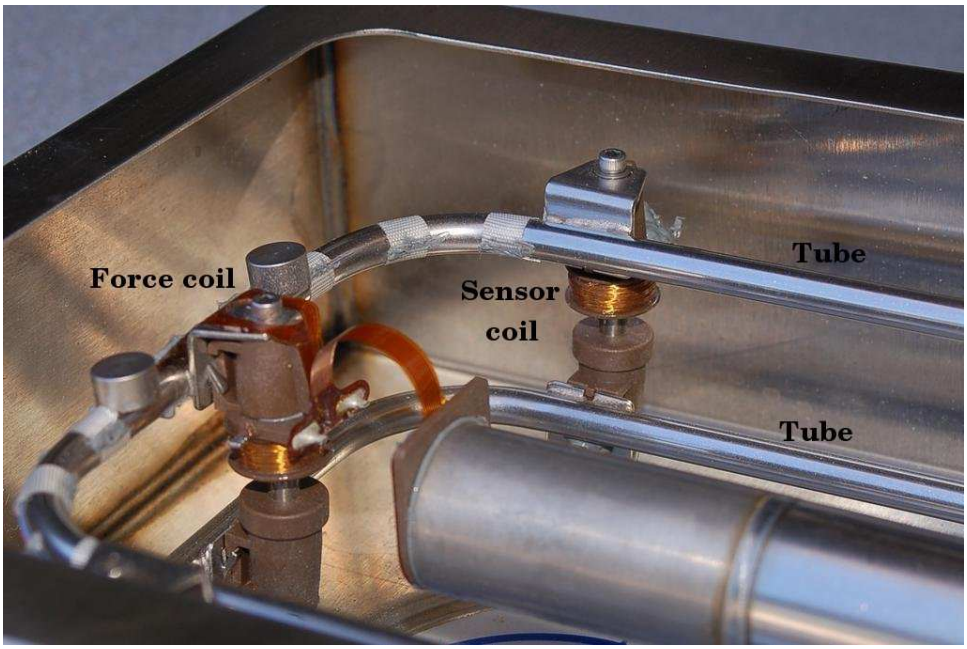
[48] For those readers with an automotive bent, this is the same principle applied in opposed-cylinder engines (e.g. Porsche "boxer" air-cooled 6-cylinder engine, Volkswagen air-cooled 4-cylinder engine, BMW air-cooled motorcycle twin engine, Citroen 2CV 2-cylinder engine, Subaru 4- and 6-cylinder opposed engines, etc.). Opposite piston pairs are *always* 180^o out of phase for the purpose of maintaining mechanical balance: both moving away from the crankshaft or both moving toward the crankshaft, at any given time.

[49] An alternative to splitting the flow is to plumb the tubes in series so they *must* share the exact same flow rate, like series-connected resistors sharing the exact same amount of electrical current.

A photograph of a Rosemount (Micro-Motion) U-tube Coriolis flowmeter demonstration unit shows the U-shaped tubes (one tube is directly above the other in this picture, so you cannot tell there are actually two U-tubes):



A closer inspection of this flowmeter shows that there are actually two U-tubes, one positioned directly above the other, shaken in complementary directions by a common electromagnetic force coil:

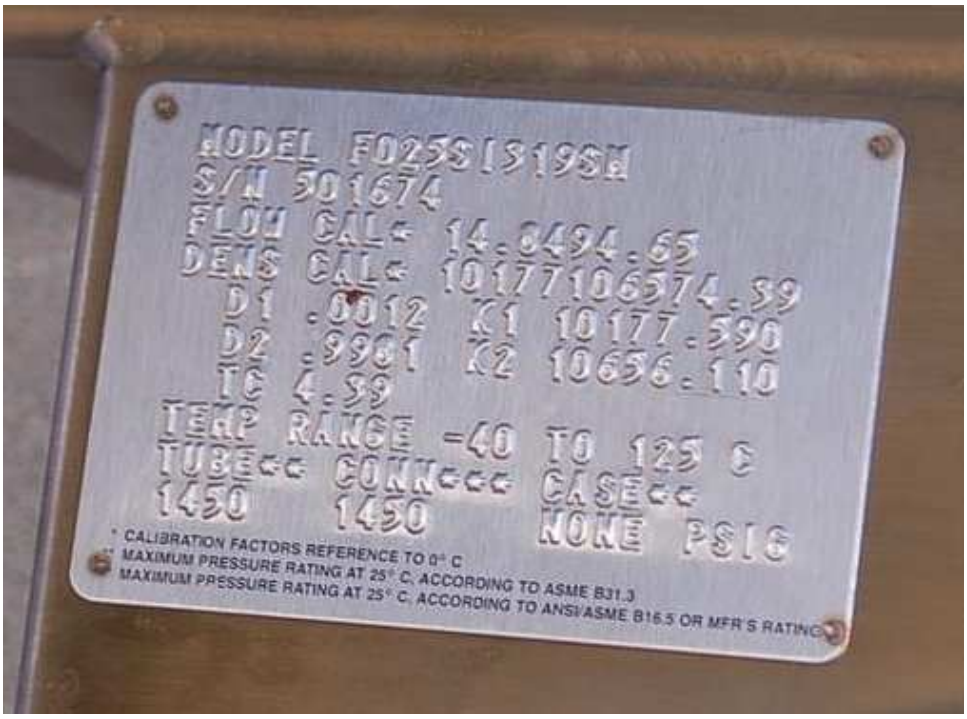


The force coil works on the same principle as an audio speaker: AC electric current passed through a wire coil generates an oscillating magnetic field, which acts against a permanent magnet's field to produce an oscillating force. In the case of an audio speaker, this force causes a lightweight cone to move, which then creates sound waves through the air. In the case of the Coriolis meter assembly, the force shakes the metal tubes back and forth.

Two magnetic displacement sensors monitor the relative motions of the tubes and transmit signals to an electronics module for digital processing. One of those sensor coils may be seen in the previous photograph. Both the force coil and the sensor coil are nothing more than permanent magnets surrounded by movable copper wire coils. The main difference between the force coil and the sensor coil is that the force coil is powered by an AC signal to impart a vibratory force to the tubes, whereas the sensor coils are both unpowered so they can detect tube motion by generating AC voltages to be sensed by the electronics module. The force coil is shown in the left-hand photograph, while one of the two sensor coils appears in the right-hand photograph:

Advances in sensor technology and signal processing have allowed the construction of Coriolis flowmeters employing straighter tubes than the U-tube unit previously illustrated and photographed. Straighter tubes are advantageous for reasons of reduced plugging potential and the ability to easily drain all liquids out of the flowmeter when needed.

The tubes of a Coriolis flowmeter are not just conduits for fluid flow, they are also precision spring elements. As such, it is important to precisely know the spring constant value of these tubes so the Coriolis force may be inferred from tube displacement (i.e. how far the tubes twist). Every Coriolis flow element is factory-tested to determine the flow tubes' mechanical properties, then the electronic transmitter is programmed with the various constant values describing those properties. The following photograph shows a close-up view of the nameplate on a Rosemount (Micro-Motion) Coriolis mass flowmeter, showing the physical constant values determined for that specific flowtube assembly at the time of manufacture:



This means every Coriolis flowmeter element (the tube and sensor assembly) is unique, with no two identical in behavior. Consequently, the transmitter (the electronics package outputting the process variable signals) must be programmed with values describing the element's behavior, and the complete flowmeter is shipped from the manufacturer as a *matched set*. You cannot interchange elements and transmitters without re-programming the transmitters with the new elements' physical constant values.

Coriolis flowmeters are equipped with RTD temperature sensors to continuously monitor the process fluid temperature. Fluid temperature is important to know because it affects certain properties of the tubes (e.g. spring constant, diameter, and length). The temperature indication is usually accessible as an auxiliary output, which means a Coriolis flowmeter may double as a (very expensive!) temperature transmitter.

Another variable is measured and (potentially) transmitted by a Coriolis flowmeter, and this variable is fluid *density*. The tubes within a Coriolis flowmeter are shaken at their mechanical

resonant frequency to maximize their shaking motion with the least amount of applied power to the force coil possible. The electronics module continuously varies the force coil's AC excitation frequency to maintain mechanical resonance. This resonant frequency happens to change with process fluid density, since the effective mass of the fluid-filled tubes changes with process fluid density[50], and mass is one of the variables influencing the resonant frequency of any physical object. Note the "mass" term in the following formula, describing the resonant frequency of a tensed string:

$$f = \frac{1}{2L}\sqrt{\frac{F_T}{\mu}}$$

Where,

f = Fundamental resonant frequency of string (Hertz)
L = String length (meters)
F_T = String tension (newtons)
μ = Unit mass of string (kilograms per meter)

This means fluid density, along with fluid temperature, is another variable measured by a Coriolis flowmeter. The ability to simultaneously measure these three variables (mass flow rate, temperature, and density) makes the Coriolis flowmeter a very versatile instrument indeed. This is especially true when the flowmeter in question communicates digitally using a "fieldbus" standard rather than an analog 4-20 mA signal. Fieldbus communication allows multiple variables to be transmitted by the device to the host system (and/or to other devices on the same fieldbus network), allowing the Coriolis flowmeter to do the job of three instruments!

An example of a Coriolis mass flowmeter being used as a multi-variable transmitter appears in the following photographs. Note the instrument tag labels in the close-up photograph (FT, TT, and DT), documenting its use as a flow transmitter, temperature transmitter, and density transmitter, respectively:

[50] If you consider each tube as a container with a fixed volume capacity, a change in fluid density (e.g. pounds per cubic foot) must result in a change in mass for each tube.

Even though a Coriolis flowmeter inherently measures *mass* flow rate, the continuous measurement of fluid density allows the meter to calculate *volumetric flow rate* if this is the preferred means of expressing fluid flow. The relationship between mass flow (W), volumetric flow (Q), and mass density (ρ) is quite simple:

$$W = \rho Q \qquad\qquad Q = \frac{W}{\rho}$$

All the flowmeter's computer must do to output a volumetric flow measurement is take the mass flow measurement value and divide that by the fluid's measured density. A simple exercise in dimensional analysis (performed with metric units of measurement) validates this concept for both forms of the equation shown above:

$$\left[\frac{\text{kg}}{\text{s}}\right] = \left[\frac{\text{kg}}{\text{m}^3}\right]\left[\frac{\text{m}^3}{\text{s}}\right] \qquad\qquad \left[\frac{\text{m}^3}{\text{s}}\right] = \frac{\left[\frac{\text{kg}}{\text{s}}\right]}{\left[\frac{\text{kg}}{\text{m}^3}\right]}$$

Coriolis mass flowmeters are very accurate and dependable. They are also completely immune to swirl and other fluid disturbances, which means they may be located nearly anywhere in a piping system with no need at all for straight-run pipe lengths upstream or downstream of the meter. Their natural ability to measure true mass flow, along with their characteristic linearity and accuracy, makes them ideally suited for custody transfer applications (where the flow of fluid represents product being bought and sold).

Perhaps the greatest disadvantage of Coriolis flowmeters is their high initial cost, especially for large pipe sizes. Coriolis flowmeters are also more limited in operating temperature than other types of flowmeters and may have difficulty measuring low-density fluids (gases) and mixed-phase[51] (liquid/vapor) flows. The bent tubes used to sense process flow may also trap process fluid inside to the point where it becomes unacceptable for hygienic (e.g. food processing, pharmaceuticals) applications. Straight-tube Coriolis flowmeter designs, and designs where the angle of the tubes is slight, fare better in this regard than the traditional U-tube Coriolis flowmeter design. However, an advantage of U-shaped tubes is that they aren't as stiff as straight tubes, and so straight-tube Coriolis flowmeters tend to be less sensitive to low flow rates than their U-tube counterparts.

[51] Significant technological progress has been made on mixed-phase Coriolis flow measurement, to the point where this may no longer be a serious consideration in the future.

21.7.2 Thermal flowmeters

Wind chill is a phenomenon common to nearly everyone who has ever lived in a cold environment. When the ambient air temperature is substantially colder than the temperature of your body, heat will transfer from your body to the surrounding air. If there is no breeze to move air past your body, the air molecules immediately surrounding your body will begin to warm up as they absorb heat from your body, which will then decrease the rate of heat loss. However, if there is even a slight breeze of air moving past your body, your body will come into contact with more cool (unheated) air molecules than it would otherwise, causing a greater rate of heat loss. Thus, your perception of the surrounding temperature will be cooler than if there were no breeze.

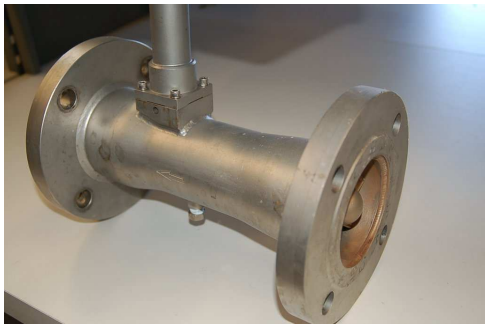
We may exploit this principle to measure mass flow rate, by placing a heated object in the midst of a fluid flowstream, and measuring how much heat the flowing fluid convects away from the heated object. The "wind chill" experienced by that heated object is a function of true mass flow rate (and not just volumetric flow rate) because the mechanism of heat loss is the rate at which fluid molecules contact the heated object, with each of those molecules having a definite mass.

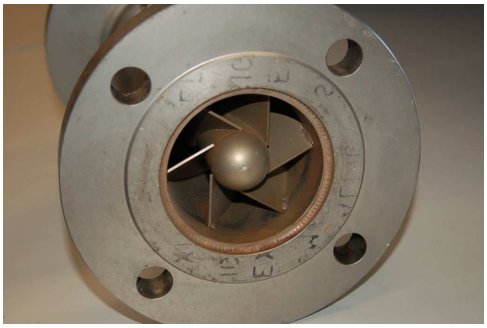
The simplest form of thermal mass flowmeter is the *hot-wire anemometer*, used to measure air speed. This flowmeter consists of a metal wire through which an electric current is passed to heat it up. An electric circuit monitors the resistance of this wire (which is directly proportional to wire temperature because most metals have a definite temperature coefficient of resistance). If air speed past the wire increases, more heat will be drawn away from the wire and cause its temperature to drop. The circuit senses this temperature change and compensates by increasing current through the wire to bring its temperature back up to setpoint. The amount of current sent through the wire becomes a representation of mass air flow rate past the wire.

Most mass air flow sensors used in automotive engine control applications employ this principle. It is important for engine control computers to measure *mass* air flow and not just volumetric air flow because it is important to maintain proper air/fuel ratio even if the air density changes due to changes in altitude. In other words, the computer needs to know how many air molecules are entering the engine per second in order to properly meter the correct amount of fuel into the engine for complete and efficient combustion. The "hot wire" mass air flow sensor is simple and inexpensive to produce in quantity, which is why it finds common use in automotive applications.

Industrial thermal mass flowmeters usually consist of a specially designed "flowtube" with two temperature sensors inside: one that is heated and one that is unheated. The heated sensor acts as the mass flow sensor (cooling down as flow rate increases) while the unheated sensor serves to compensate for the "ambient" temperature of the process fluid.

A typical thermal mass flowtube appears in the following diagrams (note the swirl vanes in the close-up photograph, designed to introduce large-scale turbulence into the flowstream to maximize the convective cooling effect of the fluid against the heated sensor element):





The simple construction of thermal mass flowmeters allows them to be manufactured in very small sizes. The following photograph shows a small device that is not only a mass flow meter, but also a mass flow *controller* with its own built-in throttling valve mechanism and control electronics. To give you a sense of scale, the tube fittings seen on the left- and right-hand sides of this device are 1/4 inch, making this photograph nearly full-size:

An important factor in the calibration of a thermal mass flowmeter is the *specific heat* of the process fluid. "Specific heat" is a measure of the amount of heat energy needed to change the temperature of a standard quantity of substance by some specified amount[52]. Some substances have much greater specific heat values than others, meaning those substances have the ability to absorb (or release) a lot of heat energy without experiencing a great temperature change. Fluids with high specific heat values make good *coolants*, because they are able to remove much heat energy from hot objects without experiencing great increases in temperature themselves. Since thermal mass

[52]For example, the specific heat of water is 1.00 kcal / kg $\cdot$ C^o, meaning that the addition of 1000 calories of heat energy is required to raise the temperature of 1 kilogram of water by 1 degree Celsius, or that we must remove 1000 calories of heat energy to cool that same quantity of water by 1 degree Celsius. Ethyl alcohol, by contrast, has a specific heat value of only 0.58 kcal / kg $\cdot$ C^o, meaning it is almost twice as easy to warm up or cool down as water (little more than half the energy required to heat or cool water needs to be transferred to heat or cool the same mass quantity of ethyl alcohol by the same amount of temperature).

flowmeters work on the principle of convective cooling, this means a fluid having a high specific heat value will elicit a greater response from a thermal mass flowmeter than the exact same mass flow rate of a fluid having a lesser specific heat value (i.e. a fluid that is not as good of a coolant).

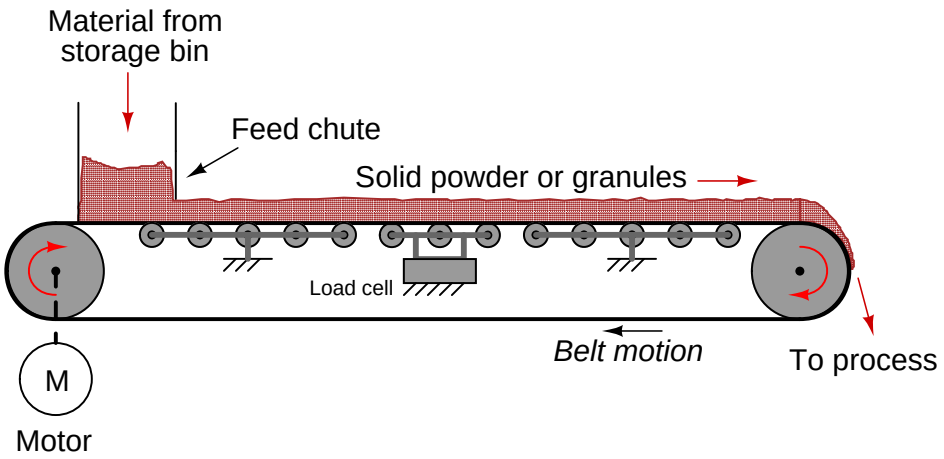
This means we must know the specific heat value of whatever fluid we plan to measure with a thermal mass flowmeter, and we must be assured its specific heat value will remain constant. For this reason, thermal mass flowmeters are not suitable for measuring the flow rates of fluid streams whose chemical composition is likely to change over time. This limitation is analogous to that of a pressure sensor used to hydrostatically measure the level of liquid in a vessel: in order for this level-measurement technique to be accurate, we must know the density of the liquid and also be assured that density will be constant over time.

Thermal mass flowmeters are simple and reliable instruments. While not as accurate or tolerant of piping disturbances as Coriolis mass flowmeters, they are far less expensive.

Perhaps the greatest disadvantage of thermal mass flowmeters is their sensitivity to changes in the specific heat of the process fluid. This makes the calibration of any thermal mass flowmeter specific for one composition of fluid only. In some applications such as automotive engine intake air flow, where the fluid composition is constant, this limitation is not a factor. In many industrial applications, however, this limitation is severe enough to prohibit the use of thermal mass flowmeters. Industrial applications for thermal mass flowmeters include natural gas flow measurement (non-custody transfer), and the measurement of purified gas flows (oxygen, hydrogen, nitrogen) where the composition is known to be very stable.

21.8 Weighfeeders

A completely different kind of flowmeter is the *weighfeeder*, used to measure the flow of solid material such as powders and grains. One of the most common weighfeeder designs consists of a conveyor belt with a section supported by rollers coupled to one or more load cells, such that a fixed length of the belt is continuously weighed:



The load cell measures the weight of a fixed-length belt section, yielding a figure of material weight per linear distance on the belt. A tachometer (speed sensor) measures the speed of the belt. The product of these two variables is the mass flow rate of solid material "through" the weighfeeder:

$$W = \frac{FS}{d}$$

Where,

W = Mass flow rate (e.g. pounds per second)
F = Force of gravity acting on the weighed belt section (e.g. pounds)
S = Belt speed (e.g. feet per second)
d = Length of weighed belt section (e.g. feet)

21.9 Change-of-quantity flow measurement

Flow, by definition, is the passage of material from one location to another over time. So far this chapter has explored technologies for measuring flow rate en route from source to destination. However, a completely different method exists for measuring flow rates: measuring how much material has either departed or arrived at the terminal locations over time.

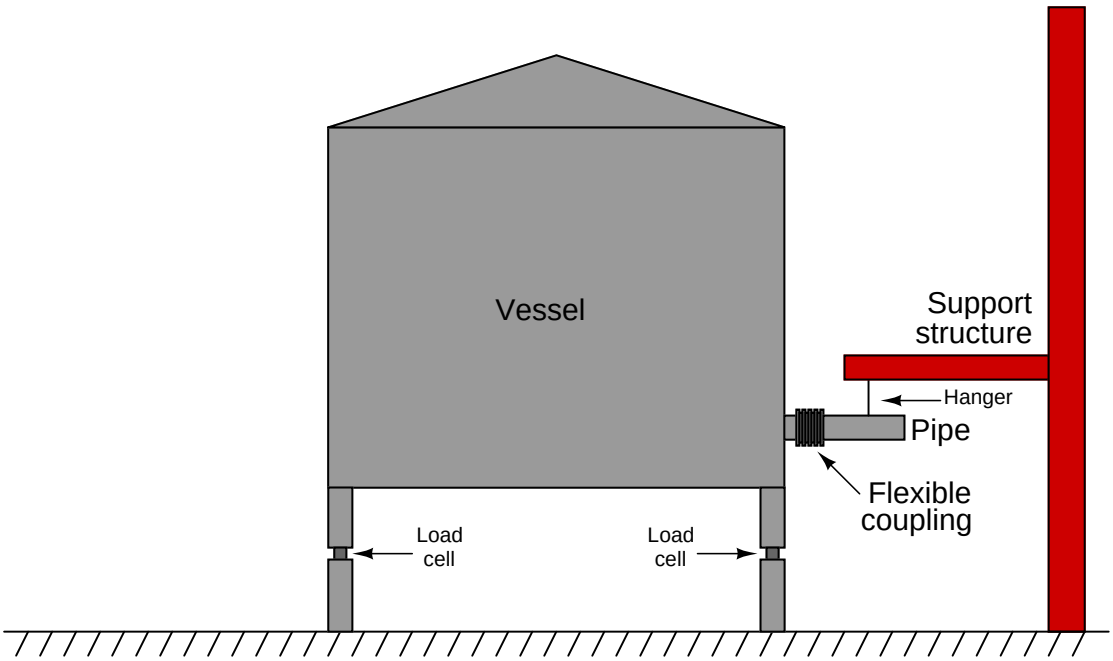
Mathematically, we may express flow as a ratio of quantity to time. Whether it is volumetric flow or mass flow we are referring to, the concept is the same: quantity of material moved per quantity of time. We may express average flow rates as ratios of changes:

$$\overline{W} = \frac{\Delta m}{\Delta t} \qquad\qquad \overline{Q} = \frac{\Delta V}{\Delta t}$$

Where,

$\overline{W}$ = Average mass flow rate
$\overline{Q}$ = Average volumetric flow rate
Δm = Change in mass
ΔV = Change in volume
Δt = Change in time

Suppose a water storage vessel is equipped with load cells to precisely measure weight (which is directly proportional to mass with constant gravity). Assuming only one pipe entering or exiting the vessel, any flow of water through that pipe will result in the vessel's total weight changing over time:



If the measured mass of this vessel decreased from 74,688 kilograms to 70,100 kilograms between 4:05 AM and 4:07 AM, we could say that the average mass flow rate of water leaving the vessel is 2,294 kilograms per minute over that time span.

$$\overline{W} = \frac{\Delta m}{\Delta t} = \frac{70100 \text{ kg} - 74688 \text{ kg}}{4{:}07 - 4{:}05} = \frac{-4588 \text{ kg}}{2 \text{ min}} = 2294 \frac{\text{kg}}{\text{min}}$$

Note that this average flow measurement may be determined without any flowmeter of any kind installed in the pipe to intercept the water flow. All the concerns of flowmeters studied thus far (turbulence, Reynolds number, fluid properties, etc.) are completely irrelevant. We may measure practically any flow rate we desire simply by measuring stored weight (or volume) over time. A computer may do this calculation automatically for us if we wish, on practically any time scale desired.

Now suppose the practice of determining average flow rates every two minutes was considered too infrequent. Imagine that operations personnel require flow data calculated and displayed more often than just 30 times an hour. All we must do to achieve better time resolution is take weight (mass) measurements more often. Of course, each mass-change interval will be expected to be less with more frequent measurements, but the amount of time we divide by in each calculation will be proportionally smaller as well. If the flow rate happens to be absolutely steady, we may sample mass as frequently as we might like and we will still arrive at the same flow rate value as before (sampling mass just once every two minutes). If, however, the flow rate is not steady, sampling more often will allow us to better see the immediate "ups" and "downs" of flow behavior.

Imagine now that we had our hypothetical "flow computer" take weight (mass) measurements at an infinitely fast pace: an infinite number of samples per second. Now, we are no longer *averaging* flow rates over finite periods of time; instead we would be calculating *instantaneous* flow rate at any given *point* in time.

Calculus has a special form of symbology to represent such hypothetical scenarios: we replace the Greek letter "delta" (Δ, meaning "change") with the roman letter "d" (meaning *differential*). A simple way of picturing the meaning of "d" is to think of it as meaning an *infinitesimal* interval of whatever variable follows the "d" in the equation[53]. When we set up two differentials in a quotient, we call the $\frac{d}{d}$ fraction a *derivative*. Re-writing our average flow rate equations in derivative (calculus) form:

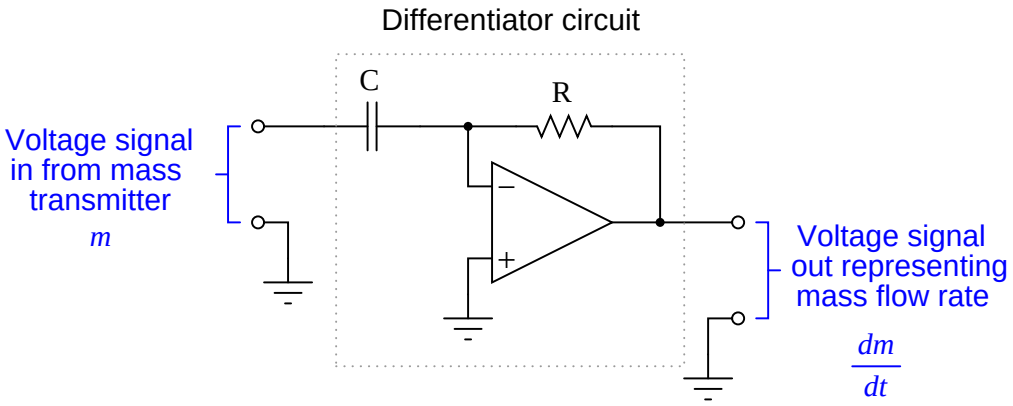
$$W = \frac{dm}{dt} \qquad\qquad Q = \frac{dV}{dt}$$

Where,

W = Instantaneous mass flow rate
Q = Instantaneous volumetric flow rate
dm = Infinitesimal (infinitely small) change in mass
dV = Infinitesimal (infinitely small) change in volume
dt = Infinitesimal (infinitely small) change in time

[53] While this may seem like a very informal definition of differential, it is actually rooted in a field of mathematics called *nonstandard analysis*, and closely compares with the conceptual notions envisioned by calculus' founders.

We need not dream of hypothetical computers capable of infinite calculations per second in order to derive a flow measurement from a mass (or volume) measurement. Analog electronic circuitry exploits the natural properties of resistors and capacitors to essentially do this very thing in real time[54]:



In the vast majority of applications you will see digital computers used to calculate average flow rates rather than analog electronic circuits calculating instantaneous flow rates. The broad capabilities of digital computers virtually ensures they will be used somewhere in the measurement/control system, so the rationale is to use the existing digital computer to calculate flow rates (albeit imperfectly) rather than complicate the system design with additional (analog) circuitry. As fast as modern digital computers are able to process simple calculations such as these anyway, there is little practical reason to prefer analog signal differentiation except in specialized applications where high speed performance is paramount.

Perhaps the single greatest disadvantage to inferring flow rate by differentiating mass or volume measurements over time is the requirement that the storage vessel have but one flow path in and out. If the vessel has multiple paths for liquid to move in and out (simultaneously), any flow rate calculated on change-in-quantity will be a *net* flow rate only. It is impossible to use this flow measurement technique to measure one flow out of multiple flows common to one liquid storage vessel.

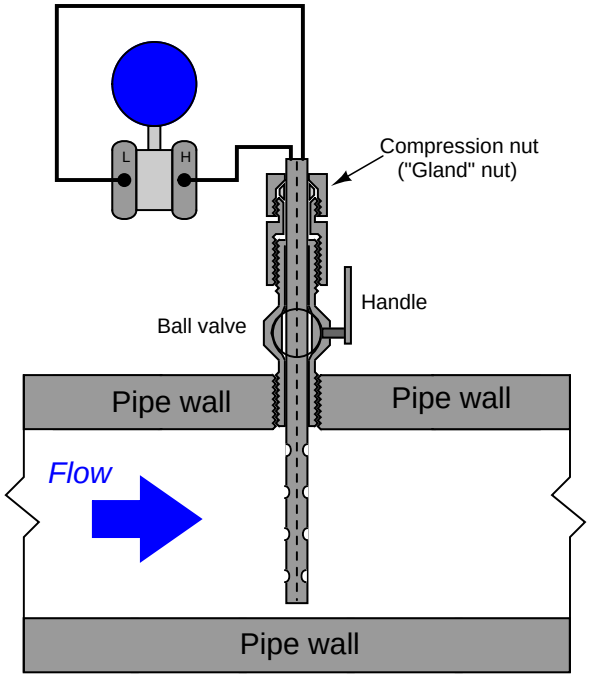
A simple "thought experiment" confirms this fact. Imagine a water storage vessel receiving a flow rate in at 200 gallons per minute. Next, imagine that same vessel emptying water out of a second pipe at the exact same flow rate: 200 gallons per minute. With the exact same flow rate both entering and exiting the vessel, the water level in the vessel will remain constant. Any change-of-quantity flow measurement system would register zero change in mass or volume over time, consequently calculating a flow rate of absolutely zero. Truly, the *net* flow rate for this vessel is zero, but this tells us nothing about the flow in each pipe, except that those flow rates are equal in magnitude and opposite in direction.

[54] To be precise, the equation describing the function of this analog differentiator circuit is: $V_{out} = -RC\frac{dV_{in}}{dt}$. The negative sign is an artifact of the circuit design – being essentially an inverting amplifier with negative gain – and not an essential element of the math.

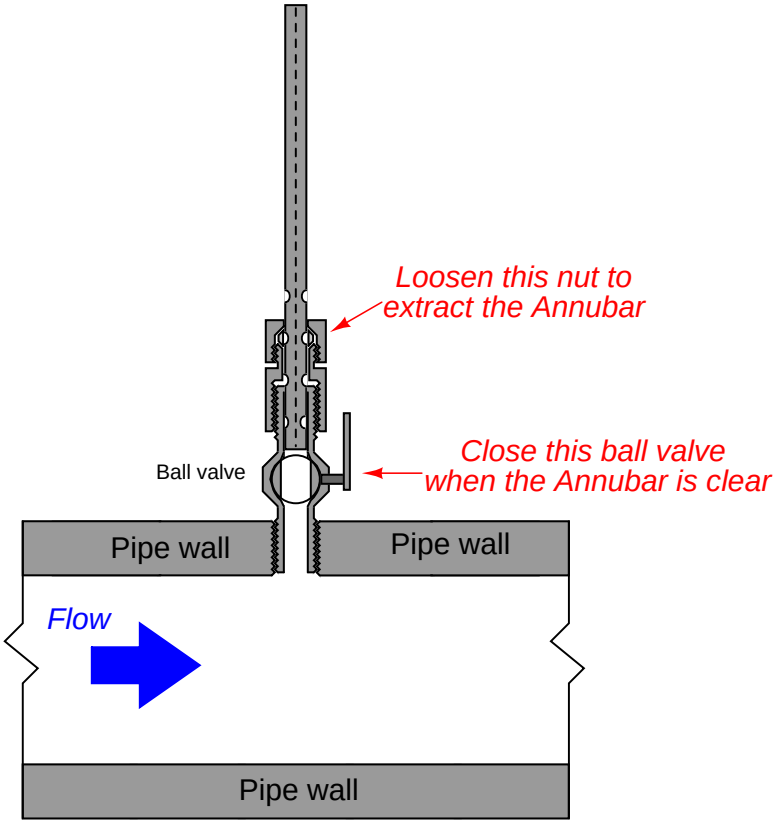
21.10 Insertion flowmeters

This section does not describe a particular type of flowmeter, but rather a design that may be implemented for several different kinds of flow measurement technologies. When the pipe carrying process fluid is large in size, it may be impractical or cost-prohibitive to install a full-diameter flowmeter to measure fluid flow rate. A practical alternative for many applications is the installation of an *insertion* flowmeter: a probe that may be inserted into or extracted from a pipe, to measure fluid velocity in one region of the pipe's cross-sectional area (usually the center).

A classic example of an insertion flowmeter element is the *Annubar*, a form of averaging pitot tube pioneered by the Dieterich Standard corporation. The Annubar flow element is inserted into a pipe carrying fluid where it generates a differential pressure for a pressure sensor to measure:

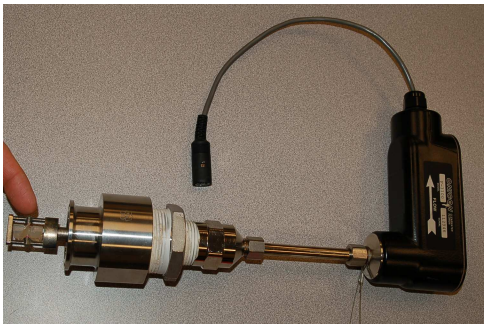


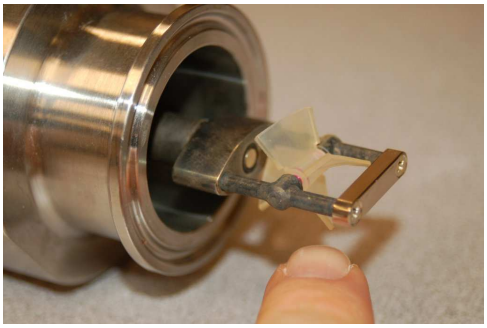
The Annubar element may be extracted from the pipe by loosening a "gland nut" and pulling the assembly out until the end passes through a hand ball valve. Once the element has been extracted this far, the ball valve may be shut and the Annubar completely removed from the pipe:



For safety reasons, a "stop" is usually built into the assembly to prevent someone from accidently pulling the element all the way out with the valve still open.

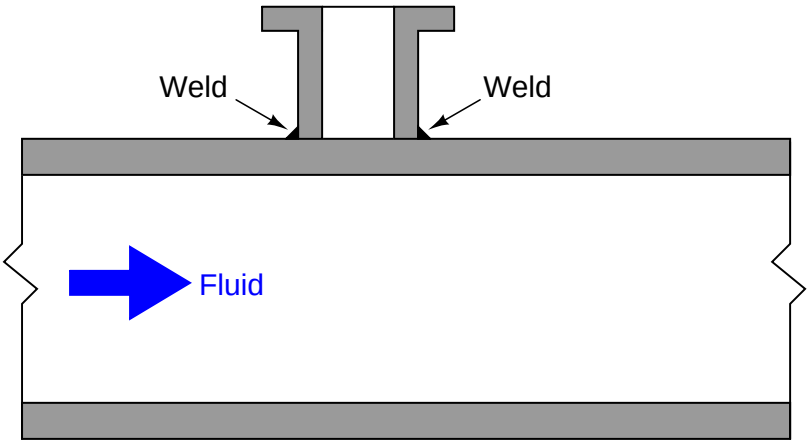
Other flowmeter technologies manufactured in insertion form include vortex, turbine, and thermal mass. An insertion-type turbine flowmeter appears in the following photographs:



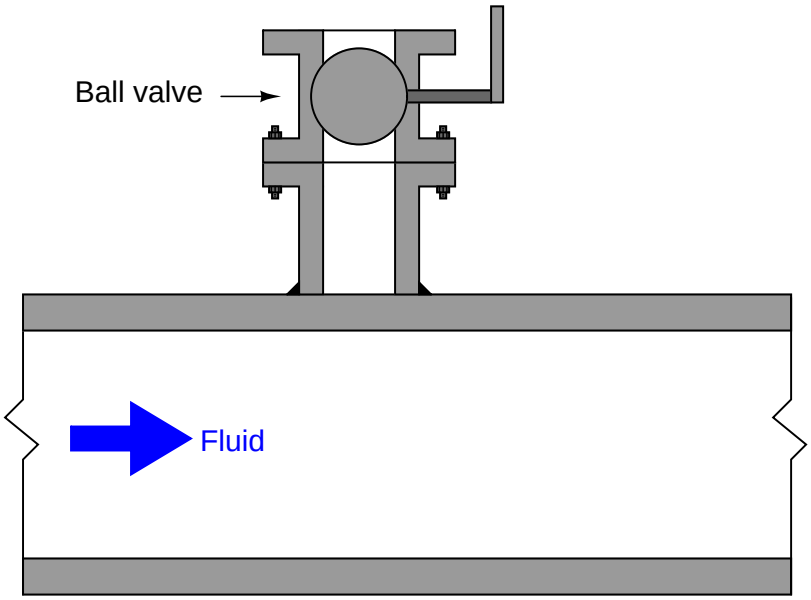


If the flow-detection element is compact rather than distributed (as is certainly the case with the turbine flowmeter shown above), care must be taken to ensure correct positioning within the pipe. Since flow profiles are never completely flat, any insertion meter element will register a greater flow rate at the center of the pipe than near the walls. Wherever the insertion element is placed in the pipe diameter, that placement must remain consistent through repeated extractions and re-insertions or else the effective calibration of the insertion flowmeter will change every time it is removed and re-inserted into the pipe. Care must also be taken to insert the flowmeter so the flow element points directly upstream, and not at an angle.

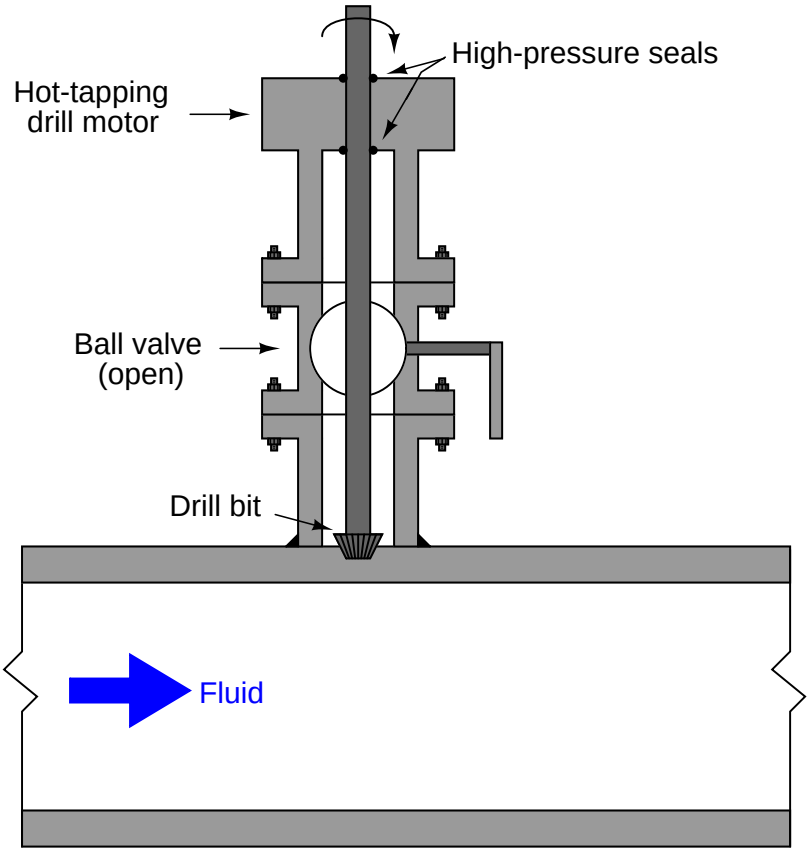
A unique advantage of insertion instruments is that they may be installed in an operating pipe by using specialized *hot-tapping* equipment. A "hot tap" is a procedure whereby a safe penetration is made into a pipe while the pipe is carrying fluid under pressure. The first step in a hot-tapping operation is to weld a "saddle tee" fitting on the side of the pipe:



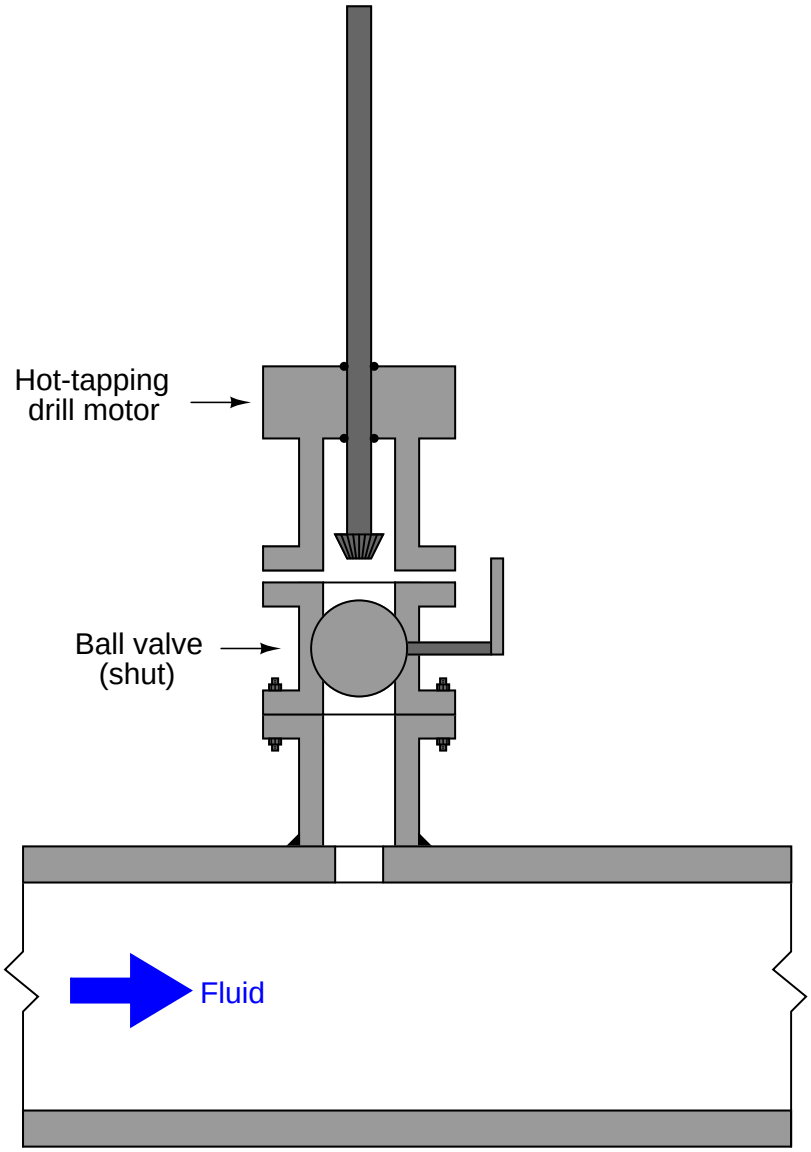
Next, a ball valve is bolted onto the saddle tee flange. This ball valve will be used to isolate the insertion instrument from the fluid pressure inside the pipe:



A special hot-tapping drill is then bolted to the open end of the ball valve. This drill uses a high-pressure seal to contain fluid pressure inside the drill chamber as a motor spins the drill bit. The ball valve is opened, then the drill bit is advanced toward the pipe wall where it cuts a hole into the pipe. Fluid pressure rushes into the empty chamber of the ball valve and hot-tapping drill as soon as the pipe wall is breached:



Once the hole has been completely drilled, the bit is extracted and the ball valve shut to allow removal of the hot-tapping drill:



Now there is a flanged and isolated connection into the "hot" pipe, through which an insertion flowmeter (or other instrument/device) may be installed.

Hot-tapping is a technical skill, with many safety concerns specific to different process fluids, pipe types, and process applications. This brief introduction to the technique is not intended to be instructional, but merely informational.

21.11 Process/instrument suitability

Every flow-measuring instrument exploits a physical principle to measure the flow rate of fluid stream. Understanding each of these principles as they apply to different flow-measurement technologies is the first and most important step in properly applying a suitable technology to the measurement of a particular process stream flow rate. The following table lists the specific operating principles exploited by different flow measurement technologies:

Flow measurement technology	**Operating principle**	**Linearity**	**2-way flow**
Differential pressure	Fluid mass self-acceleration, potential-kinetic energy exchange	$\sqrt{\Delta P}$	(some)
Laminar	Viscous fluid friction	linear	yes
Weirs & flumes	Fluid mass self-acceleration, potential-kinetic energy exchange	H^n	no
Turbine (velocity)	Fluid velocity spinning a vaned wheel	linear	yes
Vortex	von Kármán effect	linear	no
Magnetic	Electromagnetic induction	linear	yes
Ultrasonic	Sound wave time-of-flight	linear	yes
Coriolis	Fluid inertia, Coriolis effect	linear	yes
Turbine (mass)	Fluid inertia	linear	(some)
Thermal	Convective cooling, specific heat of fluid	linear	no
Positive displacement	Movement of fixed volumes	linear	(some)

A potentially important factor in choosing an appropriate flowmeter technology is energy loss caused by pressure drop. Some flowmeter designs, such as the common orifice plate, are inexpensive to install but carry a high price in terms of the energy lost in *permanent pressure drop* (the total, non-recoverable loss in pressure from the inlet of the device to the outlet, not the temporary pressure difference between inlet and vena contracta). Energy costs money, and so industrial facilities would be wise to consider the long-term cost of a flowmeter before settling on the one that is cheapest to install. It could very well be, for example, that an expensive venturi tube will cost less after years of operation than a cheap orifice plate[55].

In this regard, certain flowmeters stand above the rest: those with obstructionless flowtubes. Magnetic and ultrasonic flowmeters have no obstructions whatsoever in the path of the flow. This translates to (nearly) zero permanent pressure loss along the length of the tube, and therefore. Thermal mass and straight-tube Coriolis flowmeters are nearly obstructionless, while vortex and turbine meters are only slightly worse.

[55]This is not always the case, as primary elements are often found on throttled process lines. In such cases where a control valve normally throttles the flow rate, any energy dissipated by the orifice plate is simply less energy that the valve would otherwise be required to dissipate. Therefore, the presence or absence of an orifice plate has no net impact on energy dissipation when used on a process flow throttled by a control valve, and therefore does not affect cost over time due to energy loss.

References

AGA Report No. 3 – Orifice metering of natural gas and other related hydrocarbon fluids, Part 1 (General Equations and Uncertainty Guidelines), Catalog number XQ9017, American Gas Association and American Petroleum Institute, Washington D.C., Third Edition October 1990, Second Printing June 2003.

AGA Report No. 3 – Orifice metering of natural gas and other related hydrocarbon fluids, Part 2 (Specification and Installation Requirements), Catalog number XQ0002, American Gas Association and American Petroleum Institute, Washington D.C., Fourth Edition April 2000, Second Printing June 2003.

AGA Report No. 3 – Orifice metering of natural gas and other related hydrocarbon fluids, Part 3 (Natural Gas Applications), Catalog number XQ9210, American Gas Association and American Petroleum Institute, Washington D.C., Third Edition August 1992, Second Printing June 2003.

AGA Report No. 3 – Orifice metering of natural gas and other related hydrocarbon fluids, Part 4 (Background, Development, Implementation Procedure, and Subroutine Documentation for Empirical Flange-Tapped Discharge Coefficient Equation), Catalog number XQ9211, American Gas Association and American Petroleum Institute, Washington D.C., Third Edition October 1992, Second Printing August 1995, Third Printing June 2003.

Chow, Ven Te., *Open-Channel Hydraulics*, McGraw-Hill Book Company, Inc., New York, NY, 1959.

"Flow Measurement User Manual", Form Number A6043, Part Number D301224X012, Emerson Process Management, 2005.

Fribance, Austin E., *Industrial Instrumentation Fundamentals*, McGraw-Hill Book Company, New York, NY, 1962.

General Specifications: "EJX910A Multivariable Transmitter", Document GS 01C25R01-01E, 5th edition, Yokogawa Electric Corporation, Tokyo, Japan, 2005.

Giancoli, Douglas C., *Physics for Scientists & Engineers*, Third Edition, Prentice Hall, Upper Saddle River, NJ, 2000.

Hanlon, Paul C., *Compressor Handbook*, The McGraw-Hill Companies, New York, NY, 2001.

Hofmann, Friedrich, *Fundamentals of Ultrasonic Flow Measurement for industrial applications*, Krohne Messtechnik GmbH & Co. KG, Duisburg, Germany, 2000.

Hofmann, Friedrich, *Fundamental Principles of Electromagnetic Flow Measurement*, 3rd Edition, Krohne Messtechnik GmbH & Co. KG, Duisburg, Germany, 2003.

Improving Compressed Air System Performance – a sourcebook for industry, U.S. Department of Energy, Washington, DC, 2003.

Kallen, Howard P., *Handbook of Instrumentation and Controls*, McGraw-Hill Book Company, Inc.,

New York, NY, 1961.

Keisler, H. Jerome, *Elementary Calculus – An Infinitesimal Approach*, Second Edition, University of Wisconsin, 2000.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Measurement and Analysis Volume I*, Fourth Edition, CRC Press, New York, NY, 2003.

Miller, Richard W., *Flow Measurement Engineering Handbook*, Second Edition, McGraw-Hill Publishing Company, New York, NY, 1989.

Price, James F., *A Coriolis Tutorial*, version 3.3, Woods Hole Oceanographic Institution, Woods Hole, MA, 2006.

Spink, L. K., *Principles and Practice of Flow Meter Engineering*, Ninth Edition, The Foxboro Company, Foxboro, MA, 1967.

Tech-Spec: "SCFM (Standard CFM) vs. ACFM (Actual CFM)", Reference 15-010504.006, Sullair Corporation, 2004.

Vennard, John K., *Elementary Fluid Mechanics*, 3rd Edition, John Wiley & Sons, Inc., New York, NY, 1954.

Chapter 22

Continuous analytical measurement

In the field of industrial instrumentation and process control, the word *analyzer* generally refers to an instrument tasked with measuring the concentration of some substance, usually mixed with other substances of little or no interest to the controlled process. Unlike the other "bulk" measurement devices for sensing such general variables as pressure, level, temperature, or flow, an analytical device must *discriminately select* one material over all others present in the sample. This single problem accounts for much of the complexity of analytical instrumentation: *how do we measure the quantity of just one substance when thoroughly mixed with other substances?*

Analytical instruments generally achieve selectivity by measuring some property of the substance of interest unique to that substance alone, or at least unique to it among the possible substances likely to be found in the process sample. For example, an optically-based analyzer might achieve selectivity by measuring the intensities of only those particular wavelengths of light absorbed by the compound of interest, and absorbed by none of the other wavelengths. A "paramagnetic" oxygen gas analyzer achieves selectivity by exploiting the paramagnetic properties of oxygen gas, since no other industrial gas is as paramagnetic as oxygen. A pH analyzer achieves hydrogen ion selectivity by using a specially-prepared glass membrane intended to pass only hydrogen ions.

Problems are sure to arise if the measured property of the substance of interest is not as unique as originally thought. This may occur due to oversight on the part of the person originally choosing the analyzer technology, or it may occur as a result of changes made to the process chemistry, whether by intentional modification of the process equipment or by abnormal operating conditions. For example, a gas that happens to absorb some (or all!) of the same light wavelengths as the gas of interest will cause false measurements if not properly compensated for in the analyzer. Nitric oxide (NO) gas is one of the few gases also exhibiting significant paramagnetism, and as such will cause measurement errors if introduced into the sample inlet of a paramagnetic oxygen analyzer. A pH analyzer immersed in a liquid solution containing an abundance of sodium ions may fall victim to measurement errors, because sodium ions also happen to interact with the glass membrane of a pH electrode to generate a voltage.

For this reason, the student of analytical instrumentation must always pay close attention to the *underlying principle of measurement* for any analyzer technology, looking out for any ways that

analyzer may be "fooled" by the presence of some *other* substance than the one the analyzer was designed to measure.

22.1 Conductivity measurement

Electrical conductivity in metals is the result of free electrons drifting within a "lattice" of atomic nuclei comprising the metal object. When a voltage is applied across two points of a metal object, these free electrons immediately drift toward the positive pole (anode) and away from the negative pole (cathode).

Electrical conductivity in liquids is another matter entirely. Here, the charge carriers are *ions*: electrically imbalanced atoms or molecules that are free to drift because they are not "locked" into a lattice structure as is the case with solid substances. The degree of electrical conductivity of any liquid is therefore dependent on the ion density of the solution (how many ions freely exist per unit volume of liquid). When a voltage is applied across two points of a liquid solution, negative ions will drift toward the positive pole (anode) and positive ions will drift toward the negative pole (cathode). In honor of this directional drifting, negative ions are sometimes called *anions* (attracted to the *anode*), while positive ions are sometimes called *cations* (attracted to the *cathode*).

Electrical conductivity in gases is much the same: ions are the charge carriers. However, with gases at room temperature, ionic activity is virtually nonexistent. A gas must be superheated into a *plasma* state before substantial ions exist which can support an electric current.

22.1.1 Dissociation and ionization in aqueous solutions

Pure water is a very poor conductor of electricity. Some water molecules will "ionize" into unbalanced halves (instead of H_2O, you will find some negatively charged hydroxyl ions (OH^-) and some positively charged hydrogen ions[1] (H^+), but the percentage is extremely small at room temperature.

Any substance that enhances electrical conductivity when dissolved in water is called an *electrolyte*. This enhancement of conductivity occurs due to the molecules of the electrolyte separating into positive and negative ions, which are then free to serve as electrical charge carriers. If the electrolyte in question is an *ionically-bonded* compound[2] (table salt is a common example), the ions forming that compound naturally separate in solution, and this separation is called *dissociation*. If the electrolyte in question is a *covalently-bonded* compound[3] (hydrogen chloride is an example), the separation of those molecules into positive and negative ions is called *ionization*.

Both *dissociation* and *ionization* refer to the separation of formerly joined atoms upon entering a solution. The difference between these terms is the type of substance that splits: "dissociation" refers to the division of ionic compounds (such as table salt), while "ionization" refers to covalent-bonded (molecular) compounds such as HCl which are not ionic in their pure state.

Ionic impurities added to water (such as salts and metals) immediately dissociate and become available to act as charge carriers. Thus, the measure of a water sample's electrical conductivity is a fair estimate of ionic impurity concentration. Conductivity is therefore an important analytical measurement for certain water purity applications, such as the treatment of boiler feedwater, and the preparation of high-purity water used for semiconductor manufacturing.

It should be noted that conductivity measurement is a very *non-specific* form of analytical measurement. The conductivity of a liquid solution is a gross indication of its ionic content, but it tells us nothing specific about the *type* or *types* of ions present in the solution. Therefore, conductivity measurement is meaningful only when we have prior knowledge of the particular ionic species present in the solution (or when the purpose is to eliminate all ions in the solution such as in the case of ultra-pure water treatment, in which case we do not care about types of ions because our ideal goal is zero conductivity).

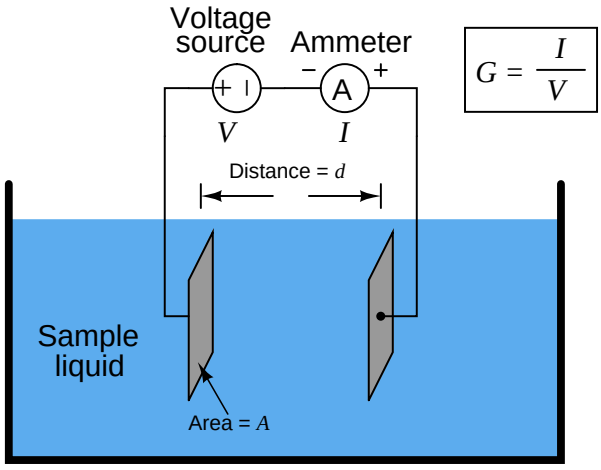
[1] Truth be told, free hydrogen ions are extremely rare in an aqueous solution. You are far more likely to find them bound to normal water molecules to form positive hydronium ions (H_3O^+). For simplicity's sake, though, professional literature often refers to these positive ions as "hydrogen" ions and even represent them symbolically as H^+.

[2] Ionic compounds are formed when oppositely charged atomic ions bind together by mutual attraction. The distinguishing characteristic of an ionic compound is that it is a conductor of electricity in its pure, liquid state. That is, it readily separates into anions and cations all by itself. Even in its solid form, an ionic compound is already ionized, with its constituent atoms held together by an imbalance of electric charge. Being in a liquid state simply gives those atoms the physical mobility needed to dissociate.

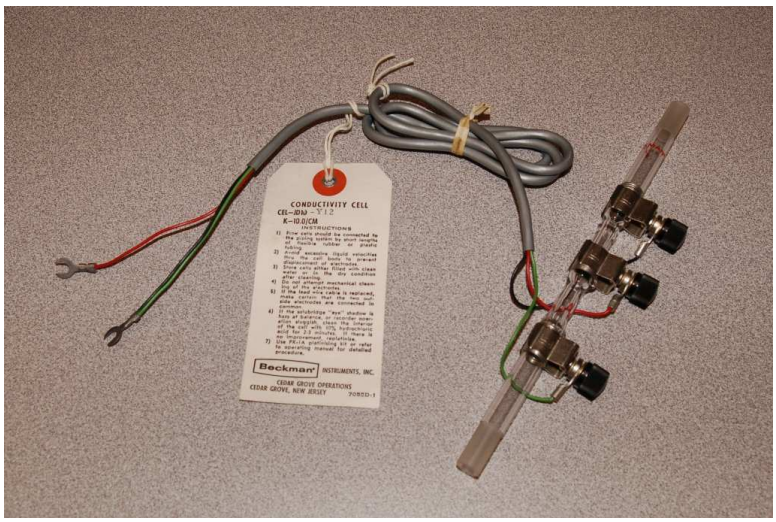
[3] Covalent compounds are formed when neutral atoms bind together by the sharing of valence electrons. Such compounds are not good conductors of electricity in their pure, liquid states.

22.1.2 Two-electrode conductivity probes

Conductivity is measured by an electric current passed through the solution. The most primitive form of conductivity sensor (sometimes referred to as a conductivity *cell*) consists of two metal electrodes inserted in the solution, connected to a circuit designed to measure conductance (G), the reciprocal of resistance ($\frac{1}{R}$):



The following photograph shows such direct-contact style of conductivity probe, consisting of stainless steel electrodes contacting the fluid flowing through a glass tube:



The conductance measured by a direct-contact conductivity instrument is a function of plate geometry (surface area and distance of separation) as well as the ionic activity of the solution. A simple increase in separation distance between the probe electrodes will result in a decreased conductance measurement (increased resistance R) even if the liquid solution's ionic properties

do not change. Therefore, conductance (G) is not particularly useful as an expression of liquid conductivity.

The mathematical relationship between conductance (G), plate area (A), plate distance (d), and the actual conductivity of the liquid (k) is expressed in the following equation[4]:

$$G = k\frac{A}{d}$$

Where,

G = Conductance, in Siemens (S)
k = Specific conductivity of liquid, in Siemens per centimeter (S/cm)
A = Electrode area (each), in square centimeters (cm^2)
d = Electrode separation distance, in centimeters (cm)

The unit of Siemens per centimeter may seem odd at first, but it is necessary to account for all the units present in the variables of the equation. A simple dimensional analysis proves this:

$$[\mathrm{S}] = \left[\frac{\mathrm{S}}{\mathrm{cm}}\right]\frac{[\mathrm{cm}^2]}{[\mathrm{cm}]}$$

For any particular conductivity cell, the geometry may be expressed as a ratio of separation distance to plate area, usually symbolized by the lower-case Greek letter Theta (θ), and always expressed in the unit of inverse centimeters (cm^{-1}):

$$\theta = \frac{d}{A}$$

Re-writing the conductance equation using θ instead of A and d, we see that conductance is the quotient of conductivity k and the cell constant θ:

$$G = \frac{k}{\theta}$$

Where,

G = Conductance, in Siemens (S)
k = Specific conductivity of liquid, in Siemens per centimeter (S/cm)
θ = Cell constant, in inverse centimeters (cm^{-1})

Manipulating this equation to solve for conductivity (k) given electrical conductance (G) and cell constant (θ), we have the following result:

$$k = G\theta$$

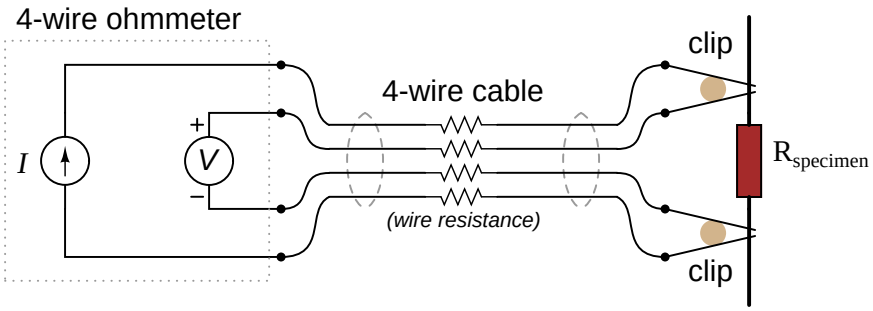
Two-electrode conductivity cells are not very practical in real applications, because mineral and metal ions attracted to the electrodes tend to "plate" the electrodes over time forming solid, insulating barriers on the electrodes. While this "electroplating" action may be substantially reduced

[4]This equation bears a striking similarity to the equation for resistance of metal wire: $R = \rho\frac{l}{A}$, where l is the length of a wire sample, A is the cross-sectional area of the wire, and ρ is the specific resistance of the wire metal.

by using AC instead of DC[5] to excite the sensing circuit, it is usually not enough. Over time, the conductive barriers formed by ions bonded to the electrode surfaces will create calibration errors by making the instrument "think" the liquid is less conductive than it actually is.

22.1.3 Four-electrode conductivity probes

A very old electrical technique known as the *Kelvin* or *four-wire* resistance-measuring method is a practical solution for this problem. Commonly employed to make precise resistance measurements for scientific experiments in laboratory conditions, as well as measuring the electrical resistance of strain gauges and other resistive sensors, the four-wire technique uses four conductors to connect the resistance under test to the measuring instrument:



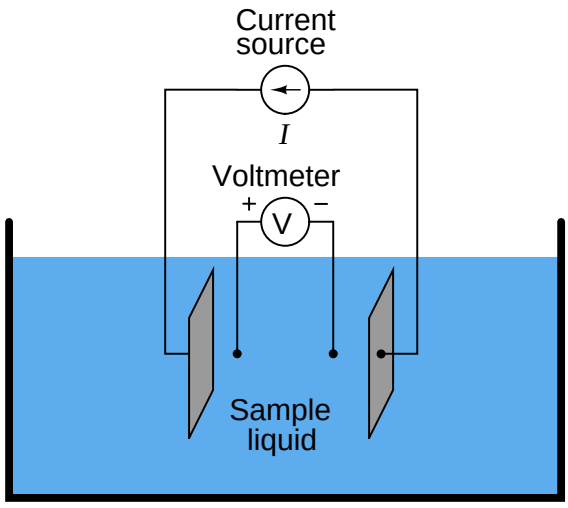
$$R_{specimen} = \frac{\text{Voltmeter indication}}{\text{Current source}}$$

Only the outer two conductors carry substantial current. The inner two conductors connecting the voltmeter to the test specimen carry negligible current (due to the voltmeter's extremely high input impedance) and therefore drop negligible voltage along their lengths. Voltage dropped across the current-carrying (outer) wires is irrelevant, since that voltage drop is never detected by the voltmeter.

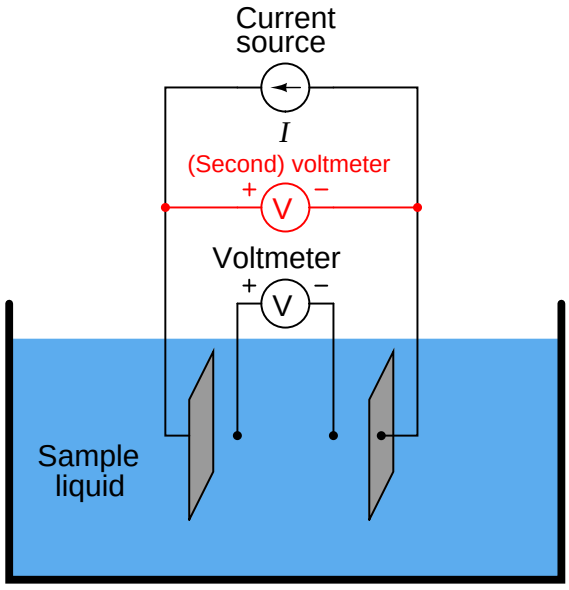
Since the voltmeter only measures voltage dropped across the specimen (the resistor under test), and not the test resistance plus wiring resistance, the resulting resistance measurement is much more accurate.

[5] The use of alternating current forces the ions to switch directions of travel many times per second, thus reducing the chance they have of bonding to the metal electrodes.

In the case of conductivity measurement, it is not wire resistance that we care to ignore, but rather the added resistance caused by plating of the electrodes. By using four electrodes instead of two, we are able to measure voltage dropped across a length of liquid solution *only*, and completely ignore the resistive effects of electrode plating:



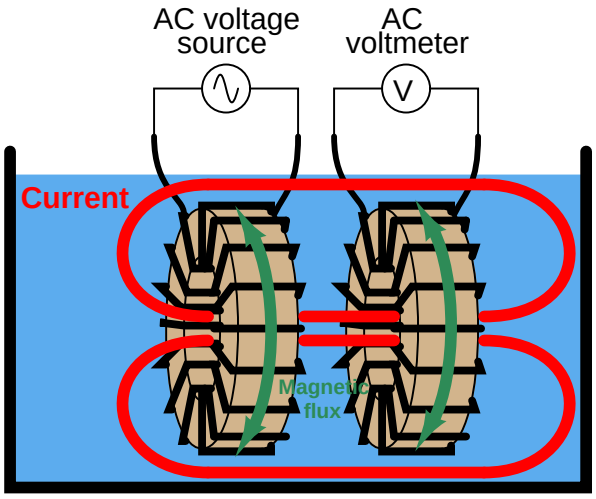
In the 4-wire conductivity cell, any electrode plating will merely burden the current source by causing it to output a greater voltage, but it will *not* affect the amount of voltage detected by the two inner electrodes as that electric current passes through the liquid. Some conductivity instruments employ a second voltmeter to measure the voltage dropped between the "excitation" electrodes, to indicate electrode fouling:



Any form of electrode fouling will cause this secondary voltage measurement to rise, thus providing an indicator that instrument technicians may use for predictive maintenance (telling them when the probes need cleaning or replacement). Meanwhile, the primary voltmeter will do its job of accurately measuring liquid conductivity so long as the current source is still able to output its normal amount of current.

22.1.4 Electrodeless conductivity probes

An entirely different design of conductivity cell called *electrodeless* uses electromagnetic induction rather than direct electrical contact to detect the conductivity of the liquid solution. This cell design enjoys the distinct advantage of virtual immunity to fouling[6], since there is no direct electrical contact between the measurement circuit and the liquid solution. Instead of using two or four electrodes inserted into the solution for conductivity measurement, this cell uses two *toroidal* inductors (one to induce an AC voltage in the liquid solution, and the other to measure the strength of the resulting current through the solution):

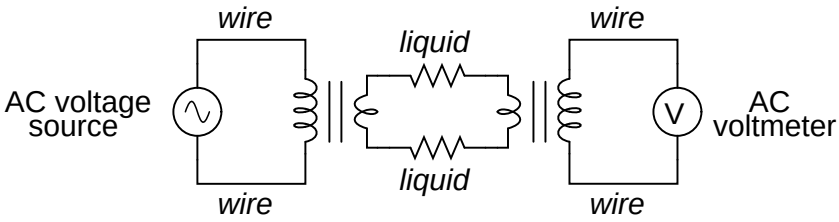


Since toroidal magnetic cores do an excellent job of containing their own magnetic fields, there will be negligible mutual inductance between the two wire coils. The *only* way a voltage will be induced in the secondary coil is if there is an AC current passing through the center of that coil, through the liquid itself. The primary coil is ideally situated to induce such a current in the solution. The more conductive the liquid solution, the more current will pass through the center of both coils (through the liquid), thus producing a greater induced voltage at the secondary coil. Secondary coil voltage therefore is directly proportional to liquid conductivity[7].

[6] Toroidal conductivity sensors may suffer calibration errors if the fouling is so bad that the hole becomes choked off with sludge, but this is an extreme condition. These sensors are far more tolerant to fouling than any form of contact-type (electrode) conductivity cell.

[7] Note that this is opposite the behavior of a direct-contact conductivity cell, which produces *less* voltage as the liquid becomes more conductive.

The equivalent electrical circuit for a toroidal conductivity probe looks like a pair of transformers, with the liquid acting as a resistive path for current to connect the two transformers together:



Toroidal conductivity cells are used whenever possible, due to their ruggedness and virtual immunity to fouling. However, they are not sensitive enough for conductivity measurement in high-purity applications such as boiler feedwater treatment and ultra-pure water treatment necessary for pharmaceutical and semiconductor manufacturing. As always, the manufacturer's specifications are the best source of information for conductivity cell applicability in any particular process.

The following photograph shows a toroidal conductivity probe along with a conductivity transmitter (to both display the conductivity measurement in millisiemens per centimeter and also transmit the measurement as a 4-20 mA analog signal):

22.2 pH measurement

pH is the measurement of the hydrogen ion activity in a liquid solution. It is one of the most common forms of analytical measurement in industry, because pH has a great effect on the outcome of many chemical processes. Food processing, water treatment, pharmaceutical production, steam generation (thermal power plants), and alcohol manufacturing are just some of the industries making extensive use of pH measurement (and control). pH is also a significant factor in the corrosion of metal pipes and vessels carrying aqueous (water-based) solutions, so pH measurement and control is important in the life-extension of these capital investments.

In order to understand pH measurement, you must first understand the chemistry of pH. Please refer to section 3.12 beginning on page 177 for a theoretical introduction to pH.

22.2.1 Colorimetric pH measurement

One of the simplest ways to measure the pH of a solution is by color. Certain specific chemicals dissolved in an aqueous solution will change color if the pH value of that solution falls within a certain range. *Litmus paper* is a common laboratory application of this principle, where a color-changing chemical substance infused on a paper strip changes color when dipped in the solution. Comparing the final color of the litmus paper to a reference chart yields an approximate pH value for the solution.

A natural example of this phenomenon is well-know to flower gardeners, who recognize that hydrangea blossoms change color with the pH value of the soil. In essence, these plants act as organic litmus indicators[8]. This hydrangea plant indicates acidic soil by the violet color of its blossoms:

[8] Truth be told, the color of a hydrangea blossom is only indirectly determined by soil pH. Soil pH affects the plant's uptake of aluminum, which is the direct cause of color change. Interestingly, the pH-color relationship of a hydrangea plant is exactly opposite that of common laboratory litmus paper: red litmus paper indicates an acidic solution while blue litmus paper indicates an alkaline solution; whereas red hydrangea blossoms indicate alkaline soil while blue (or violet) hydrangea blossoms indicate acidic soil.

22.2.2 Potentiometric pH measurement

Color-change is a common pH test method used for manual laboratory analyses, but it is not well-suited to continuous process measurement. By far the most common pH measurement method in use is *electrochemical*: special pH-sensitive electrodes inserted into an aqueous solution will generate a voltage dependent upon the pH value of that solution.

Like all other potentiometric (voltage-based) analytical measurements, electrochemical pH measurement is based on the *Nernst equation*, which describes the electrical potential by ions migrating through a permeable membrane:

$$V = \frac{RT}{nF} \ln \left(\frac{C_1}{C_2} \right)$$

Where,

V = Voltage produced across membrane due to ion exchange, in volts (V)

R = Universal gas constant (8.315 J/mol·K)

T = Absolute temperature, in Kelvin (K)

n = Number of electrons transferred per ion exchanged (unitless)

F = Faraday constant, in coulombs per mole (96,485 C/mol e^-)

C_1 = Concentration of ion in measured solution, in moles per liter of solution (M)

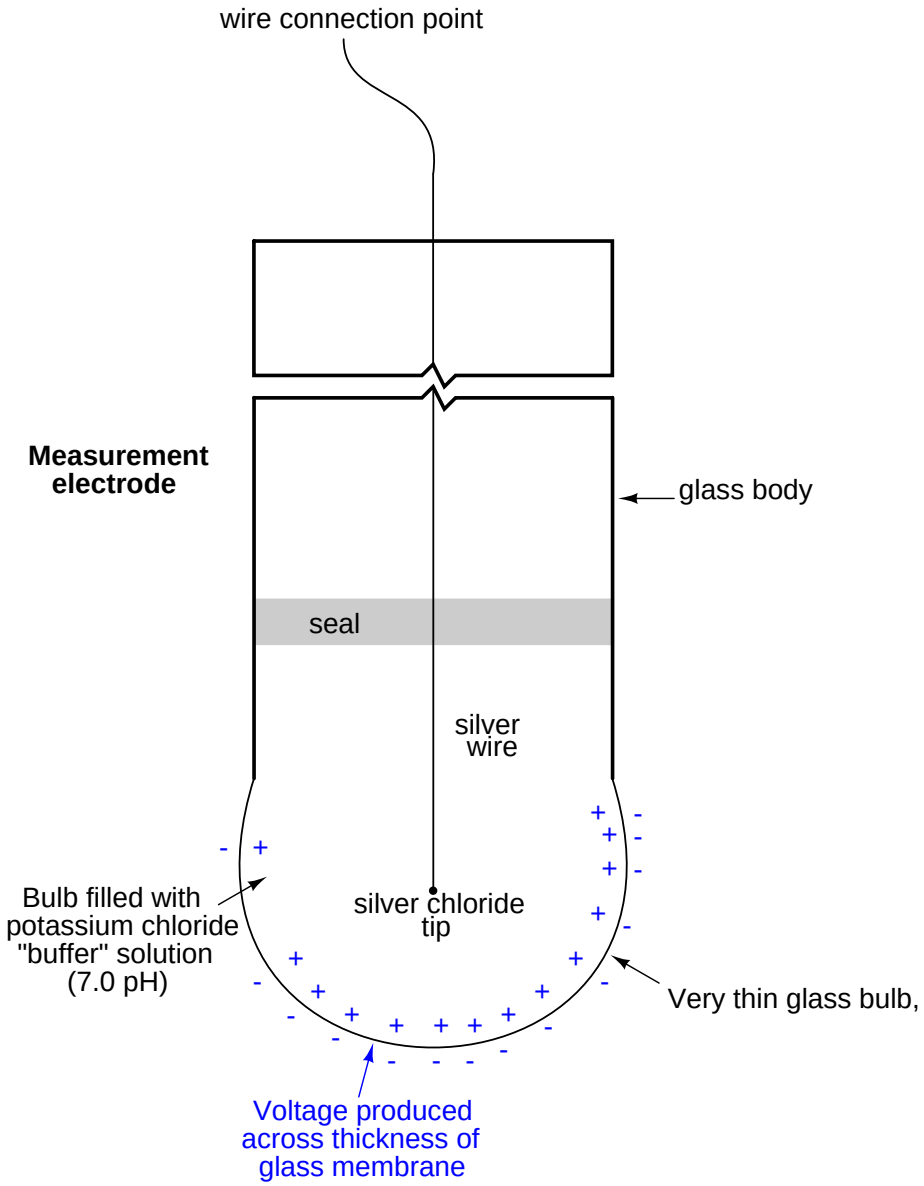
C_2 = Concentration of ion in reference solution (on other side of membrane), in moles per liter of solution (M)

We may also write the Nernst equation using of common logarithms instead of natural logarithms, which is usually how we see it written in the context of pH measurement:

$$V = \frac{2.303RT}{nF} \log \left(\frac{C_1}{C_2} \right)$$

Both forms of the Nernst equation predict a greater voltage developed across the thickness of a membrane as the concentrations on either side of the membrane differ to a greater degree. If the ionic concentration on both sides of the membrane are equal, no Nernst potential will develop.

In the case of pH measurement, the Nernst equation describes the amount of electrical voltage developed across a special *glass* membrane due to hydrogen ion exchange between the process liquid solution and a *buffer solution* inside the bulb formulated to maintain a constant pH value of 7.0 pH. Special pH-measurement electrodes are manufactured with a closed end made of this glass, a small quantity of buffer solution contained within the glass bulb:



Any concentration of hydrogen ions in the process solution differing from the hydrogen ion concentration in the buffer solution ($[H^+] = 1 \times 10^{-7}\ M$) will cause a voltage to develop across the

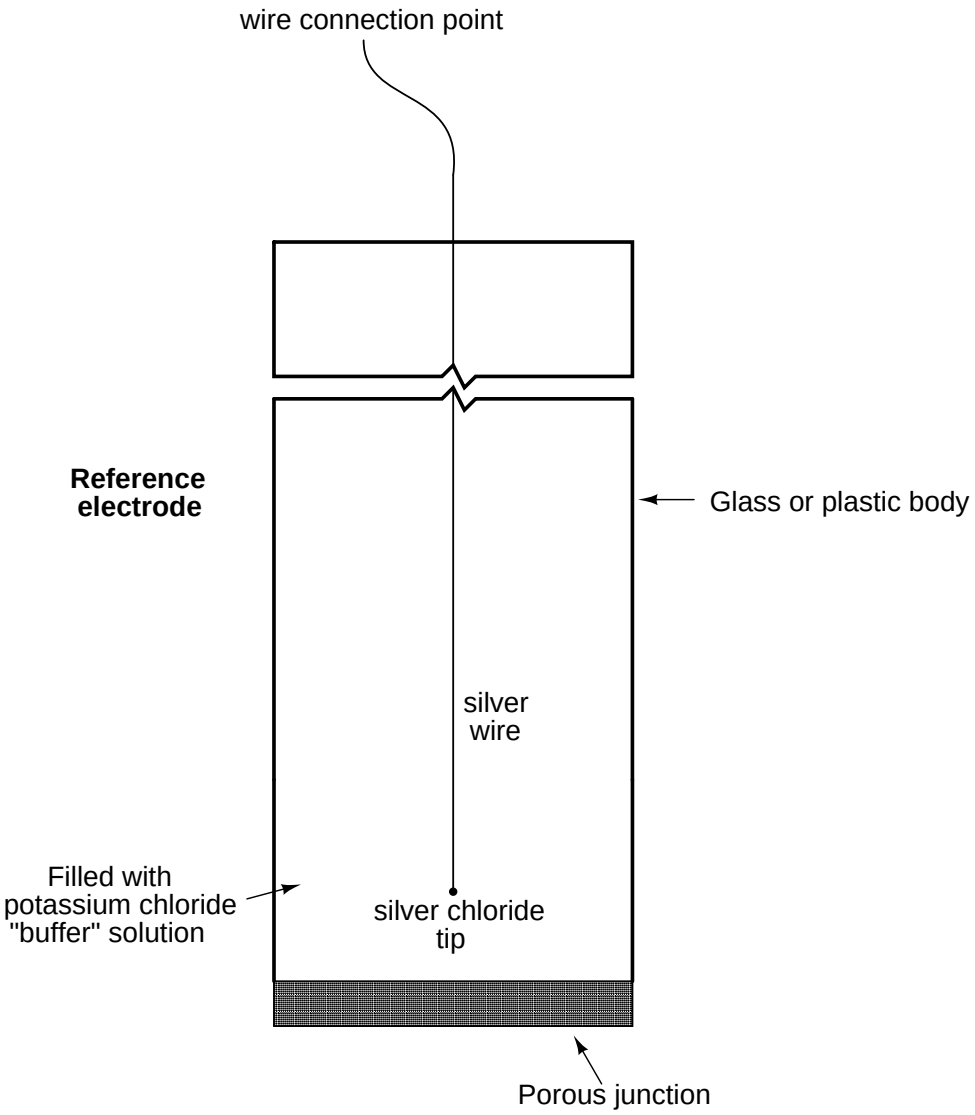
thickness of the glass. Thus, a standard pH measurement electrode produces no potential when the process solution's pH value is exactly 7.0 pH (equal in hydrogen ion activity to the buffer solution trapped within the bulb).

The glass used to manufacture this electrode is no ordinary glass. Rather, it is specially manufactured to be *selectively permeable* to hydrogen ions[9]. If it were not for this fact, the electrode might generate voltage as it contacted any number of different ions in the solution. This would make the electrode non-specific, and therefore useless for pH measurement.

Manufacturing processes for pH-sensitive glass are highly guarded trade secrets. There seems to be something of an art to the manufacture of an accurate, reliable, and long-lived pH electrode. A variety of different measurement electrode designs exist for different process applications, including high pressure and high temperature services.

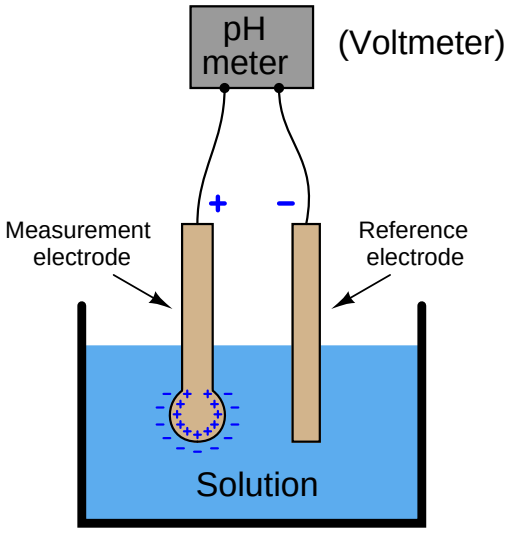
[9] It is a proven fact that sodium ions in relatively high concentration (compared to hydrogen ions) will also cause a Nernst potential across the glass of a pH electrode, as will certain other ion species such as potassium, lithium, and silver. This effect is commonly referred to as *sodium error*, and it is usually only seen at high pH values where the hydrogen ion concentration is extremely low. Like any other analytical technology, pH measurement is subject to "interference" from species unrelated to the substance of interest.

Actually measuring the voltage developed across the thickness of the glass electrode wall, however, presents a bit of a problem: while we have a convenient electrical connection to the solution inside the glass bulb, we do not have any place to connect the other terminal of a sensitive voltmeter to the solution outside the bulb[10]. In order to establish a complete circuit from the glass membrane to the voltmeter, we must create a zero-potential electrical junction with the process solution. To do this, we use another special electrode called a *reference electrode* immersed in the same liquid solution as the measurement electrode:



[10]Remember that voltage is always measured *between two points!*

Together, the measurement and reference electrodes provide a voltage-generating element sensitive to the pH value of whatever solution they are submerged in:



The most common configuration for modern pH probe sets is what is called a *combination electrode*, which combines both the glass measurement electrode and the porous reference electrode in a single unit. This photograph shows a typical industrial combination pH electrode:

The red-colored plastic cap on the right-hand end of this combination electrode covers and

protects a gold-plated coaxial electrical connector, to which the voltage-sensitive pH indicator (or transmitter) attaches.

Another model of pH probe appears in the next photograph. Here, there is no protective plastic cap covering the probe connector, allowing a view of the gold-plated connector bars:

A close-up photograph of the probe tip reveals the glass measurement bulb, a weep hole for process liquid to enter the reference electrode assembly (internal to the white plastic probe body), and a metal *solution ground* electrode:

It is extremely important to always keep the glass electrode wet. Its proper operation depends on complete *hydration* of the glass, which allows hydrogen ions to penetrate the glass and develop the Nernst potential. The probes shown in these photographs are shown in a dry state only because they have already exhausted their useful lives and cannot be damaged any further by dehydration.

The process of hydration – so essential to the working of the glass electrode – is also a mechanism of wear. Layers of glass "slough" off over time if continuously hydrated, which means that glass pH electrodes have a limited life whether they are being used to measure the pH of a process solution (continuously wet) or if they are being stored on a shelf (maintained in a wet state by a small quantity of potassium hydroxide held close to the glass probe by a liquid-tight cap). It is therefore impossible to extend the shelf life of a glass pH electrode indefinitely.

A common installation for industrial pH probe assemblies is to simply dip them into an open vessel containing the solution of interest. This arrangement is very common in water treatment applications, where the water mostly flows in open vessels by gravity at the treatment facility. A photograph showing a pH measurement system for the "outfall" flow of water from an industrial facility appears here:

Water flowing from the discharge pipe of the facility enters an open-top stainless steel tank where the pH probe hangs from a bracket. An overflow pipe maintains a maximum water level in the tank as water continuously enters it from the discharge pipe. The probe assembly may be easily removed for maintenance:

An alternative design for industrial pH probes is the *insertion* style, designed to install in a pressurized pipe. Insertion probes are designed to be removed while the process line remains pressurized, to facilitate maintenance without interrupting continuous operation:

The probe assembly inserts into the process line through the open bore of a 90^o turn ball valve. The left-hand photograph (above) shows the retaining nut loosened, allowing the probe to slide up and out of the pipe. The right-hand photograph shows the ball valve shut to block process liquid pressure from escaping, while the technician unlatches the clamps securing the probe to the pipe fitting.

Once the clamp is unlatched, the probe assembly may be completely detached from the pipe, allowing cleaning, inspection, calibration, repair, and/or replacement:

The voltage produced by the measurement electrode (glass membrane) is quite modest. A calculation for voltage produced by a measurement electrode immersed in a 6.0 pH solution shows this. First, we must calculate hydrogen ion concentration (activity) for a 6.0 pH solution, based on the definition of pH being the negative logarithm of hydrogen ion molarity:

$$\text{pH} = -\log[\text{H}^+]$$

$$6.0 = -\log[\text{H}^+]$$

$$-6.0 = \log[\text{H}^+]$$

$$10^{-6.0} = 10^{\log[\text{H}^+]}$$

$$10^{-6.0} = \text{H}^+$$

$$\text{H}^+ = 1 \times 10^{-6}\ M$$

This tells us the concentration of hydrogen ions in the 6.0 pH solution (hydrogen ion *concentration* being practically the same as hydrogen ion *activity* for dilute solutions). We know that the buffer solution inside the glass measurement bulb has a stable value of 7.0 pH (hydrogen ion concentration of $1 \times 10^{-7}\ M$, or 0.0000001 moles per liter), so all we need to do now is plug these values in to the Nernst equation to see how much voltage the glass electrode should generate. Assuming a solution temperature of 25^o C (298.15 K), and knowing that n in the Nernst equation will be equal to 1 (since each hydrogen ion has a single-value electrical charge):

$$V = \frac{2.303RT}{nF} \log\left(\frac{C_1}{C_2}\right)$$

$$V = \frac{(2.303)(8.315)(298.15)}{(1)(96485)} \log\left(\frac{1 \times 10^{-6}\ M}{1 \times 10^{-7}\ M}\right)$$

$$V = (59.17\ \text{mV})(\log 10) = 59.17\ \text{mV}$$

If the measured solution had a value of 7.0 pH instead of 6.0 pH, there would be no voltage generated across the glass membrane since the two solutions' hydrogen ion activities would be equal. Having a solution with one decade (ten times more: exactly one "order of magnitude") greater hydrogen ions activity than the internal buffer solution produces 59.17 millivolts at 25 degrees Celsius. If the pH were to drop to 5.0 (two units away from 7.0 instead of one unit), the output voltage would be double: 118.3 millivolts. If the solution's pH value were more alkaline than the internal buffer (for example, 8.0 pH), the voltage generated at the glass bulb would be the opposite polarity (e.g. 8.0 pH = -59.17 mV ; 9.0 pH = -118.3 mV, etc.).

The following table shows the relationship between hydrogen ion activity, pH value, and probe voltage[11]:

Hydrogen ion activity	pH value	Probe voltage (at 25^o C)
$1 \times 10^{-3}\ M = 0.001\ M$	3.0 pH	236.7 mV
$1 \times 10^{-4}\ M = 0.0001\ M$	4.0 pH	177.5 mV
$1 \times 10^{-5}\ M = 0.00001\ M$	5.0 pH	118.3 mV
$1 \times 10^{-6}\ M = 0.000001\ M$	6.0 pH	59.17 mV
$1 \times 10^{-7}\ M = 0.0000001\ M$	7.0 pH	0 mV
$1 \times 10^{-8}\ M = 0.00000001\ M$	8.0 pH	-59.17 mV
$1 \times 10^{-9}\ M = 0.000000001\ M$	9.0 pH	-118.3 mV
$1 \times 10^{-10}\ M = 0.0000000001\ M$	10.0 pH	-177.5 mV
$1 \times 10^{-11}\ M = 0.00000000001\ M$	11.0 pH	-236.7 mV

This numerical progression is reminiscent of the *Richter scale* used to measure earthquake magnitudes, where each ten-fold (decade) multiplication of power is represented by one more increment on the scale (e.g. a 6.0 Richter earthquake is ten times more powerful than a 5.0 Richter earthquake). The logarithmic nature of the Nernst equation means that pH probes – and in fact all potentiometric sensors based on the same dynamic of voltage produced by ion exchange across a membrane – have astounding rangeability: they are capable of representing a wide range of conditions with a modest signal voltage span.

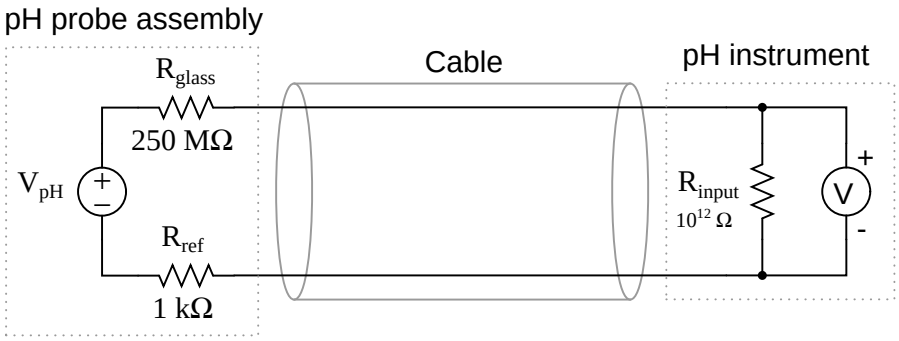
Of course, the disadvantage of high rangeability is the potential for large pH measurement errors if the voltage detection within the pH instrument is even just a little bit inaccurate. The problem is made even worse by the fact that the voltage measurement circuit has an extremely high impedance due to the presence of the *glass* membrane[12]. The pH instrument measuring the voltage produced by a pH probe assembly must have an input impedance that is orders of magnitude greater yet, or else the probe's voltage signal will become "loaded down" by the voltmeter and not register accurately.

[11] The mathematical sign of probe voltage is arbitrary. It depends entirely on whether we consider the reference (buffer) solution's hydrogen ion activity to be C_1 or C_2 in the equation. Which ever way we choose to calculate this voltage, though, the polarity will be opposite for acidic pH values as compared to alkaline pH values

[12] Glass is a very good insulator of electricity. With a thin layer of glass being an essential part of the sensor circuit, the typical impedance of that circuit will lie in the range of *hundreds* of mega-ohms!

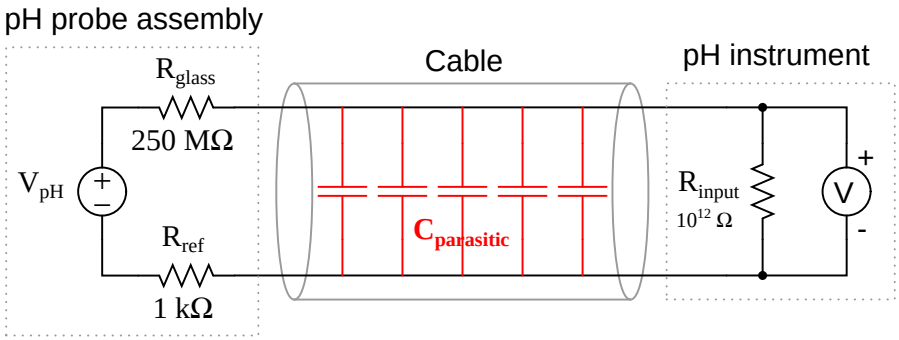
Fortunately, modern operational amplifier circuits with field-effect transistor input stages are sufficient for this task[13]:

Equivalent electrical circuit of a pH probe and instrument



The voltage sensed by the pH instrument very nearly equals V_{pH} because $(R_{glass} + R_{ref}) << R_{input}$

Even if we use a high-input-impedance pH instrument to sense the voltage output by the pH probe assembly, we may still encounter a problem created by the impedance of the glass electrode: an RC time constant created by the parasitic capacitance of the probe cable connecting the electrodes to the sensing instrument. The longer this cable is, the worse the problem becomes due to increased capacitance:



[13] Operational amplifier circuits with field-effect transistor inputs may easily achieve input impedances in the *tera-ohm* range (1×10^{12} Ω).

This time constant value may be significant if the cable is long and/or the probe resistance is abnormally large. Assuming a combined (measurement and reference) electrode resistance of 700 MΩ and a 30 foot length of RG-58U coaxial cable (at 28.5 pF capacitance per foot), the time constant will be:

$$\tau = RC$$

$$\tau = (700 \times 10^6\ \Omega)\left((28.5 \times 10^{-12}\ \text{F/ft})(30\ \text{ft})\right)$$

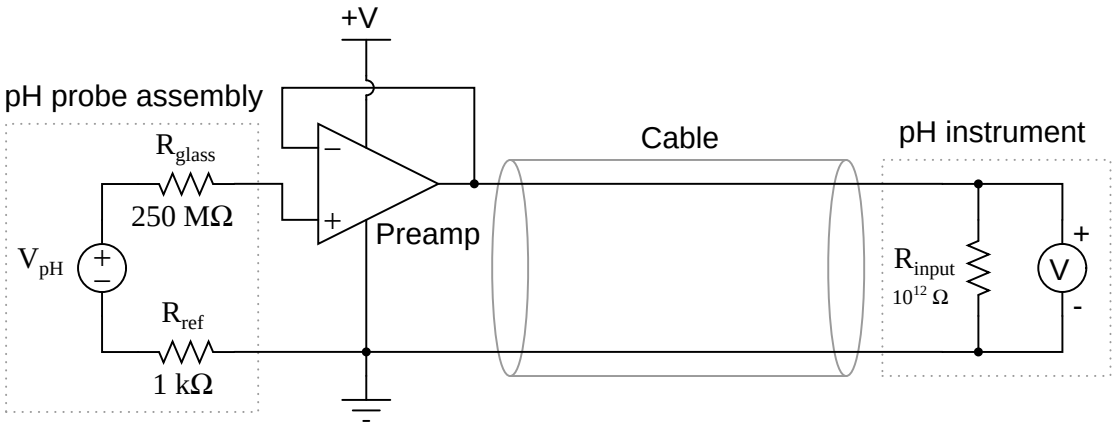
$$\tau = (700 \times 10^6\ \Omega)(8.55 \times 10^{-10}\ \text{F})$$

$$\tau = 0.599 \text{ seconds}$$

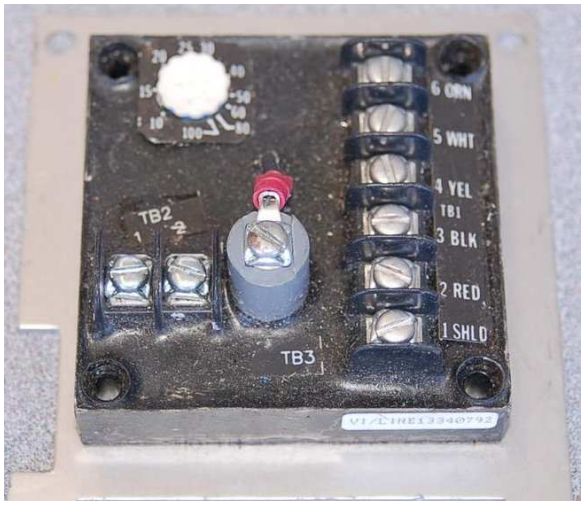
Considering the simple approximation of 5 time constants being the time necessary for a first-order system such as this to achieve within 1% of its final value after a step-change, this means a sudden change in voltage at the pH probe caused by a sudden change in pH will not be fully registered by the pH instrument until almost 3 seconds after the event has passed!

It may seem impossible for a cable with capacitance measured in *pico*farads to generate a time constant easily within the range of human perception, but it is indeed reasonable when you consider the exceptionally large resistance value of a glass pH measurement electrode. For this reason, and also for the purpose of limiting the reception of external electrical "noise," it is best to keep the cable length between pH probe and instrument as short as possible.

When short cable lengths are simply not practical, a *preamplifier* module may be connected between the pH probe assembly and the pH instrument. Such a device is essentially a unity-gain (gain = 1) amplifier designed to "repeat" the weak voltage output of the pH probe assembly in a much stronger (i.e. lower-impedance) form so the effects of cable capacitance will not be as severe. A unity-gain operational amplifier "voltage buffer" circuit illustrates the concept of a preamplifier:



A preamplifier module appears in this next photograph:



The preamplifier does not boost the probes' voltage output at all. Rather, it serves to decrease the impedance (the Thévenin equivalent resistance) of the probes by providing a low-resistance (relatively high-current capacity) voltage output to drive the cable and pH instrument. By providing a voltage gain of 1, and a very large current gain, the preamplifier practically eliminates RC time constant problems caused by cable capacitance, and also helps reduce the effect of induced electrical noise. As a consequence, the practical cable length limit is extended by orders of magnitude.

Referring back to the Nernst equation, we see that temperature plays a role in determining the amount of voltage generated by the glass electrode membrane. The calculations we performed earlier predicting the amount of voltage produced by different solution pH values all assumed the same temperature: 25 degrees Celsius (298.15 Kelvin). If the solution is not at room temperature, however, the voltage output by the pH probe will not be 59.17 millivolts per pH unit. For example, if a glass measurement electrode is immersed in a solution having a pH value of 6.0 pH at 70 degrees Celsius (343.15 Kelvin), the voltage generated by that glass membrane will be 68.11 mV rather than 59.17 mV as it would be at 25 degrees Celsius. That is to say, the *slope* of the pH-to-voltage function will be 68.11 millivolts per pH unit rather than 59.17 millivolts per pH unit as it was at room temperature.

The portion of the Nernst equation to the left of the logarithm function defines this slope value:

$$\text{Slope} = \frac{2.303RT}{nF}$$

Recall that R and F are fundamental constants, and n is fixed at a value of 1 for pH measurement (since there is exactly one electron exchanged for every H^+ ion migrating through the membrane). This leaves temperature (T) as the only variable capable of influencing the theoretical slope of the function.

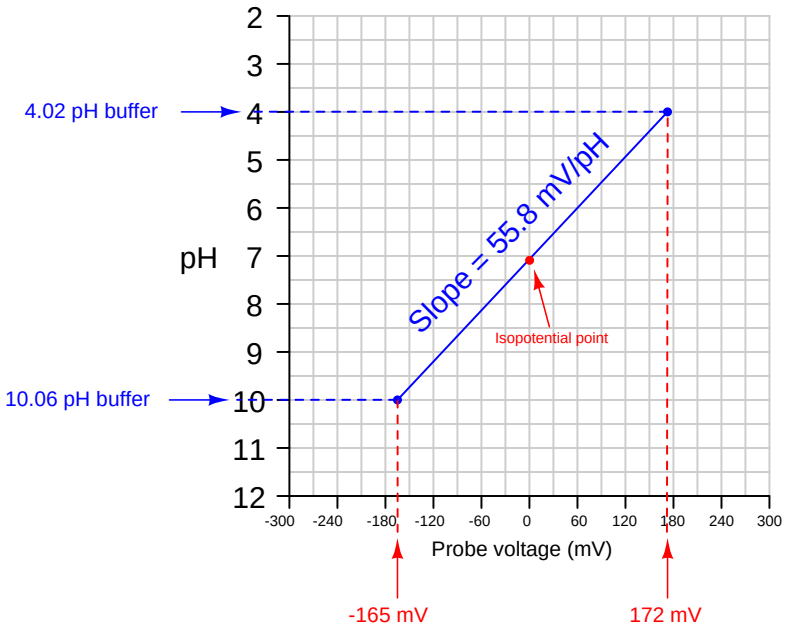
In order for a pH instrument to accurately infer a solution's pH value from the voltage generated by a glass electrode, it must "know" the expected slope of the Nernst equation. Since the only variable in the Nernst equation beside the two ion concentration values (C_1 and C_2) is temperature

(T), a simple temperature measurement will provide the pH instrument the information it needs to function accurately. For this reason, many pH instruments are constructed to accept an RTD input for solution temperature sensing, and many pH probe assemblies have built-in RTD temperature sensors ready to sense solution temperature.

While the theoretical slope for a pH instrument depends on no variable but temperature, the *actual* slope also depends on the condition of the measurement electrode. For this reason, pH instruments need to be calibrated for the probes they connect to.

A pH instrument is generally calibrated by performing a two-point test using *buffer solutions* as the pH calibration standard. A buffer solution is a specially formulated solution that maintains a stable pH value even under conditions of slight contamination. For more information on pH buffer solutions, see section 17.9.5 on page 762. The pH probe assembly is inserted into a cup containing a buffer solution of known pH value, then the pH instrument is "standardized" to that pH value[14]. After standardizing at the first calibration point, the pH probe is removed from the buffer, rinsed, then placed into another cup containing a second buffer with a different pH value. After another stabilization period, the pH instrument is standardized to this second pH value.

It only takes two points to define a line, so these two buffer measurements are all that is required by a pH instrument to define the linear transfer function relating probe voltage to solution pH:



Most modern pH instruments will display the calculated slope value after calibration. This value should (ideally) be 59.17 millivolts per pH unit at 25 degrees Celsius, but it will likely be a bit less than this. The voltage-generating ability of a glass electrode decays with age, so a low slope value may indicate a probe in need of replacement.

[14] With all modern pH instruments being digital in design, this standardization process usually entails pressing a pushbutton on the faceplate of the instrument to "tell" it that the probe is stabilized in the buffer solution.

Another informative feature of the voltage/pH transfer function graph is the location of the *isopotential* point: that point on the graph corresponding to zero probe voltage. In theory, this point should correspond to a pH value of 7.0 pH. However, if there exist stray potentials in the pH measurement circuit – for example, voltage differences caused by ion mobility problems in the porous junction of the reference electrode – this point will be shifted. Sufficient contamination of the buffer solution inside the measurement electrode (enough to drive its pH value from 7.0) will also cause an isopotential point shift, since the Nernst equation predicts zero voltage when ion concentrations on both sides of the membrane are equal.

A quick way to check the isopotential point of any calibrated pH instrument is to short the input terminals together (forcing V_{input} to be equal to 0 millivolts) and note the pH indication on the instrument's display[15]. This test should be performed *after* standardizing the instrument using accurate pH buffer solutions.

When calibrating a pH instrument, you should choose buffers that most closely "bracket" the expected range of pH measurement in the process. The most common buffer pH values are 4, 7, and 10 (nominal). For example, if you expect to measure pH values in the process ranging between 7.5 and 9, for example, you should calibrate that pH instrument using 7 and 10 buffers.

[15] A more obvious test would be to directly measure the pH probe assembly's voltage while immersed in 7.0 pH buffer solution. However, most portable voltmeters lack sufficient input impedance to perform this measurement, and so it is easier to standardize the pH instrument in 7.0 pH buffer and then check *its* zero-voltage pH value to see where the isopotential point is at.

22.3 Chromatography

Imagine a major marathon race, where hundreds of runners gather in one place to compete. When the starting gun is fired, all the runners begin running the race, starting from the same location (the starting line) at the same time. As the race progresses, the faster runners distance themselves from the slower runners, resulting in a dispersion of runners along the race course over time.

Now imagine a marathon race where certain runners share the exact same running speeds. Suppose a group of runners in this marathon all run at exactly 8 miles per hour (MPH), while another group of runners in the race run at exactly 6 miles per hour, and another group of runners plod along at exactly 5 miles per hour. What would happen to these three groups of runners over time, supposing they all begin the race at the same location and at the exact same time?

As you can probably imagine, the runners within each speed group will stay with each other throughout the race, with the three groups becoming further spread apart over time. The first of these three groups to cross the finish line will be the 8 MPH runners, followed by the 6 MPH runners a bit later, and then followed by the 5 MPH runners after that. To an observer at the very start of the race, it would be difficult to tell exactly how many 6 MPH runners there were in the crowd, but to an observer at the finish line with a stop watch, it would be very easy to tell how many 6 MPH runners competed in the race (by counting how many runners crossed the finish line at the exact time corresponding to a speed of 6 MPH).

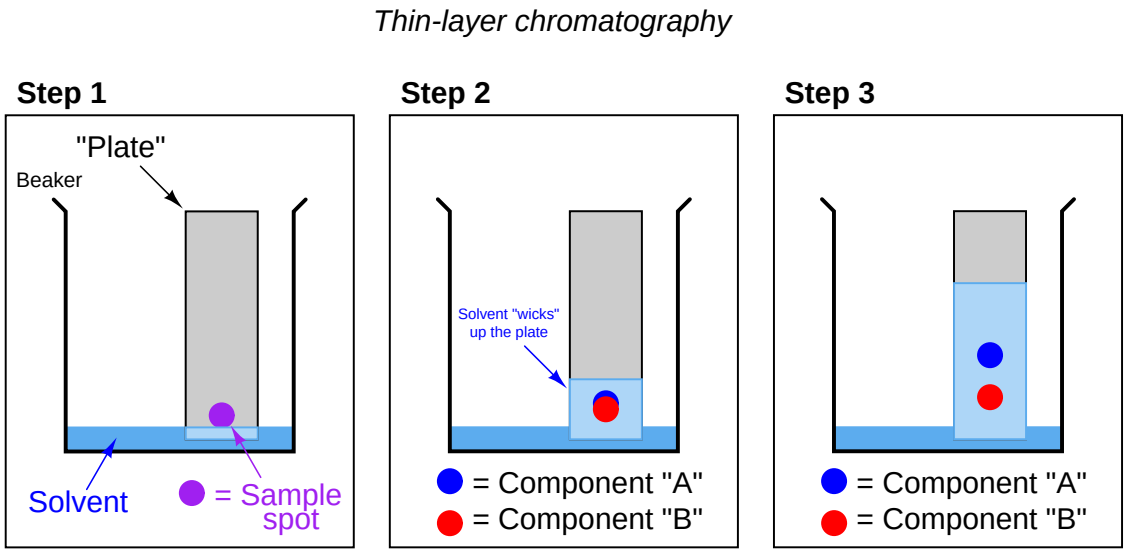
Now imagine a mixture of chemicals in a fluid state traveling through a very small-diameter "capillary" tube filled with an inert, porous material such as sand. Some of those fluid molecules will find it easier to progress down the length of the tube than others, with similar molecules sharing similar propagation speeds. Thus, a small sample of that chemical mixture injected into such a capillary tube, and carried along the tube by a continuous flow of solvent (gas or liquid), will tend to separate into its constituent components over time just like the crowd of marathon runners separate over time according to running speed. Slower-moving molecules will experience greater *retention time* inside the capillary tube, while faster-moving molecules experience less. A detector placed at the outlet of the capillary tube, configured to detect any chemical different from the solvent, will indicate the different components exiting the tube at different times. If the retention time of each chemical component is known from prior tests, this device may be used to identify the composition of the original chemical mix (and even how much of each component was present in the injected sample).

This is the essence of *chromatography*: the technique of chemical separation by time-delayed travel down the length of a stationary medium (called a *column*). In chromatography, the chemical solution traveling down the column is called the *mobile phase*, while the solid and/or liquid substance residing within the column is called the *stationary phase*. Chromatography was first applied to chemical analysis by a Russian botanist named Tswett, who was interested in separating mixtures of plant pigments. The colorful bands left behind in the stationary phase by the separated pigments gave rise to the name "chromatography," which literally means "color writing."

Modern chemists often apply chromatographic techniques in the laboratory to purify chemical samples, and/or to measure the concentrations of different chemical substances within mixtures. Some of these techniques are manual (such as in the case of *thin-layer chromatography*, where liquid solvents carry liquid chemical components along a flat plate covered with an inert coating such as alumina, and the positions of the chemical drops after time distinguishes one component from another). Other techniques are automated, with machines called *chromatographs* performing the

timed analysis of chemical travel through tightly-packed tubular columns.

An illustrated sequence showing thin-layer chromatography appears here:

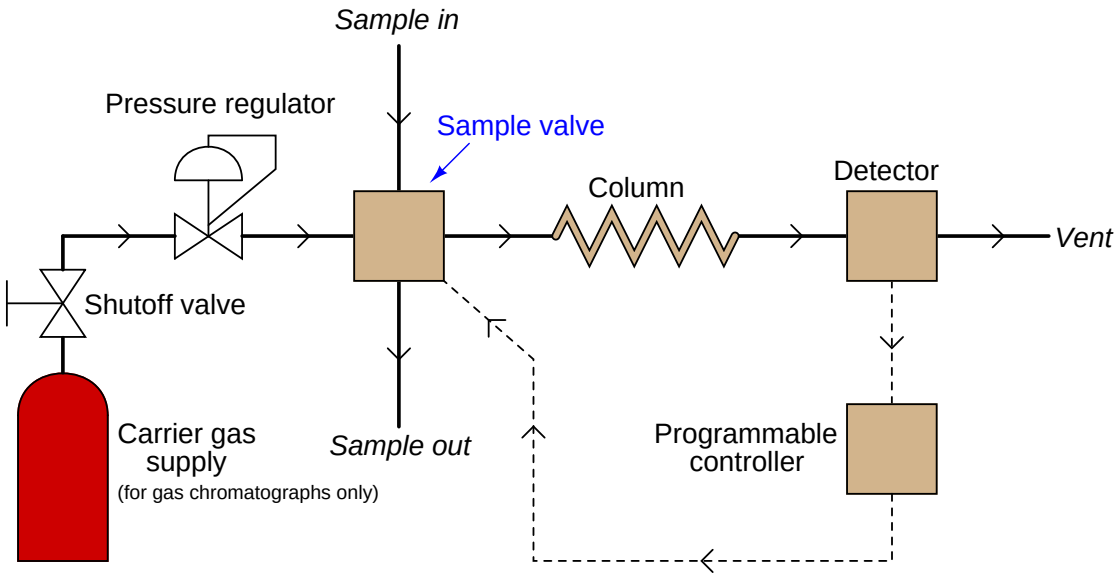


As solvent wicks up the surface of the plate, it carries along with it all components of the sample spot. Each component travels at a different speed, separating the components along the plate over time.

The simplest forms of chromatography reveal the chemical composition of the analyzed mixture as residue retained by the stationary phase. In the case of thin-layer chromatography, the different liquid components of the mobile phase remain embedded in the stationary phase at distinct locations after sufficient "developing" time. The same is true in *paper-strip chromatography* where a simple strip of filter paper serves as the stationary phase through which the mobile phase (liquid sample and solvent) travels: the different components of the sample remain in the paper as residue, their relative positions along the paper's length indicating their extent of travel during the test period. If the components have different colors, the result will be a stratified pattern of colors on the paper strip[16].

[16] This effect is particularly striking when paper-strip chromatography is used to analyze the composition of *ink*. It is really quite amazing to see how many different colors are contained in plain "black" ink!

Most chromatography techniques, however, allow the sample to completely wash through a packed column, relying on a detector at the end of the column to indicate when each component has exited the column. A simplified schematic of a process gas chromatograph (GC) shows how this type of analyzer functions:



The *sample valve* periodically injects a very precise quantity of sample into the entrance of the column tube and then shuts off to allow the constant-flow *carrier* gas to wash this sample through the length of the column tube. Each component of the sample travels through the column at different rates, exiting the column at different times. All the *detector* needs to do is be able to tell the difference between pure carrier gas and carrier gas mixed with anything else (components of the sample).

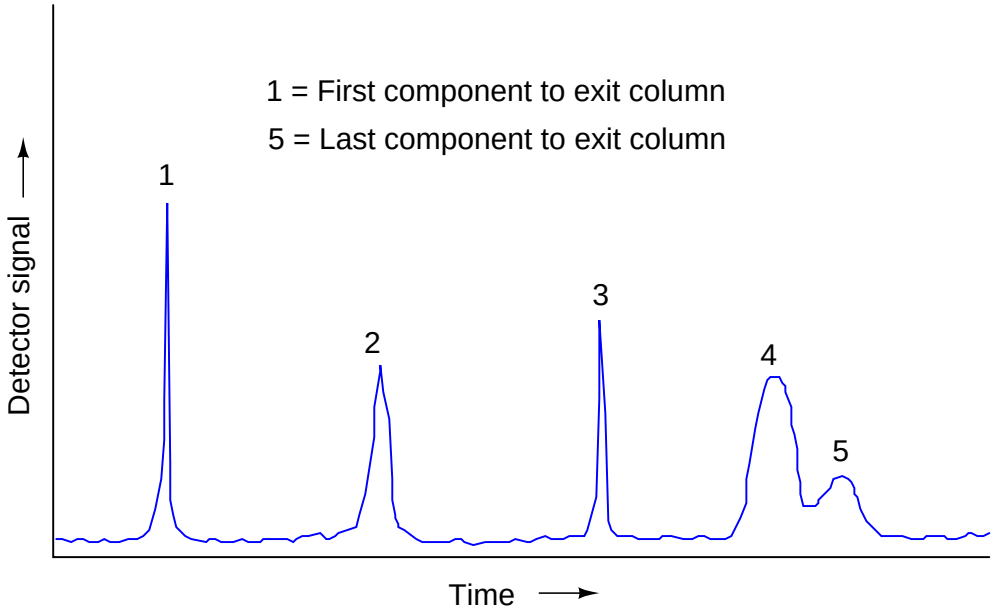
Several different detector designs exist for process gas chromatographs. The two most common are the *flame ionization detector* (FID) and the *thermal conductivity detector* (TCD). Other detector types include the *flame photometric detector* (FPD), Nitrogen-Phosphorus Detector (NPD), and *electron capture detector* (ECD). All chromatograph detectors exploit some physical difference between the solutes (sample components dissolved within the carrier gas) and the carrier gas itself which acts as a gaseous solvent, so that the detector may be able to tell the difference between pure carrier and carrier mixed with solute.

Flame ionization detectors work on the principle of ions liberated in the combustion of the sample components. A permanent flame (usually fueled by hydrogen gas which produces negligible ions in combustion) serves to ionize any gas molecules exiting the chromatograph column that are not carrier gas. Common carrier gases used with FID sensors are helium and nitrogen. Gas molecules containing carbon easily ionize during combustion, which makes the FID sensor well-suited for GC analysis in the petrochemical industries, where hydrocarbon content analysis is the most common form of analytical measurement[17].

[17] In fact, FID sensors are sometimes referred to as *carbon counters*, since their response is almost directly

Thermal conductivity detectors work on the principle of heat transfer by convection (gas cooling). Recall the dependence of a thermal mass flowmeter's calibration on the specific heat value of the gas being measured[18]. This dependence upon specific heat meant that we needed to know the specific heat value of the gas whose flow we intend to measure, or else the flowmeter's calibration would be in jeopardy. Here, in the context of chromatograph detectors, we exploit the impact specific heat value has on thermal convection, using this principle to detect compositional change for a constant-flow gas rate. The temperature change of a heated RTD or thermistor caused by exposure to a gas mixture with changing specific heat value indicates when a new sample component exits the chromatograph column.

If we plot the response of the detector on a graph, we see a pattern of peaks, each one indicating the departure of a component "group" exiting the column. This graph is typically called a *chromatogram*:



Narrow peaks represent compact bunches of molecules all exiting the column at nearly the same time. Wide peaks represent more diffuse groupings of similar (or identical) molecules. In this chromatogram, you can see that components 4 and 5 are not clearly differentiated over time. Better separation of components may be achieved by altering the sample volume, carrier gas flow rate, carrier gas pressure, type of carrier gas, column packing material, and/or column temperature.

Changes in column temperature (called *temperature programming*) are very commonly used to alter the retention times of different components during an analysis cycle, working on the principle

proportional to the number of carbon atoms passing through the flame.

[18] See section 21.7.2, on page 1116. The greater the specific heat value of a gas, the more heat energy it can carry away from a hot object through convection, all other factors being equal.

of a fluid's viscosity being dependent on temperature[19]. Since the flow regime of the mobile phase through a chromatograph column is definitely laminar (not turbulent), fluid viscosity plays a large role in determining flow rate.

If the relative propagation speeds of each component is known in advance, the chromatogram peaks may be used to identify the presence (and quantities of) those components. The quantity of each component present in the original sample may be determined by applying the calculus technique of *integration* to each chromatogram peak, calculating the area underneath each curve. The vertical axis represents detector signal, which is proportional to component concentration[20] which is proportional to flow rate given a fixed carrier flow rate. This means the height of each peak represents mass flow rate of each component (W, in units of micrograms per minute, or some similar units). The horizontal axis represents time, so therefore the integral (sum of infinitesimal products) of the detector signal over the time interval for any specific peak (time t_1 to t_2) represents a mass quantity that has passed through the column. In simplified terms, a mass flow rate (micrograms per minute) multiplied by a time interval (minutes) equals mass in micrograms:

$$m = \int_{t_1}^{t_2} W \, dt$$

Where,

m = Mass of sample component in micrograms

W = Instantaneous mass flow rate of sample component in micrograms per minute

t = Time in minutes (t_1 and t_2 are the interval times between which total mass is calculated)

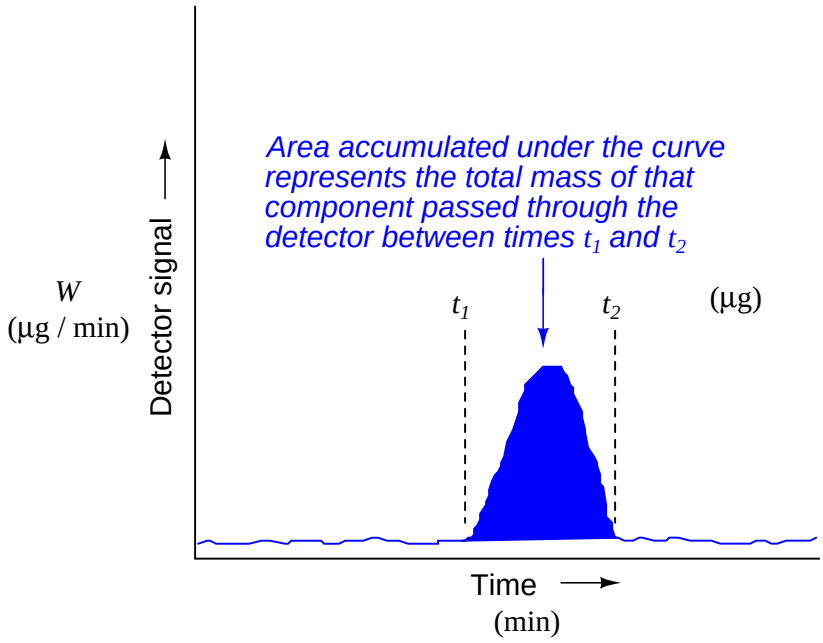
As is the case with all examples of integration, the unit of measurement for the totalized result is the *product* of the units within the integrand: flow rate (W) in units of micrograms per minute multiplied by increments of time (dt) in the unit of minutes, summed together over an interval ($\int_{t_1}^{t_2}$), result in a mass quantity (m) expressed in the unit of micrograms. Integration is really nothing more than the sum of products, with dimensional analysis working as it does with any product of two physical quantities:

$$\left(\frac{[\mu g]}{[\text{min}]} \right) [\text{min}] = [\mu g]$$

[19] Liquids generally become less viscous when heated, and gases become more viscous when heated. Therefore, raising column temperature in a gas chromatograph will "slow down" the later components (raise retention time) to achieve better separation.

[20] Detector response also varies substantially with the type of substance being detected, and not just its concentration. A flame ionization detector (FID), for instance, yields different responses for a given mass flow rate of butane (C_4H_{10}) than it does for the same mass flow rate of methane (CH_4), due to the differing carbon count per mass ratios of the two compounds. This means the same raw signal from an FID sensor generated by a concentration of butane versus a concentration of methane actually represents different concentrations of butane versus methane in the carrier. The inconsistent response of a chromatograph detector to different sampled components is not as troubling a problem as one might think, though. Since the chromatograph column does a good job separating each component from the other over time, we may program the computer to re-calibrate itself for each component at the specific time(s) each component is expected to exit the column. So long as we know in advance the characteristic detector response for each expected compound separated by the chromatograph, we may easily compensate for those variations in real time so the chromatogram consistently and accurately represents component concentrations over the entire analysis cycle.

This mathematical relationship may be seen in graphical form by shading the area underneath the peak of a chromatogram:



Since process chromatographs have the ability to independently analyze the quantities of multiple components in a chemical sample, these instruments are inherently *multi-variable*. A single analog output signal (e.g. 4-20 mA) would only be able to transmit information about the concentration of any one component (any one peak) in the chromatogram. This is perfectly adequate if only one component concentration is worth knowing about in the process[21], but some form of multi-channel digital (or multiple analog outputs) transmission is necessary to make full use of a chromatograph's ability.

All modern chromatographs are "smart" instruments, containing one or more digital computers which execute the calculations necessary to derive precise measurements from chromatogram data. The computational power of modern chromatographs may be used to further analyze the process sample, beyond simple determinations of concentration or quantity. Examples of more abstract analyses include approximate octane value of gasoline (based on the relative concentrations of several components), or the heating value of natural gas (based on the relative concentrations of methane, ethane, propane, butane, carbon dioxide, helium, etc. in a sample of natural gas).

[21] It is not uncommon to find chromatographs used in processes to measure the concentration of a single chemical component, even though the device is capable of measuring the concentrations of multiple components in that process stream. In those cases, chromatography is (or was at the time of installation) the most practical analytical technique to use for quantitative detection of that substance. Why else use an inherently multi-variable analyzer when you could have used a single-variable technology that was simpler? By analogy, it is possible to use a Coriolis flowmeter to measure nothing but fluid density, even though such a device is fully capable of measuring fluid density *and* mass flow rate *and* temperature.

The following photograph shows a gas chromatograph (GC) fulfilling precisely this purpose – the determination of heating value for natural gas[22]:

This particular GC is used by a natural gas distribution company as part of its pricing system. The heating value of the natural gas is used as data to calculate the selling price of the natural gas (dollars per standard cubic foot), so the customers pay only for the actual benefit of the gas (i.e. its ability to function as a fuel) and not just volumetric or mass quantity. No chromatograph can directly measure the heating value of natural gas, but the analytical process of chromatography can determine the relative concentrations of compounds within the natural gas. A computer, taking those concentration measurements and multiplying each one by the respective heating value of each compound, derives the gross heating value of the natural gas.

Although the column cannot be seen in the photograph of the GC, several high-pressure steel "bottles" may be seen in the background holding carrier gas used to wash the natural gas sample through the column.

[22] Since the heat of combustion is well-known for various components of natural gas (methane, ethane, propane, etc.), all the chromatograph computer needs to do is multiply the different heat values by their respective concentrations in the gas flowstream, then average the total heat value per unit volume (or mass) of natural gas.

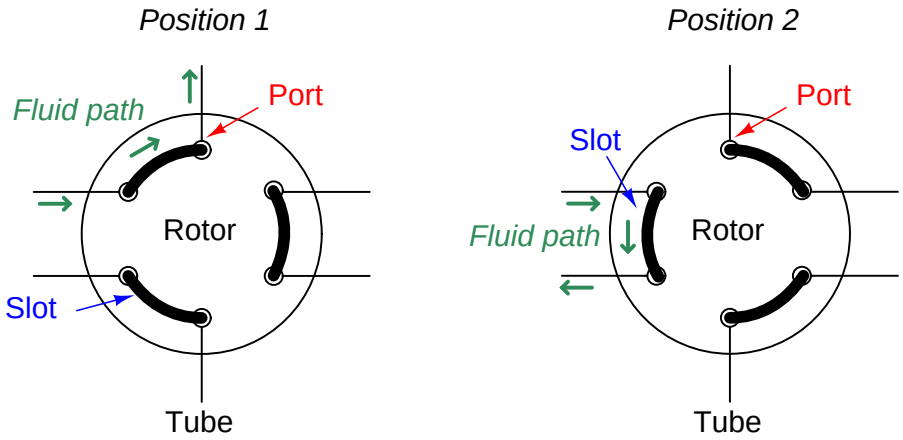
A typical gas chromatograph column appears in the next photograph. It is nothing more than a stainless-steel tube packed with an inert, porous filling material:

This particular GC column is 28 feet long, with an *outside* diameter of only 1/8 inch (the tube's inside diameter is even less than that). Column geometry and packing material vary greatly with application. The many choices intrinsic to column design are best left to specialists in the field of chromatography, not the average technician or even the average process engineer.

Arguably, the most important component of a process gas chromatograph is the sample valve. Its purpose is to inject the exact same sample quantity into the column at the beginning of each cycle. If the sample quantity is not repeatable, the measured quantities exiting the column will change from cycle to cycle even if the sample composition does not change. If the valve's cycle time is not repeatable, component separation efficiency will vary from cycle to cycle. If the sample valve leaks such that a small flow rate of sample continuously enters the column, the result will be an

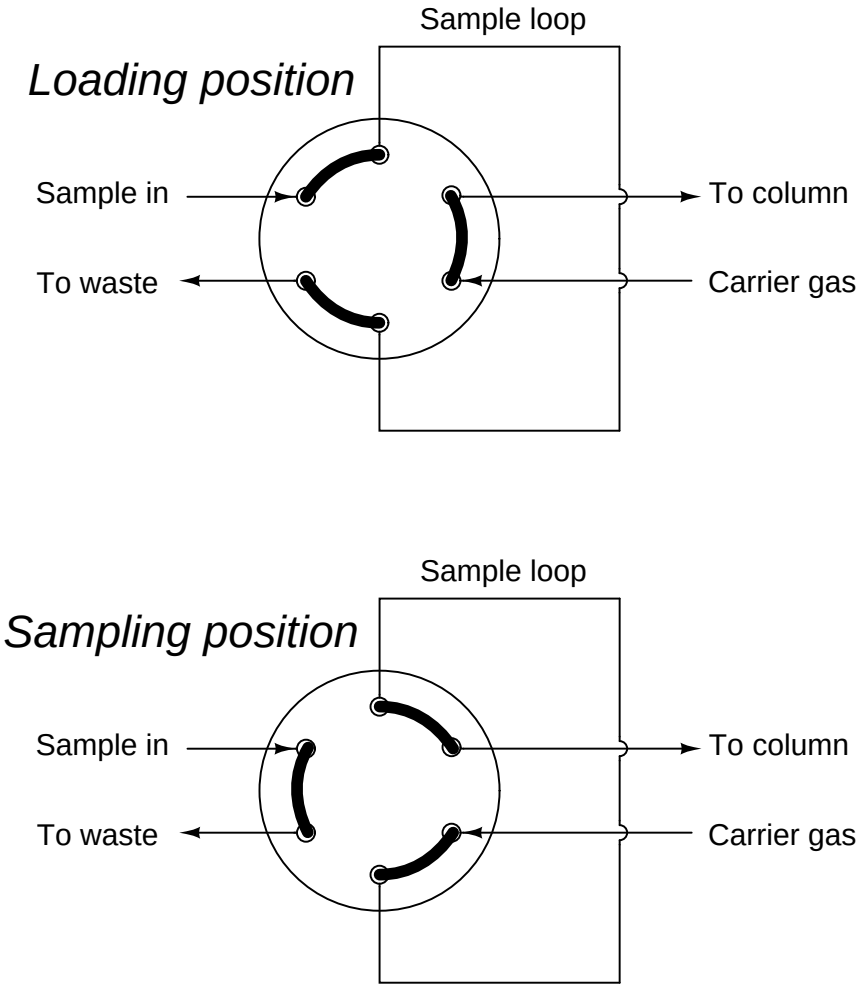
altered "baseline" signal at the detector (at best) and total corruption of the analysis (at worst). Many process chromatograph problems are caused by irregularities in the sample valve(s).

A common form of sample valve uses a rotating element to switch port connections between the sample gas stream, carrier gas stream, and column:



Three slots connect three pairs of ports together. When the rotary valve actuates, the port connections switch, redirecting gas flows.

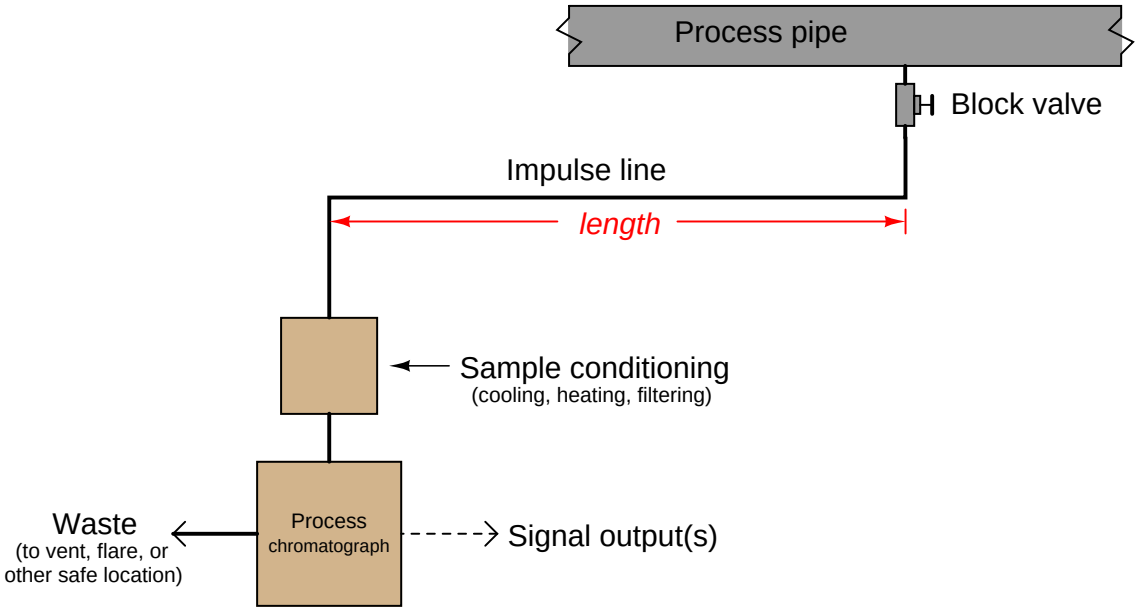
Connected to a sample stream, carrier stream, and column, the rotary sample valve operates in two different modes. The first mode is a "loading" position where the sample stream flows through a short length of tubing (called a *sample loop*) and exits to a waste discharge port, while the carrier gas flows through the column to wash the last sample through. The second mode is a "sampling" position where the volume of sample gas held in the sample loop tubing gets injected into the column by a flow of carrier gas behind it:



The purpose of the sample loop tube is to act as a holding reservoir for a fixed volume of sample gas. When the sample valve switches to the sample position, the carrier gas will flush the contents of the sample loop into the front of the column. This valve configuration guarantees that the injected sample volume does not vary with inevitable variations in sample valve actuation time. The sample valve need only remain in the "sampling" position long enough to completely flush the sample loop tube, and the proper volume of injected sample gas is guaranteed.

While in the loading position, the stream of gas sampled from the process continuously fills the

sample loop and then exits to a waste port. This may seem unnecessary but it is in fact essential for practical sampling operation. The volume of process gas injected into the chromatograph column during each cycle is so small (typically measured in units of *micro*liters!) that a continuous flow of sample gas to waste is necessary to purge the impulse line connecting the analyzer to the process, which in turn is necessary for the analyzer to obtain analyses of current conditions. If it were not for the continuous flow of sample to waste, it would take a *very long time* for a sample of process gas to make its way through the long impulse tube to the analyzer to be sampled!



Even with continuous flow in the impulse line, process chromatographs exhibit substantial dead time in their analyses for the simple reason of having to wait for the next sample to progress through the entire length of the column. It is the basic nature of a chromatograph to separate components of a chemical stream over time, and so a certain amount of dead time will be inevitable. However, dead time in any measuring instrument is an undesirable quality. Dead time in a feedback control loop is especially bad, as it greatly increases the chances of instability.

One way to reduce the dead time of a chromatograph is to alter some of its operating parameters during the analysis cycle in such a way that it speeds up the progress of the mobile phase during periods of time where slowness of elution is not as important for fine separation of components. The flow rate of the mobile phase may be altered, the temperature of the column may be ramped up or down, and even different columns may be switched into the mobile phase stream. In chromatography, we refer to this on-line alteration of parameters as *programming*. Temperature programming is an especially popular feature of process gas chromatographs, due to the direct effect temperature has on the viscosity of a flowing gas[23]. Carefully altering the operating temperature of a GC column

[23] Whereas most liquids decrease in viscosity as temperature rises, gases increase in viscosity as they get hotter. Since the flow regime through a chromatograph column is most definitely laminar and not turbulent, viscosity has a great effect on flow rate.

while a sample washes through it is an excellent way to optimize the separation and time delay properties of a column, effectively realizing the high separation properties of a long column with the reduced dead time of a much shorter column.

22.4 Optical analyses

Light is known to interact with matter in very specific ways, which may be exploited as a means of measuring the composition of gases and liquids. Either a sample of substance to be analyzed is stimulated into emitting light (optical *emission*), or light from a stable source is passed through a transparent sample or reflected off an opaque sample (optical *absorption*). The specific frequencies (colors) of light obtained from these analyses serve to identify the chemical elements and/or compounds present in the sample, and the relative intensities of each spectral pattern indicate the concentrations of those elements and/or compounds.

The theoretical basis for optical analysis is the interaction between charged particles of matter and light, which may be modeled both as a particle (called a *photon*) and as an electromagnetic wave possessing a frequency (f) and a wavelength (λ). Thanks to the work of the physicists Max Planck and Albert Einstein at the beginning of the 20th century, we now know there is a definite proportionality between the frequency of a light wave and the amount of energy each photon carries (E). This proportionality is *Planck's constant*, or h:

$$E = hf$$

Where,

E = Energy carried by a single "photon" of light (joules)
h = Planck's constant (6.626×10^{-34} joule-seconds)
f = Frequency of light wave (Hz, or 1/seconds)

If the amount of energy carried by a photon happens to match the energy required to make an atomic electron "jump" from one energy level to another, the photon will be consumed in the work of that task when it strikes the atom. Conversely, when the electron returns to its original (lower) energy level in the atom, it releases a photon of the same frequency as the original photon that dislodged the electron.

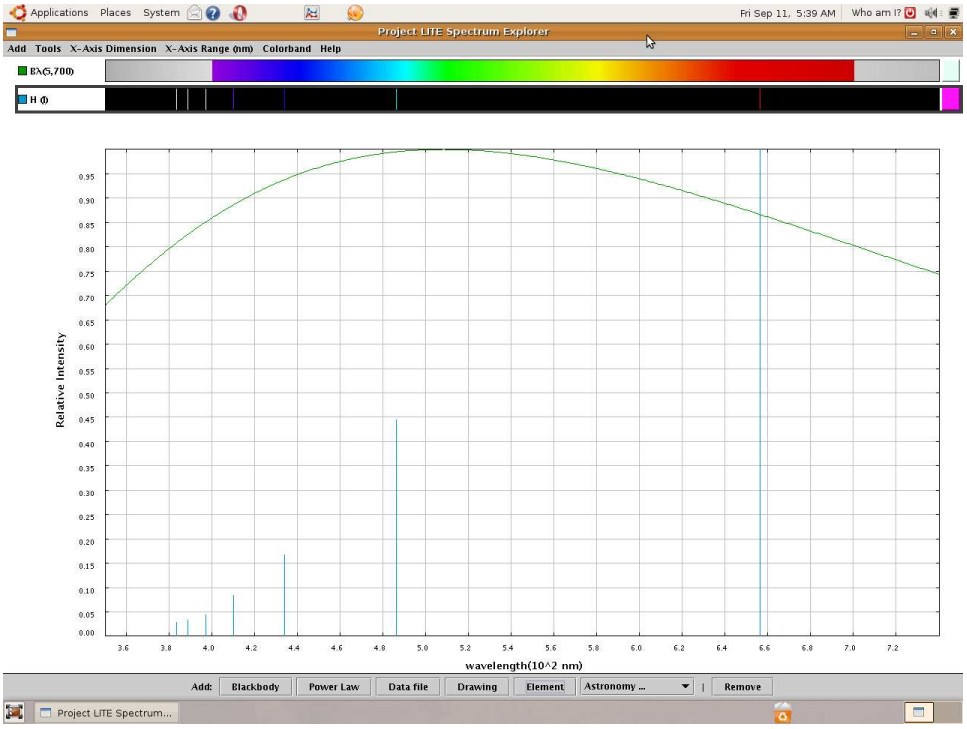
Since each element's electron configuration is unique, each element's electrons respond differently to light. Both the colors (frequencies) of light required to boost electron energy levels and the colors (frequencies) of light emitted by those atoms as their electrons fall back to their original energy levels constitute a unique "optical fingerprint" for identifying elements.

If we examine the visible light spectrum (a range of wavelengths spanning 700 nm to 400 nm, corresponding to a range of frequencies spanning 4.29×10^{14} Hz to 7.5×10^{14} Hz) emitted by a *blackbody*[24] heated to a temperature of 5700 Kelvin, we see a continuous spectrum of color from violet on the left (short wavelength, high frequency, high energy) to red on the right (long wavelength, low frequency, low energy). Here, I am using a computer program called *Spectrum Explorer* (SPEX) to map both the color spectrum and the intensity of radiation across a range of wavelengths:

Unless the light from a heated blackbody is passed through some device to separate it into its constituent colors, the human eye blends all the colors together and only sees *white*. Thus, we use the term "white light" to refer to an equal mixture of light frequencies covering the visible spectrum. The grey areas to the far left and far right of the spectrum represent the ultraviolet and infrared regions, respectively, that lie outside of the human vision range. A blackbody heated to 5700 K emits substantial quantities of both ultraviolet and infrared radiation, but this radiation is invisible to the human eye.

[24] In physics, a "blackbody" is a perfect emitter of electromagnetic radiation (photons) as it is heated. The intensity of light emitted as a function of wavelength (λ) and temperature (T) is $I = \frac{2\pi hc^2\lambda^{-5}}{e^{hc/\lambda kT}-1}$.

If we take a sample of pure hydrogen gas and heat it using an electric arc (inside a glass tube), the hydrogen atoms' electrons will be forced into higher energy states by the passage of electric current through the gas. As those electrons fall back to lower original energy levels, they emit photons of characteristic wavelengths (color). These wavelengths do *not* cover the visible spectrum as they do for blackbody objects, but rather reveal themselves as thin "lines" on the visible spectrum range, and as "peaks" on the intensity plot:



Viewed with the unaided human eye, the light emitted from a hydrogen gas discharge tube looks bright red, because that is the predominant wavelength emitted. The other colors tend to be overshadowed by the red, but we can still view them if we pass the light through a prism or through a diffraction grating to split it up into its constituent colors.

This particular set of "lines" is unique for the element hydrogen, and may serve as an identifying "fingerprint" for hydrogen if found in the emission spectrum for any chemical sample generated by the same method.

An alternative to electrically stimulating a quantity of hydrogen gas to make it generate specific colors of light is to pass white light through a sample of hydrogen gas and then look for which colors are *absorbed* by the gas. As mentioned previously, photons having the necessary energies (frequencies) will become consumed in the work of elevating the hydrogen atoms' electrons to higher energy levels, leaving dark lines in an otherwise unbroken spectrum of colors from violet to red. This is called the *absorption spectrum* for an element, in contrast to the *emission spectrum* obtained by electrically energizing atoms of that element to emit light.

The following illustration shows three different spectra: the *full-color* (white light) spectrum of white light (top), the *emission* spectrum of hydrogen gas (middle), and the *absorption* spectrum of hydrogen gas (bottom). Note how the dark gaps in the absorption spectrum precisely match the positions and colors of the bright lines in the emission spectrum, because the wavelengths of light *absorbed* from white light as it passes through hydrogen gas are the exact same wavelengths *emitted* by hydrogen gas when stimulated by an electric spark in a glass tube:

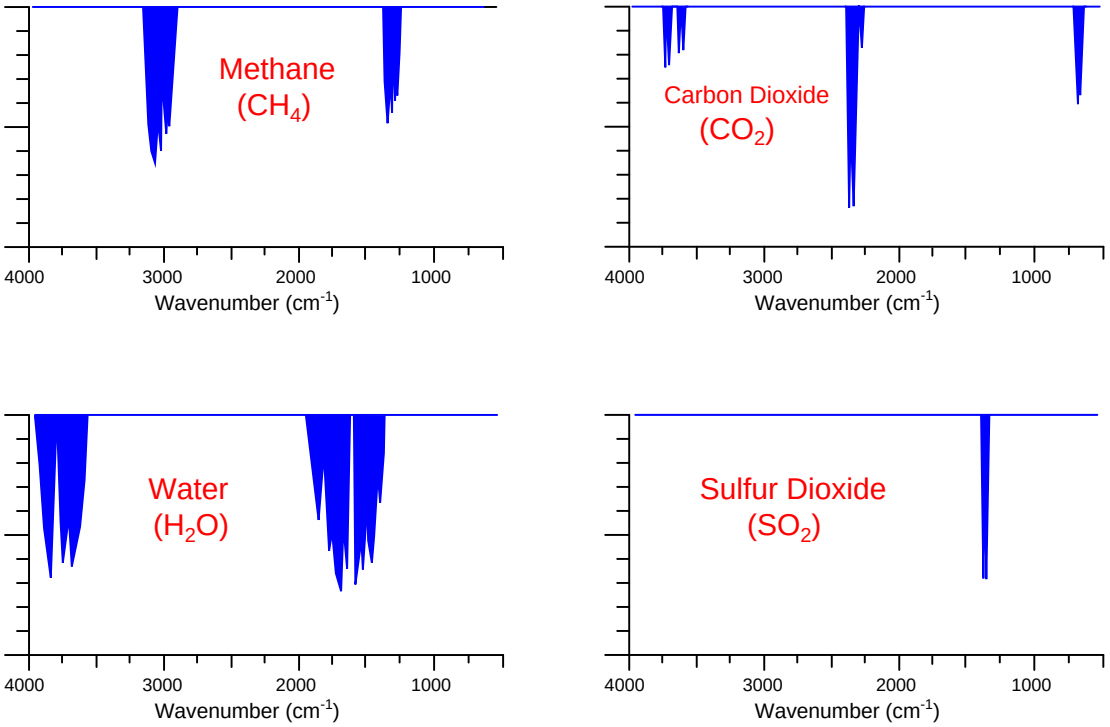
The dark lines found in the absorption spectrum constitute a distinctive "fingerprint" for the element hydrogen, and may be used to detect the presence of hydrogen in gas samples through which white light is passed.

Usually in industrial analysis we are more concerned with the quantifiable presence of certain *compounds* in a process sample than we are in the presence of certain elements. Fortunately, molecules have their own distinctive interactions with light. Sometimes, these interactions take the form of molecular electrons being boosted into higher energy levels, much the same as with individual atoms. Other molecular-optical interactions take the form of *vibrations* and *rotations* set up between the atoms of a molecule, usually with photons in the infrared range[25]. As an infrared photon of the correct wavelength (energy value) strikes an appropriate molecule, its frequency resonates with the bonded atoms, almost as if they acted as miniscule masses connected together by coil springs. This causes a transfer of energy from the photon to the molecule, where the vibration eventually dissipates that energy in the form of heat.

Thus, shining infrared and/or ultraviolet light through a sample of process gas, and analyzing the wavelengths absorbed by that gas sample, can provide quantitative measurements of the concentrations of certain gases types in that sample.

[25] These photons have wavelengths longer than 700 nm, and so have energy values too low to boost electrons into higher levels. However, the attractive bonds *between* atoms in a molecule may be subject to the energy of these infrared photons, and so may dissipate the photons' energy and thereby attenuate a beam of infrared light.

A few different infrared absorption spectra[26] for common industrial compounds are shown here, with the frequency shown in units of *wavenumber* (the number of wavelengths per centimeter). It should be noted that these absorption spectra are not drawn in scale to each other; rather, they are each drawn to their own scale to better show the relative sizes of the different absorption "dips" across the spectrum for each substance:



Note that the pattern of each absorption spectrum is unique. Each compound tends to absorb infrared light in its own way, and these "signature" absorption patterns provide us with a means to selectively identify the presence of various compounds in a process fluid sample.

Molecule types most effective at absorbing infrared light are those comprised of different atom types, such as carbon monoxide (CO), carbon dioxide (CO_2), sulfur dioxide (SO_2), water vapor (H_2O), and oxides of nitrogen (NO_x). Molecules formed of two atoms of the same type such as molecular oxygen (O_2), nitrogen (N_2), and hydrogen (H_2) exhibit negligible interaction with infrared light. This is a fortuitous quality of infrared analysis, because many process monitoring applications focus specifically on the former compounds to the exclusion of the latter. Monitoring the exhaust emissions of a large combustion system, for example, is an application where the

[26] In an absorption spectrum diagram, a non-absorbing substance results in a straight line at the 100% mark. Compounds absorbing specific wavelengths of light will produce low "dips" in the graph at those wavelength values, showing how less light (of those wavelengths) is able to pass un-absorbed through the sample to be detected at the other end. By contrast *emission* spectra are usually plotted with the characteristic wavelengths shown as high "peaks" in a graph that normally resides at 0%.

concentration(s) of CO, CO_2, SO_2, and/or NO_x are relevant but the concentration of nitrogen (N_2) is not. As with all chemical analyses, the "trick" is to find some property of measurement applicable only to the substance you are interested in measuring, and not to any others. This is the only way an analytical instrument may discriminate between the substance of interest and the other "background" substances.

Between optical emission and optical absorption, absorption analysis seems to be the more popular in modern industrial use, with optical emission analysis limited mostly to laboratory applications. One reason for this is the necessity of heating a sample to a high enough temperature where it emits light: an energy-intensive and potentially hazardous endeavor. Absorption analyzers need only shine a beam of light through an unheated sample chamber, then measure how much of specific wavelengths were absorbed by the sample. Another important reason for the prevalence of absorption analyzers in industry is the necessity of a sophisticated computer and algorithm to sort the line spectra of substances generated in emission-type analyzers. Inventors have devised clever ways to quantify the absorption spectra of different process substances without resorting to automated pattern-matching of spectra.

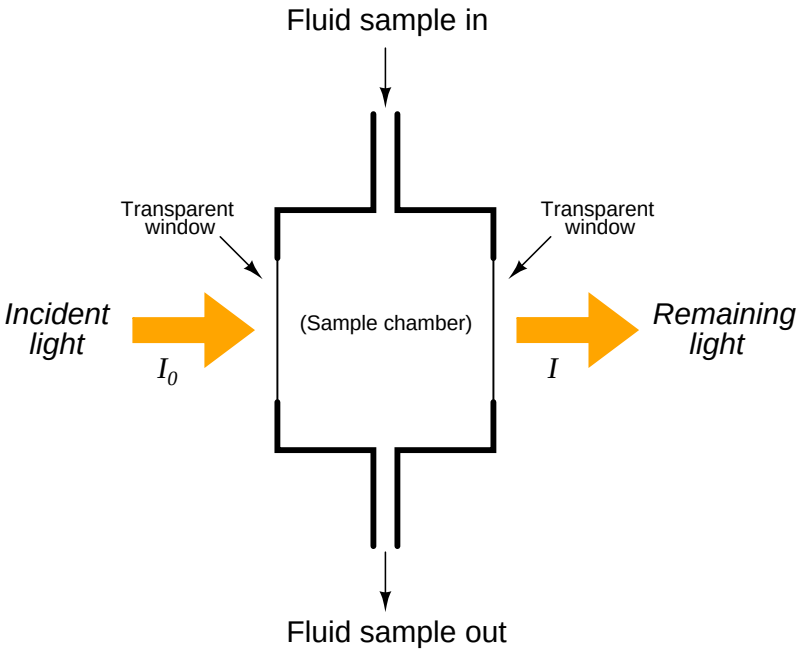
In every absorption-type optical analyzer, the fundamental equation relating photon absorption to substance concentration is the *Beer-Lambert Law* (sometimes called the *Lambert-Beer Law*):

$$A = abc = \log\left(\frac{I_0}{I}\right)$$

Where,

A = Absorbance
a = Extinction coefficient for photon-absorbing substance(s)
b = Path length of light traveling through the sample
c = Concentration of photon-absorbing substance in the sample
I_0 = Intensity of source (incident) light
I_0 = Intensity of received light after passing through the sample

A typical arrangement for exposing a fluid sample (liquid or air) to light is shown in this diagram:

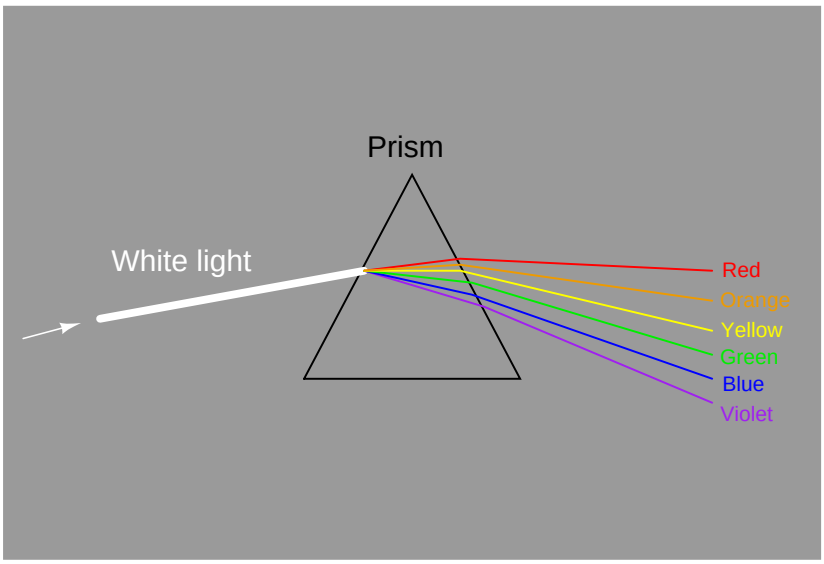


As indicated by the Beer-Lambert equation, greater sensitivity will be achieved with a longer path length. In some applications where the substance of interest is an atmospheric pollutant, the light beam is simply shot through open air (usually reflecting on a mirror) before returning to the instrument for analysis. If the light source happens to be a laser, the distance may be quite large – one such analyzer I saw in industry has a path length of a quarter-mile (1320 feet), to better measure extremely low concentrations of a gas!

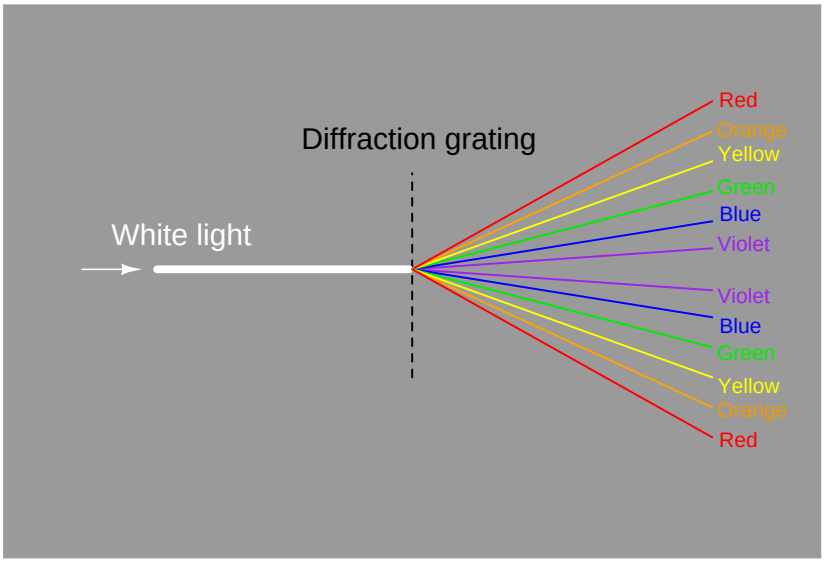
Once the light has been passed through (or reflected off of) the process sample, it must be analyzed for attenuated wavelengths. Two major types of wavelength analysis exist: *dispersive* (where the light is split up into its constituent wavelengths) and *nondispersive* (where the spectral distribution of the wavelengths is detected without separating colors). These two optical analysis methods form the subject of the next two subsections.

22.4.1 Dispersive spectroscopy

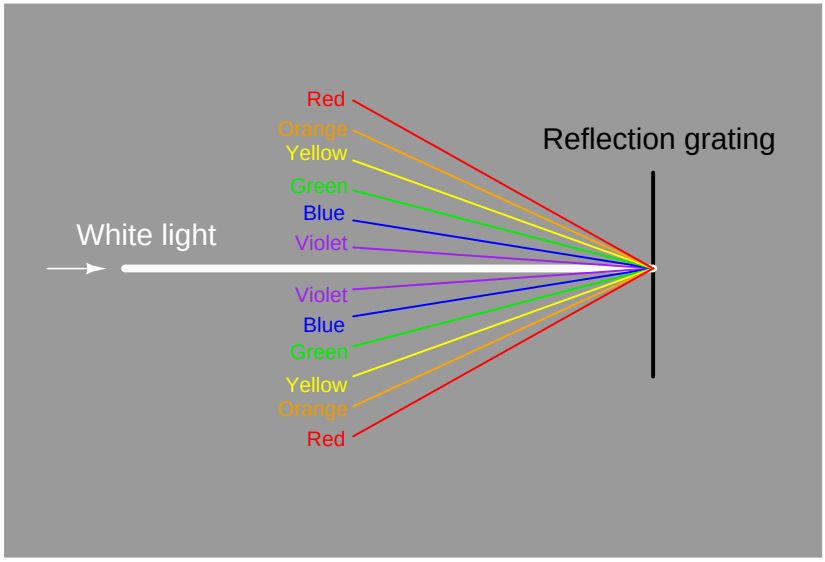
The dispersion of visible light into its constituent colors goes all the way back to the 17th century with Isaac Newton's experiments, taking a glass *prism* and generating the characteristic "rainbow" of colors:



A modern variation on the theme of a solid glass prism is a thin *diffraction grating*, causing light of different wavelengths to "bend" as they pass through a series of very thin slits:



Some dispersive analyzers use a *reflection grating* instead of a refraction grating. Reflection gratings use fine lines etched on a reflective (mirror) surface to produce an equivalent dispersive effect to a diffraction grating[27]:



In 1814, the German physicist Joseph von Fraunhofer closely analyzed the spectrum of colors obtained from sunlight and noticed the existence of several dark bands in the otherwise uninterrupted spectrum where specific colors seemed to be attenuated. Later that century, experiments by the French physicist Jean Bernard Léon Foucault and the German physicist Gustav Robert Kirchoff confirmed the same effect when white light was passed through a vapor of sodium. They correctly reasoned that the sun's core produced a continuous spectrum[28] of light (all wavelengths) due to its

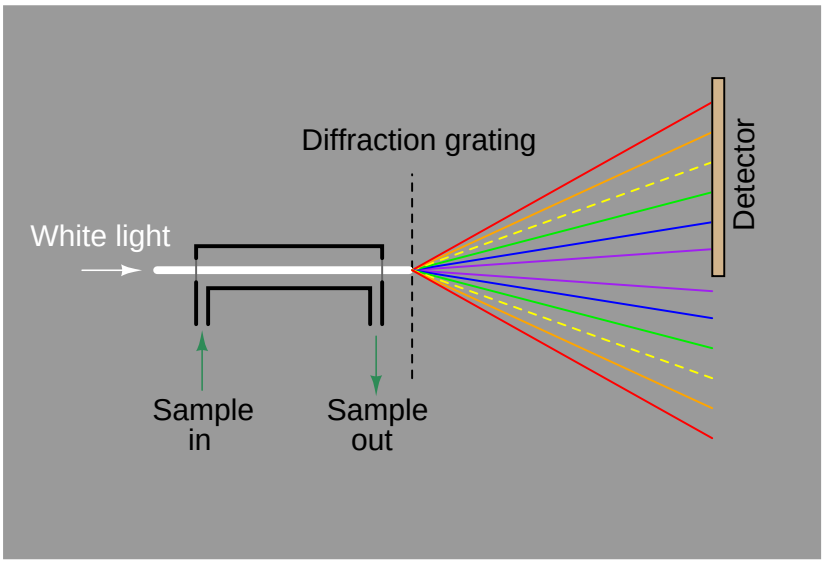
[27] You may use an old compact disk (CD) as a simple reflection and refraction grating. Holding the CD with the reflective (shiny) surface angled toward you, light reflected from a bright source such as a lamp (avoid using the sun, as you can easily damage your eyes viewing reflected sunlight!) will split into its constituent colors by reflection off the CD's surface. Lines in the plastic of the CD perform the dispersion of wavelengths. You will likely have to experiment with the angle you hold the CD, pointing it more perpendicular to the lamp's direction and more angled to your eyes, before you see the image of the lamp "smeared" as a colorful spectrum. To use the CD as a diffraction grating, you will have to carefully peel the reflective aluminum foil off the front side of the disk. Use a sharp tool to scribe the disk's front surface from center to outer edge (tracing a radius line), then use sticky tape to carefully peel the scribed foil off the plastic disk. When you are finished removing all the foil, you may look *through* the transparent plastic and see spectra from light sources on the other side. Once again, experimentation is in order to find the optimum viewing angle, and be sure to avoid looking at the sun this way!

[28] One might wonder why the sun does not produce a line-type emission spectrum of all its constituent elements, instead of the continuous spectrum it does. The answer to this question is that emission spectra are produced only when the "excited" atoms are in relative isolation from each other, such as is the case in a low-pressure gas. In solids, liquids, and high-pressure gases, the close proximity of the atoms to each other creates many different opportunities for electrons to "jump" to lower energy levels. With all those different alternatives, the electrons emit a whole range of different wavelength photons as they seek lower energy levels, not just the few wavelengths associated with the limited energy levels offered by an isolated atom. We see the same effect on Earth when we heat metals: the electrons in a solid or liquid metal sample have so many different optional energy levels to "fall" to, they end up emitting a broad spectrum of wavelengths instead of just a few. In this way, a molten metal is a good approximation of a blackbody photon source.

intense heat, but that certain gaseous elements (including sodium) in the cooler, outer "atmosphere" of the sun were absorbing some of the wavelengths to cause the *Fraunhofer lines* in the observed spectrum. These scientists noted the same patterns of absorption (dark lines) in the sun's spectrum that appeared in laboratory absorption tests with sodium. The implication of these scientists' experiments are truly staggering, as they were able to correctly identify gaseous elements *93 billion miles away from Earth!*.

This sort of spectrographic analysis is called *dispersive*, because it relies on a device such as a prism or diffraction grating to *disperse* the different wavelengths of light from each other so they may be independently measured.

A dispersive analyzer for process fluids would be constructed in this manner, introducing incident light to a windowed sample chamber where some wavelengths of that light would be attenuated by interaction with the process fluid molecules. In this case, I have hypothesized a sample that absorbs some of the yellow light wavelengths, resulting in less yellow light reaching the detector array:



The light source need not output white light if the wavelengths of interest do not span the entire visible spectrum. For example, if the absorption spectrum of a particular substance is known to primarily span infrared light and not the visible range, it may be sufficient to use a dispersive analyzer with an infrared light source rather than a "broad spectrum" light source covering both the infrared and visible ranges.

A necessary component of any dispersive analyzer is a computer connected to the detector array with the ability to recognize all expected emission spectra patterns, and quantify them based on the relative strengths of the detected wavelengths. This is a level of sophistication far beyond most industrial measuring instruments, which is one reason dispersive analyzers are not as popular (yet!) for industrial process use. However, once such a computer and necessary software are in place to perform the analyses, measurement of multiple substances from the same absorption spectrum becomes possible. Like chromatographs, dispersive optical analyzers naturally function as multi-component measurement devices.

22.4.2 Non-dispersive spectroscopy

Non-dispersive analysis, while newer in discovery than dispersive analysis (Newton's 17th-century prism), has actually seen far earlier application as continuous process analyzers. The basic design was developed during the years 1937-1938 by Dr. Luft and Dr. Lehrer in the laboratories of the German chemical company *I.G. Farbenindustrie*. By the end of World War II, over four hundred of these innovative instruments were in service in German chemical plants.

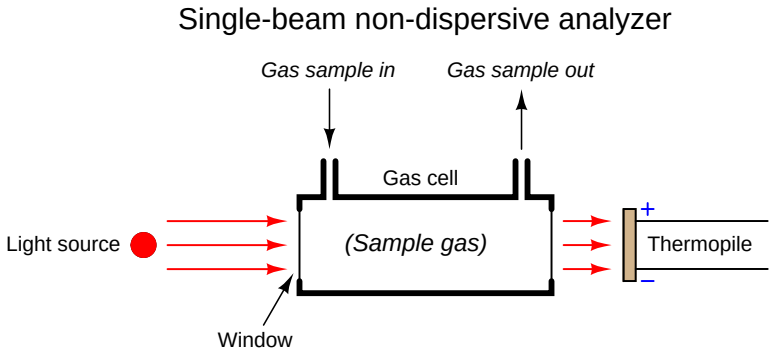
Industrial non-dispersive analyzers typically use either infrared or ultraviolet light sources, because most substances of interest absorb wavelengths in those regions rather than in the visible light spectrum. Non-dispersive spectroscopy using infrared light is usually abbreviated *NDIR*, while non-dispersive spectroscopy using ultraviolet light is abbreviated *NDUV*. Historically, NDIR is the more prevalent of the two technologies. Also, *gas* analysis is the more common application in industry, as opposed to *liquid* analysis, which is why all the examples in this portion of the book assume the analysis of a process gas.

The "trick" of non-dispersive spectrographic analysis is how to achieve selectivity, where the analyzing instrument responds to just one substance rather than to *any* substance that absorbs light. Dispersive spectrographs achieve selectivity by "disassembling" the spectrum into individual wavelengths and measuring them one by one, but a non-dispersive analyzer must somehow distinguish different spectral responses without this "disassembly" of wavelengths.

Many variations of non-dispersive analyzers exist – too many in fact to cover within the scope of this textbook. The important point as always, dear reader, is to grasp the general *concepts* involved with any type of technology, for all distinctions past that point are merely variations on a common theme.

A simple single-beam analyzer

Like dispersive analysis, non-dispersive analysis begins with a light beam passing through a sample substance, often enclosed in a windowed sample chamber (typically called a *cell*). Certain "species" of gas introduced into this cell absorb part of the incident light, leaving the light exiting the cell partially depleted of specific wavelengths. As the concentration of any light-absorbing gas increases in the cell, a detector placed at the other end of the cell receives less and less light of the absorbed wavelengths. The simplest style of non-dispersive analyzer uses a single light source, shining continuously through a single gas cell, and eventually falling on a small thermopile (converting the received infrared light into heat, and then into a voltage signal):



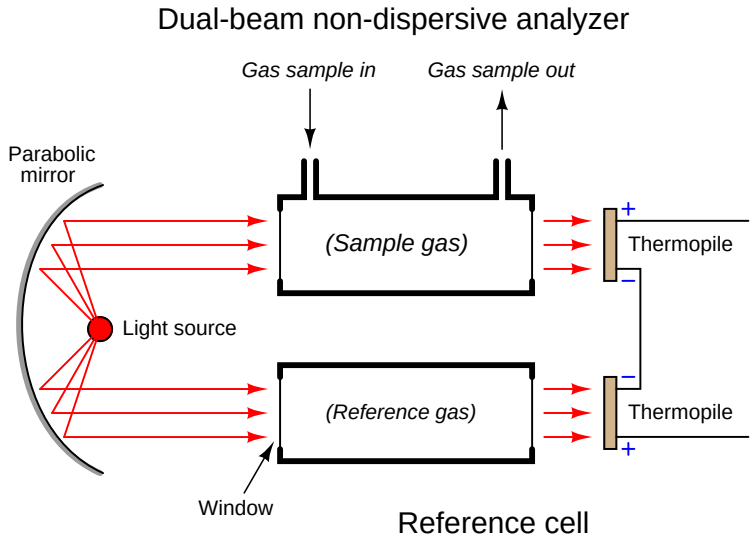
This crude analyzer suffers from multiple problems. First, it is non-selective: *any* light-absorbing gas entering the sample cell will cause a change in the detector's signal, regardless of the species. It might work well enough in an application where *only* the gas of interest absorbs light in the wavelength range of the source, but most industrial analyzer applications are not like this. In most cases, our process sample contains multiple species of gases capable of absorbing light within a similar range of wavelengths, but we are only interested in measuring one of them. An example would be the measurement of carbon dioxide (CO_2) concentration in the exhaust gas of a combustion furnace: most of the gases exiting the furnace do not absorb infrared light (nitrogen, oxygen) and CO_2 does, but carbon monoxide (CO), water vapor (H_2O), and sulfur dioxide (SO_2) also absorb infrared light, and are all normally present in the exhaust gas of a furnace to varying degrees. Our crude infrared analyzer cannot tell the difference between carbon dioxide and any of the other infrared-absorbing gases present in the exhaust gas.

Another significant problem with this analyzer design is that any variations in the light source's output cause both a zero shift and a span shift in the instrument's calibration. Since light sources tend to change output with age, this flaw necessitates frequent re-calibration of the analyzer.

Finally, since the detector is a thermopile, its output will be affected not just by the light falling on it, but also by ambient temperature, causing the analyzer's output to vary in ways completely unrelated to the sample composition.

A simple dual-beam analyzer

One way to improve on the single-beam analyzer design is to split the light beam into two equal halves, then pass each half-beam through its own cell. Only one of these cells will hold the process gas to be analyzed – the other cell is sealed, containing a "reference" gas such as nitrogen that does not absorb infrared light. At the end of each cell we will place a matched pair of thermopile detectors, connecting these detectors in series-opposing fashion so equal voltages will cancel:



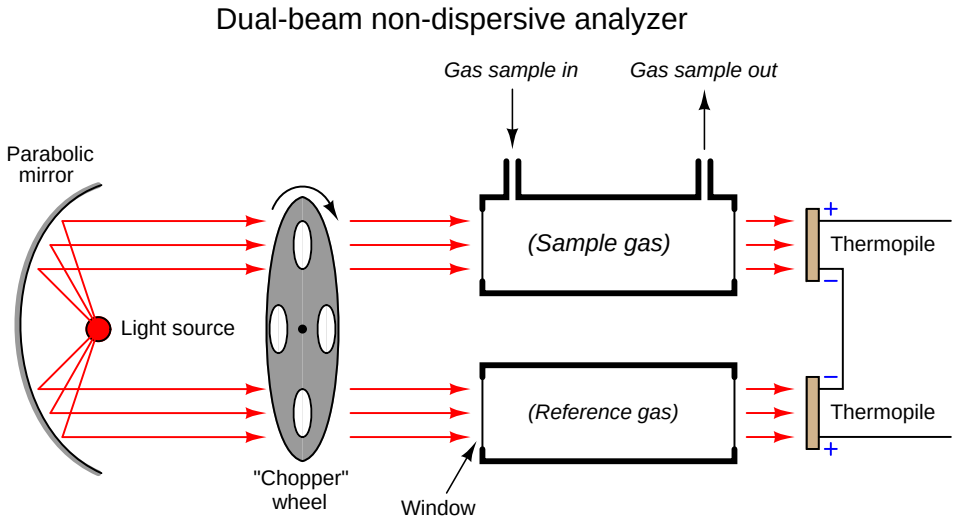
If the sample gas is non-absorbing of infrared light just like the reference gas, the opposed detector pair will generate no voltage signal. If, however, the sample contains some concentration of an infrared-absorbing gas, the two thermopile detectors will receive differing intensities of infrared light. This temperature difference will cause the pair to be out of balance, generating a net voltage signal we can measure as an indication of light-absorbing gas concentration.

This modification completely eliminates the ambient temperature problem. If the analyzer's temperature happens to rise or fall, the voltages output by *both* thermopiles will rise and fall equally, canceling each other out so that the only voltage produced by the series-opposing pair will be that produced by differences in received light intensity.

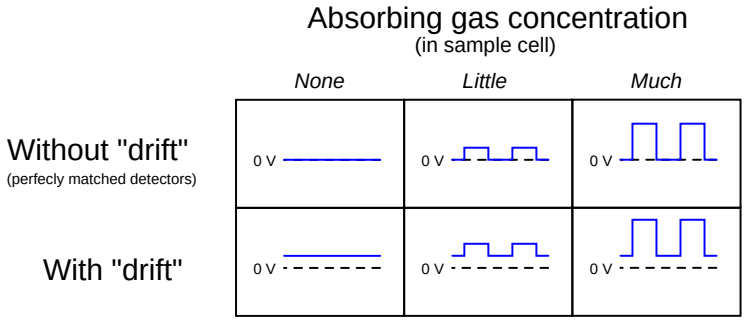
The dual-detector design also eliminates the problem of "zero drift" as the light source ages. As time progresses and the light source becomes dimmer, *both* detectors see less light than before. Since the detector pair measures the difference between the two light beam intensities, any degradation common to both beams will be ignored[29].

[29] There will still be a *span* shift resulting from degradation of the light source, but this is inevitable. At least with this design, the zero-shift problem is eliminated.

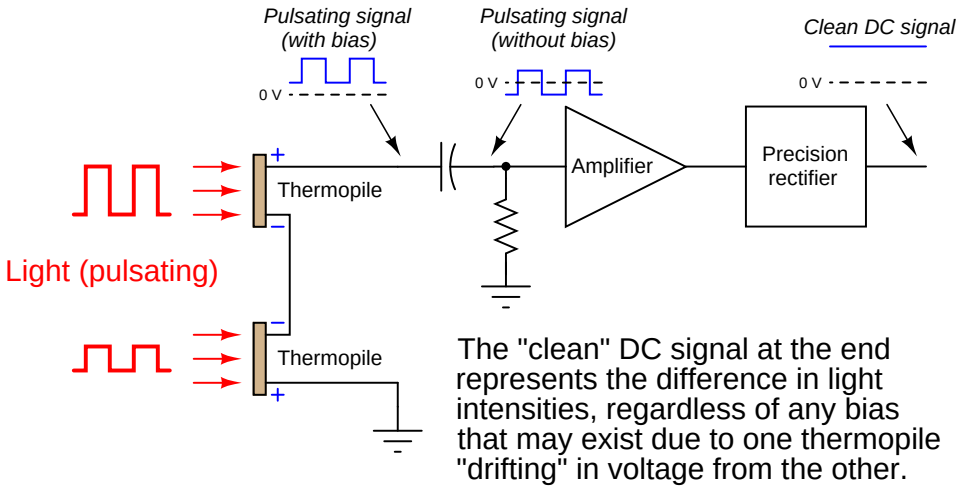
Another detector problem still remains, in that an imbalance will develop if one detector happens to "drift" in voltage apart from the other, so they are no longer in perfect counter-balance even with the same received light intensities. This might happen if one of the thermopiles becomes warmer than the other, perhaps due to heat from the hot process sample gas entering one cell and not the other. An ingenious solution to this problem is to insert a spinning metal "chopper" wheel in the path of both light beams, causing the light beams to *pulse* through the sample and reference cells at a low frequency (typically a few pulses per second):



The effect of the "chopper" is to make the detector assembly output a *pulsating* ("AC") voltage signal rather than a steady voltage signal. The peak-to-peak amplitude of this pulsating signal represents the difference in light intensity between the two detectors, but any "drift" will manifest itself as a constant or very slowly-changing ("DC") bias voltage. The following table illustrates the detector assembly signal for three different gas concentrations (none, little, and much) both with and without a mis-match in detector signals due to thermal drift:



This DC bias voltage is very easy to filter in the amplifier section of the analyzer. All we need is capacitive coupling between the detector assembly and the amplifier, and the amplifier will never "see" the DC bias voltage:

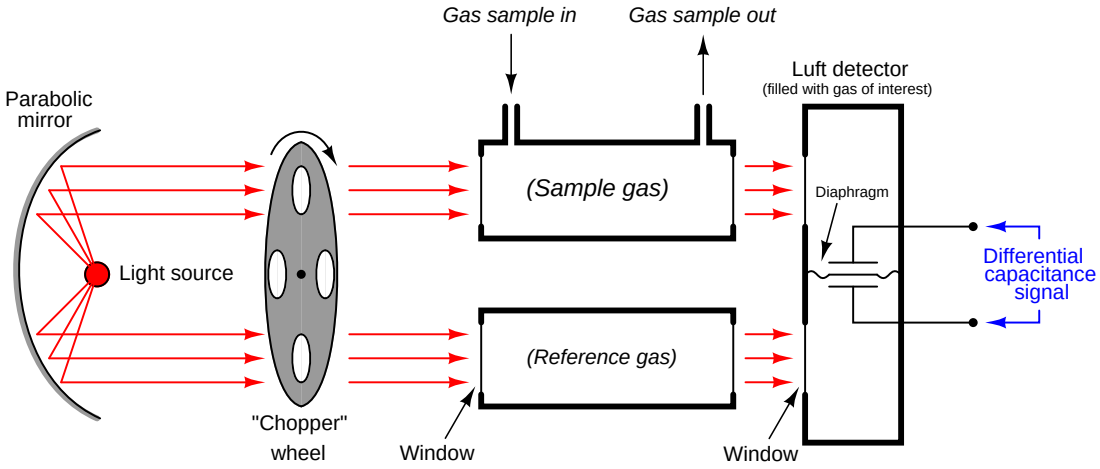


With the detector assembly producing an "AC" (pulsing) signal instead of a "DC" signal, and by using capacitive coupling to the amplifier, the electronic circuit responds only to changes in the amplitude of the AC waveform and not to its DC bias. This means the analyzer will only respond to changes in detector temperature resulting from changes in light absorbance (i.e. gas concentration), and not from any other factor such as ambient temperature drift. In other words, since the amplifier has been built to only amplify pulsing signals, and the only thing pulsing in this instrument is the light, the electronics will only measure the effects generated by the light, rejecting all other stimuli.

Despite the design improvement of the chopper wheel and AC-coupled amplifier circuit, another significant problem remains with this analyzer: it is still a non-selective instrument. *Any* infrared-absorbing gas entering the sample cell will cause the detector pair to generate a signal, regardless of the type of gas. While this may suffice for some industrial applications, it will not for most where a mixture of infrared-absorbing gases coexist. What we need is a way to make this instrument *selective* to just one type of gas, in order that it be a useful analyzer in a wider variety of process applications.

Luft-detectors

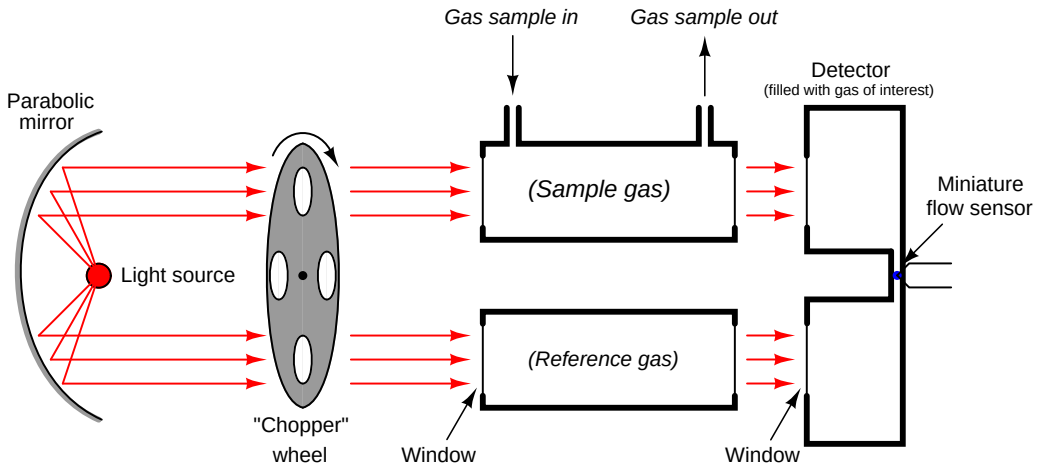
A very clever way to gain some selectivity is to replace the thermopiles with a different style of detector more sensitive to the wavelengths absorbed by the gas of interest than to the wavelengths absorbed by any other ("interfering") gas. A German inventor by the name of Luft produced just such a detector using two gas chambers and a thin diaphragm to measure the difference in light intensity exiting the sample and reference cells. This style of detector is still known today as a *Luft detector*:



As infrared light enters the dual chambers of the detector, the resulting heat causes the gas inside to expand, pressing against the thin diaphragm. If the light intensities are equal, the pressures will be equal and no diaphragm motion will result. If the light intensities are unequal (due to the sample cell absorbing some of the wavelengths), the gas pressure developed inside that half of the Luft detector will be less, causing the thin diaphragm to bow in that direction. A set of fixed metal plates senses the diaphragm's position using the differential capacitance technique (just like many modern differential pressure sensors). With the "chopper" wheel working to pulsate light through the sample and reference gas cells, the diaphragm inside the Luft detector will likewise pulsate, and the resulting "AC" pulse signal may be filtered and amplified to represent absorbing gas concentration.

What makes the Luft detector selective is that it is filled with a 100% concentration of the gas we are interested in measuring. This means only those wavelengths of light absorbed by the gas of interest will develop heat (and pressure) inside the detector chambers. Different wavelengths of light absorbed by other ("interfering") gases in the sample won't be absorbed the same way by the gas inside the Luft detector, and therefore the pressure pulses inside the Luft detector will be primarily a function of our interest-gas concentration and not of the interfering-gas concentration(s).

A modern variation on the Luft detector design replaces the diaphragm with a narrow channel and a highly sensitive thermal flow detector connecting the two gas-filled chambers. Any difference in expansion between the gases of the two chambers when heated by light causes gas to move past the flow detector, thus generating a signal:

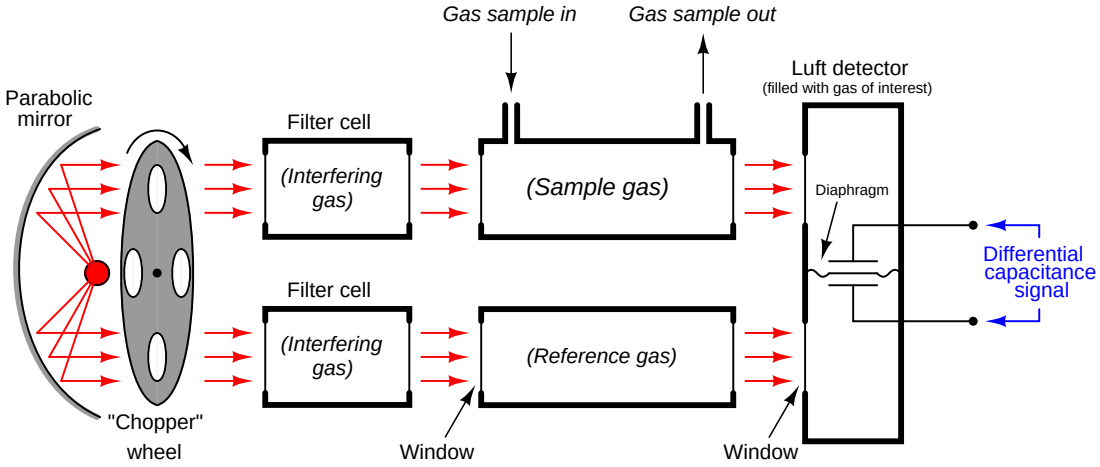


The advantage of a diaphragm-less detector is that it is just as insensitive to mechanical vibration as a thermopile (having no moving parts), but retains the spectral sensitivity of the traditional Luft-style detector (being filled with the gas of interest).

Use of filter cells

If the interfering gas(es) do not absorb any of the light wavelengths as our gas of interest, the selectivity will be total: the gas-filled detector will *only* respond to the presence of the gas we are interested in. Usually, though, process applications are not this simple. In most applications, the interfering gases have absorption spectra overlapping portions of the interest-gas spectrum. This means changes in interference gas concentration will be sensed by the detector (though not as strongly as changes in the concentration of the gas of interest) because part of the light spectrum absorbed by the interfering gas(es) will have a heating effect on the pure gas inside the detector.

One more addition to our NDIR instrument helps eliminate this problem: we add two more gas cells in the path of the light beams, each one filled with 100% concentrations of the interfering gas:

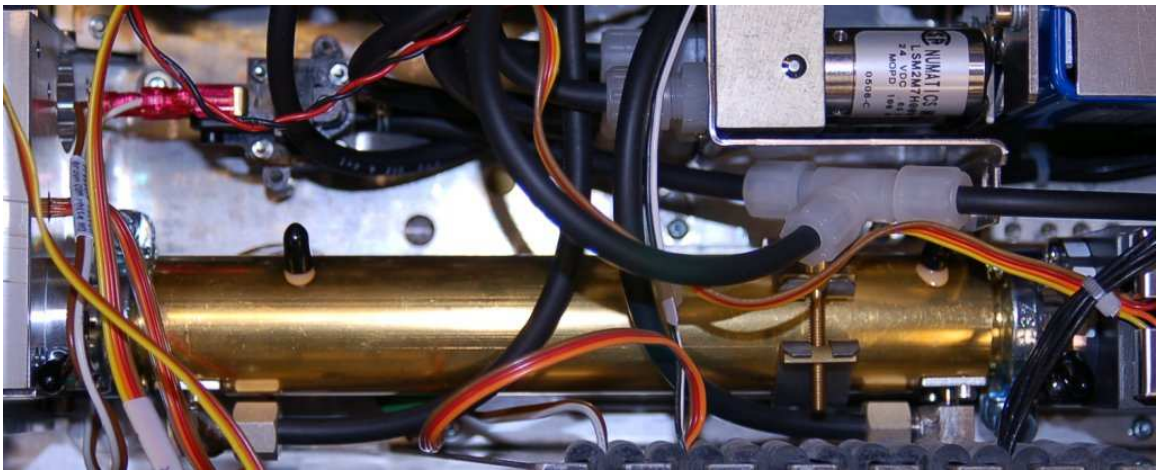


These *filter cells* purge the light of those wavelengths normally absorbed by the interfering gas. As a result, no concentration of that interfering gas in the sample cell will have any effect on the light exiting the sample cell, because those wavelengths have already been eliminated by the filter cells. So long as our gas of interest exhibits absorption wavelengths *not shared by the interfering gas* (i.e. wavelengths of light unique to the gas of interest alone), these wavelengths will pass through the filter cells and into the sample cell where they will change intensity as the gas of interest varies in concentration. Thus, the detector now *only* responds to changes in the gas of interest, and not to changes in the interfering gas.

As effective as this filtering technique is, it has the limitation of only working for one interfering gas at a time. If multiple interfering gases exist in the sample stream, we must use multiple filter cells to block those light wavelengths[30].

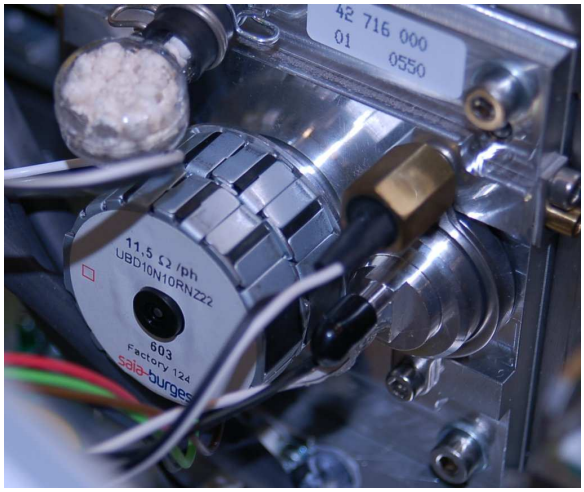
[30] And hopefully after all this filtering we still have some (unfiltered) wavelengths unique to the gas of interest we seek to measure. Otherwise, there will be no wavelengths of light remaining to be absorbed by our gas of interest inside the sample cell, which means we will have no means of spectroscopically measuring its concentration!

A photograph showing a dual-beam NDIR analyzer appears here:



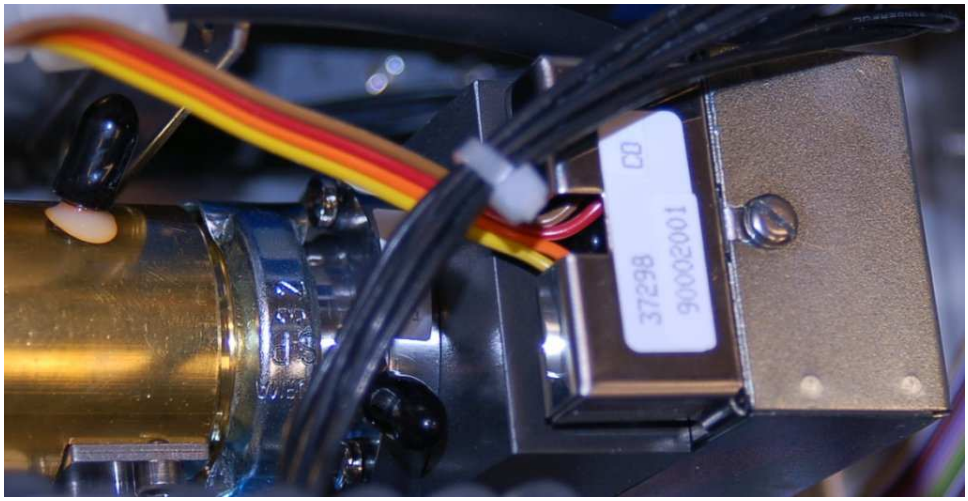
What looks like a single gold-colored gas cell is actually two cells (a divider separating the tube lengthwise into two chambers), one for the sample gas and the other for reference. Black-colored hoses pass sample gas through the bottom half of the tube, while the top half is filled with nitrogen gas (the tube connections capped and sealed with black-colored plastic). The light source and chopper assembly appears on the left-hand side of the tube, while the detector resides on the right-hand side.

In this particular instrument (a Rosemount Analytical model X-STREAM X2), the chopper wheel is driven by a stepper motor:



The head of the infrared light source appears just to the right of the chopper wheel motor.

The detector used in the X-STREAM NDIR analyzer is a modern variant of the Luft detector, with a micro-flow sensing element detecting pulses of gas flow between two chambers. In this particular analyzer the detector chambers are filled with carbon monoxide (CO) gas, to sensitize it to that species:



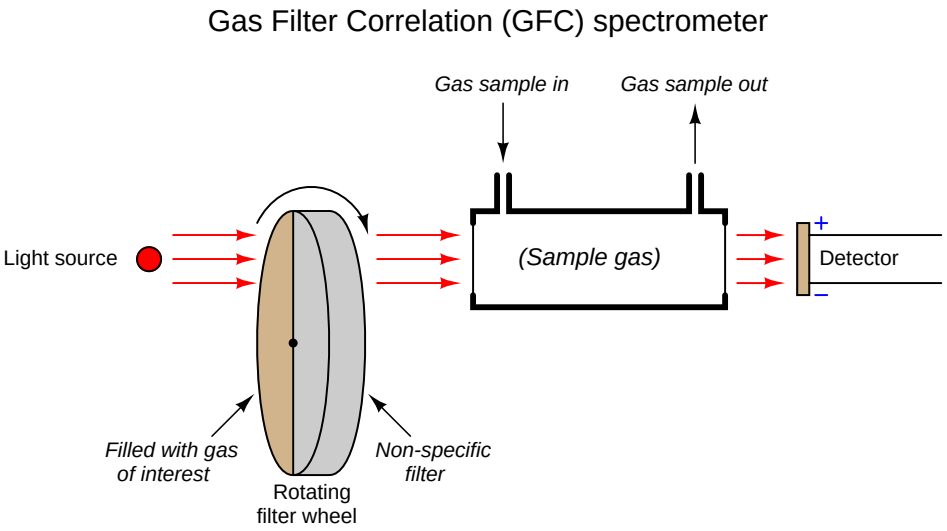
This instrument's maximum detection range happens to be 0 to 1000 ppm of carbon monoxide, with the ability to turn down to a range of 0 to 400 ppm.

Gas Filter Correlation (GFC) analyzers

Using filter cells to eliminate wavelengths associated with interfering gases is called *positive filtering* in the field of spectroscopy. You may think of this as filtering out all the wavelengths the instrument should *not* "care about." In order for positive filtering to be completely effective, the analyzer must filter out *all* wavelengths associated with all interfering species. In some applications, this may require multiple filters stacked in "series," each one filtering out wavelengths for a different interfering gas. Not only is this technique cumbersome when multiple interfering "species" are present in the sample, but it is completely useless when the interfering species are unknown.

A different filtering technique called *negative filtering* does just the opposite: placing a filter cell in the path of the light to absorb all the wavelengths associated with the gas of interest, leaving all other wavelengths unattenuated. One application of this technique is called *Gas Filter Correlation*, or *GFC* spectroscopy. This same technique is alternatively referred to as *Interference Filter Correlation*, or *IFC* spectroscopy.

Gas filter correlation analyzers use a single gas cell rather than dual cells (sample and reference), through which a light beam of alternating spectrum is passed. A rotating filter wheel creates this alternating spectrum:



The filter wheel consists of two transparent halves: one containing a high concentration of the gas of interest, and the other designed to consistently attenuate every light wavelength (i.e. the entire spectrum) emitted by the source. The attenuation factor of the "neutral" half of this filter wheel is precisely adjusted so that the same gross intensity of infrared light enters the sample gas cell at all times, regardless of the filter wheel's position.

If the sample gas chamber contains nothing but non-absorbing gases, the detector will generate a steady (unchanging) signal[31] because it receives the same total light intensity during each half of the filter wheel's rotation.

If some of our gas of interest enters the sample cell, it will begin to absorb some of the light during the time when the "neutral" filter aligns in front of the cell. During the other half of the filter's rotation (when the light must pass through the high gas concentration chamber), our gas of interest inside the sample cell has no effect, because all those wavelengths of light have been eliminated by the filter. The result is a changing signal at the detector[32], the amplitude of oscillation proportional to the concentration of "correlating" gas (matching the absorption spectrum of the rotating filter's gas) inside the sample cell.

[31] Real GFC analyzers also have a chopper wheel preceding the filter wheel to create a pulsating light beam. This causes the detector signal to pulsate as well, allowing the analyzer to electronically filter out sensor "drift" just as in the dual-beam NDIR analyzer design. The chopper wheel has been eliminated from this diagram (and from the discussion) for simplicity. If it were not for the chopper wheel, the GFC analyzer would be prone to measurement errors caused by detector drift.

[32] As previously mentioned, real GFC analyzers have a chopper wheel preceding the filter wheel to make the light beam pulse in addition to changing its spectral composition. This chopper wheel generates multiple light pulses per rotation of the filter wheel. Thus, the signal output by the detector is actually an *amplitude-modulated* waveform, with the "carrier" frequency being the chopper wheel's pulsing and the slower "modulating" frequency being the filter wheel's rotation cycle. Hopefully by now you see why I decided to omit the chopper wheel "for simplicity."

The effect of "interfering" gases in the sample cell depends on the nature of those gases. An "interfering" gas having the exact same absorption spectrum as the gas of interest would be indistinguishable from the gas of interest by this instrument – we would say this gas has a *positive* interference. Such a gas would absorb wavelengths of light from the beam during the time light passes through the "neutral" filter, and it would absorb no wavelengths during the time light passes through the gas filter, behaving just like our gas of interest. A different "interfering" gas absorbing completely different wavelengths of light than our gas of interest would absorb light at all times regardless of the filter wheel's position. However, given an equal *percentage* of absorption in a region of the spectrum untouched by the gas filter side of the wheel, but uniformly attenuated by the "neutral" side of the wheel, means that the effect of this gas would be to absorb more light during the gas-filtered part of the wheel's rotation and less light through the "neutral-filtered" part of the wheel's rotation – just the opposite of positive interference. Thus, a gas with an absorption spectrum wholly different from our gas of interest will have a *negative* interference effect.

In order to avoid interference of any kind from gases other than the one we are interested in measuring, the effects of positive and negative correlation interference must cancel. Fortunately for this technique, most interfering gases partially overlap spectra with most gases of interest. If the degree of overlap is approximately even, the positive and negative interferences will indeed cancel each other, resulting in little or no interference from the "interfering" gas.

To re-phrase this principle: if the absorption spectrum of a gas perfectly correlates with the spectrum for our gas of interest, the effect will be "positive," making the analyzer think there is a greater concentration of the gas of interest than there actually is. If the absorption spectrum of a gas is perfectly *anti-correlated* with the spectrum of our gas of interest, the effect will be "negative," making the analyzer think there is a weaker concentration of our gas of interest than there actually is. However, if the absorption spectrum of any gas is completely *uncorrelated* (i.e. random overlap of spectra) with the spectrum for our gas of interest, there interference will be neutral (little or no effect).

This makes the Gas Filter Correlation (GFC) analyzer ideally suited to distinguish gases whose spectra overlap over the same general ranges but differ in fine detail (i.e. where the individual "peaks" and "dips" in the different spectra randomly intersect). One such practical application for GFC analyzers is combustion exhaust gas analysis for carbon monoxide (CO) in the presence of carbon dioxide (CO_2) and water vapor. Unlike the dual-beam "Luft detector" style of analyzer, the GFC analyzer does not require individual filter cells for each interfering species of gas. This is a major advantage where multiple interfering gases coexist with the gas of interest.

Being a single-beam style of analyzer, GFC instruments are much easier to implement as open-air gas analyzers than dual-beam designs. In other words, the light beam may be passed through open air (or through the diameter of an exhaust stack, for example) to sense gases anywhere in that region, rather than be limited to gases enclosed in a gas cell. Recall from the Beer-Lambert law that absorbance increases in direct proportion to the path length of the light:

$$A = abc = \log\left(\frac{I_0}{I}\right)$$

Where,

A = Absorbance

a = Extinction coefficient for photon-absorbing substance(s)

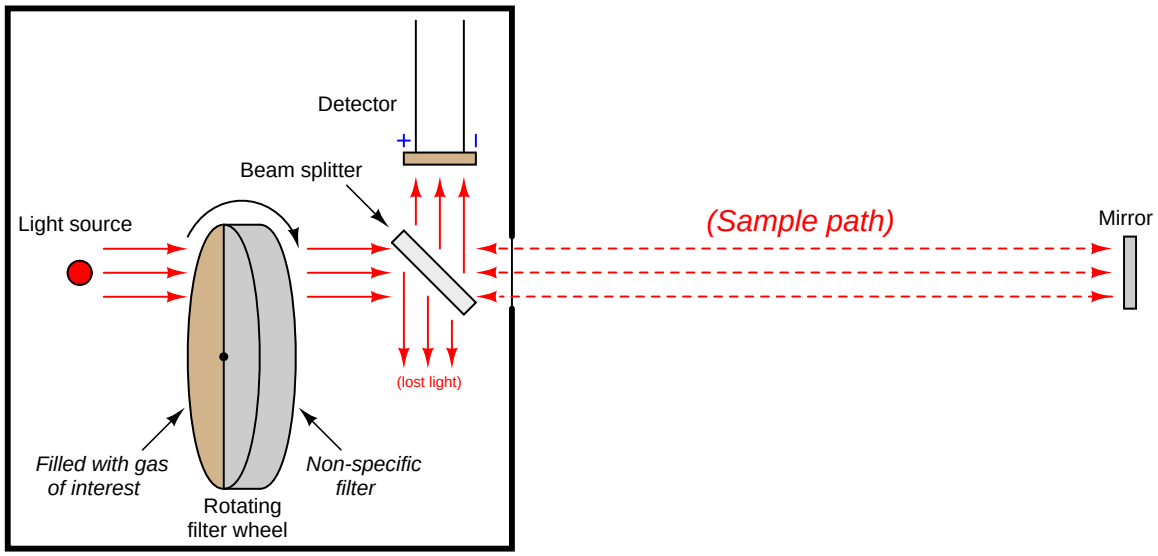
b = Path length of light traveling through the sample

c = Concentration of photon-absorbing substance in the sample
I_0 = Intensity of source (incident) light
I_0 = Intensity of received light after passing through the sample

The longer the path length, the more light will be absorbed by the gas, all other factors being equal. This increases the analyzer's sensitivity to low concentrations, which is especially important when measuring gas concentrations in the low parts-per-million (ppm) or even parts-per-billion (ppb) range.

An example diagram for a GFC analyzer used to measure gas concentrations in open air is shown here:

GFC spectrometer used in open-air measurement



Light passing through the rotating filter wheel strikes a *beam splitter* (a partially-silvered glass plate angled at 45^o) where approximately half the light passes through to the sample space and the other half is lost to reflection. At the far end of the sample space, a full-silvered mirror reflects all the light back to the analyzer, where it strikes the beam splitter again, with approximately half of that light reflecting off at 90^o to reach the detector. With this arrangement, the path length (b in the Beer-Lambert Law) is equal to *twice* the distance between the analyzer and the mirror, since light must travel one way to reach the mirror, then return the same distance back to the analyzer. As you might imagine, extremely long path lengths are easy to achieve with this style of open-air analyzer.

22.4.3 Fluorescence

When a high-energy photon strikes an atom, it may eject one of the lower-level electrons from its shell, leaving a vacancy to be filled by one of the electrons already residing in a higher-level shell. When this happens, the electron filling that lower-level vacancy emits a photon of less energy than the one responsible for ejecting the original electron. Thus, a high-energy photon strikes the atom, and in turn the atom releases a low-energy photon. This phenomenon is known as *fluorescence*.

The relationship between a photon's energy and its frequency (and correspondingly, its wavelength) is a well-defined proportionality of Planck's constant h:

$$E = hf \qquad \text{or} \qquad E = \frac{hc}{\lambda}$$

Where,

E = Energy carried by a single "photon" of light (joules)
h = Planck's constant (6.626×10^{-34} joule-seconds)
f = Frequency of light wave (Hz, or 1/seconds)
c = Velocity of light in a vacuum ($\approx 3 \times 10^{8}$ meters per second)
λ = Wavelength of light (meters)

Therefore, the high-energy photon necessary for ejecting a low-level electron from an atom must be a photon of high frequency (short wavelength), and the low-energy photon emitted by the atom must be one of low frequency (long wavelength).

Photons with enough energy to eject low-level electrons from atoms typically exist in the ultraviolet range and above. The lower-energy photons emitted by the excited atoms often fall within the visible light spectrum. Thus, what we see here is a mechanism for ultraviolet (invisible) light to cause a substance to glow with visible colors.

Fluorescence is commonly used for entertainment purposes in the form of a *black light*: an electrical bulb designed to emit ultraviolet light. Many different organic compounds readily fluoresce under such a light source, producing an eerie glow. Chemical substances present in white paper, certain inks, and certain types of clothing detergents exhibit strong fluorescent properties, as do many bodily fluids[33]. In fact, the presence of fluorescent compounds in paper, inks, and detergents is often intentional, to enhance appearance when viewed in natural sunlight containing ultraviolet light.

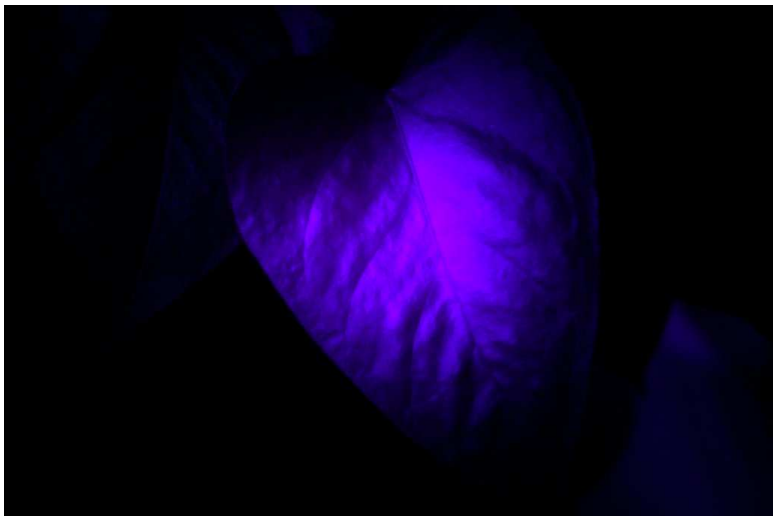
[33] Blood, urine, semen, and various proteins are known to fluoresce in the visible spectrum, making fluorescence a useful tool for crime-scene investigations.

A variety of common food substances fluoresce. Quinine, an ingredient contained in "tonic water," glows yellow-green when exposed to ultraviolet light:

Olive oil is another example of a food substance fluorescing easily under ultraviolet light. In this case, the color of the emitted light is amber in tone:

Molasses fluoresces a deep green color when exposed to ultraviolet light:

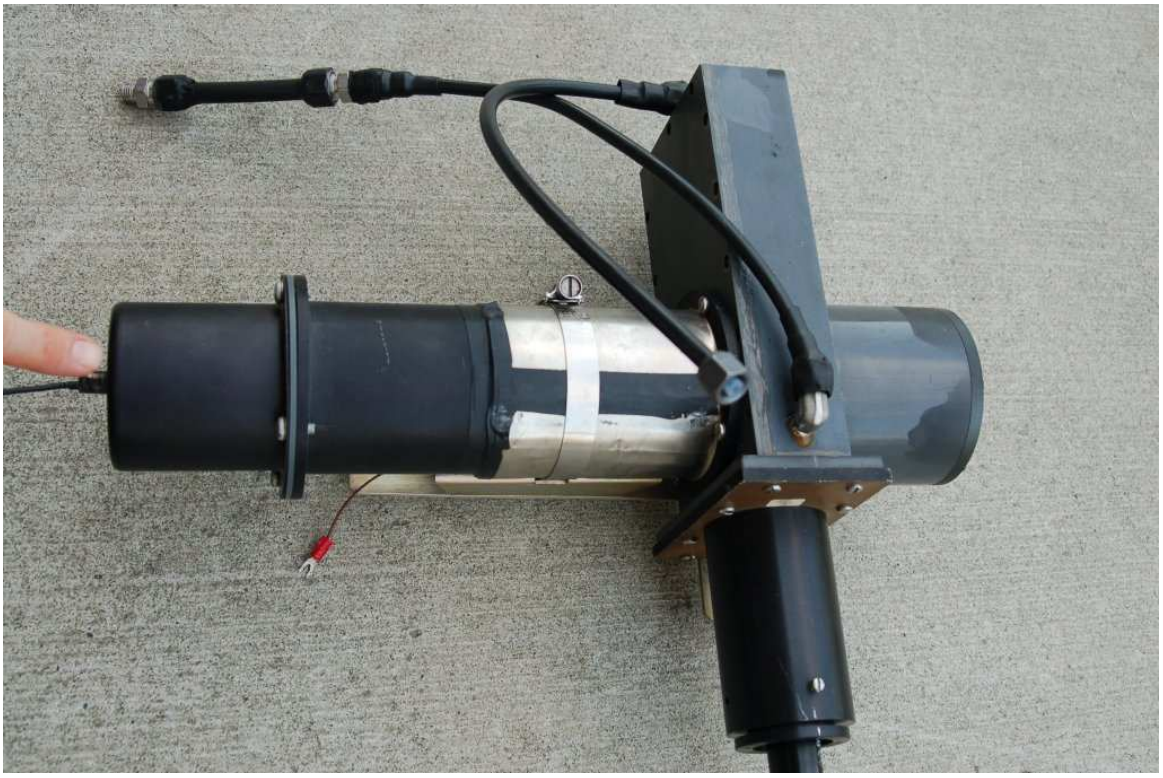
Chlorophyll is an example of a substance (occurring naturally in the tissues of green plants) capable of fluorescence when exposed to ultraviolet light. The color of its fluorescence is red, as shown in this photograph of a house-plant leaf illuminated by a black light:



Fluorescent dyes are often used as "invisible ink," marking items in such a way as to be invisible under normal light, but plainly visible when exposed to concentrated ultraviolet light. Such ink is used to mark modern United States currency, such as this $20 bill. The fluorescent stripe shown in this close-up photograph contains text reading "USA TWENTY":

Not all substances fluoresce as easily as others. If a substance present within an industrial sample happens to fluoresce, and all other substances in the sample stream do not (or at least do not to any significant degree), we may apply fluorescence as an analytical technique for the selective measurement of that substance.

Sulfur dioxide (SO_2) is an atmospheric pollutant formed by the combustion of fuels containing sulfur. This gas also happens to exhibit fluorescence under ultraviolet light. A photograph of the fluorescence chamber taken from a Thermo Electron model 43 sulfur dioxide analyzer appears here:

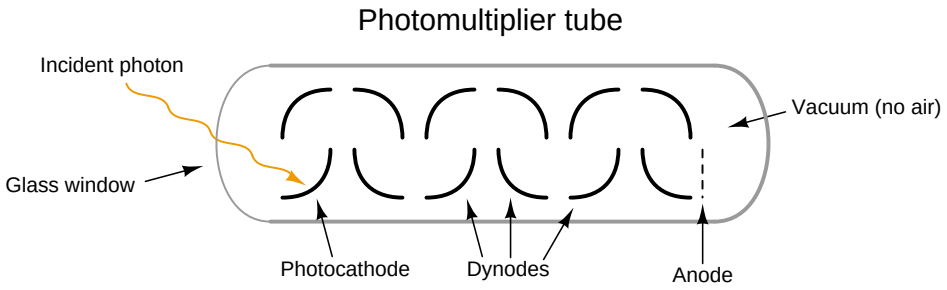


A steady flow of sample gas enters and exits the chamber through the black plastic tubes. Ultraviolet light enters the chamber from a special lamp situated on the right, and a highly sensitive device called a *photomultiplier tube* situated at the bottom detects light emitted when SO_2 molecules inside the chamber fluoresce. The greater the concentration of SO_2 molecules in the gas mixture, the more light will be sensed by the photomultiplier tube for any given amount of ultraviolet light.

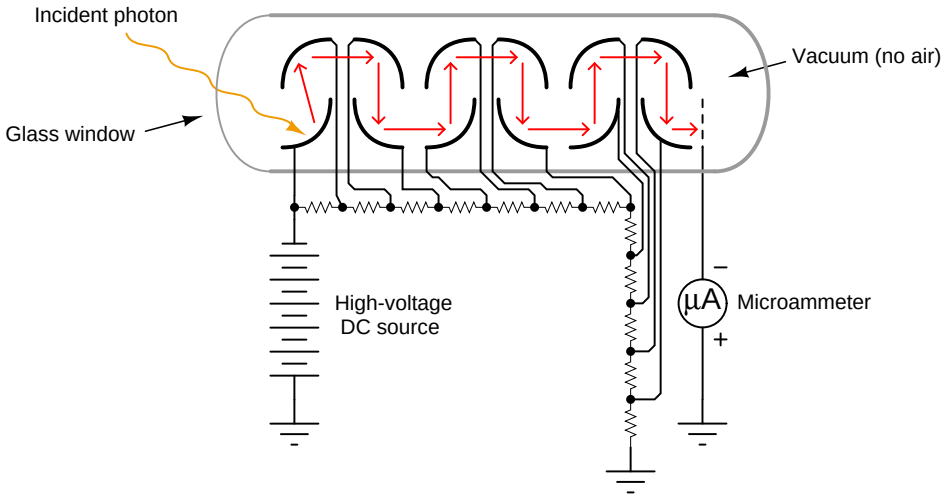
The incident ultraviolet light from the lamp cannot directly reach the photomultiplier tube, because there is no straight-line path from the lamp to the tube, and the interior walls of the chamber are non-reflective. The *only* way for the tube to receive light is if molecules inside the chamber fluoresce when excited by the lamp's ultraviolet light. This ensures the instrument will truly measure fluorescence, and produce a "zero" output when no fluorescent molecules are present.

A close-up view of the ultraviolet emitter shows it to be a gas-discharge lamp. When an oscillating source of high-voltage electricity energizes the electrodes inside this lamp, an arc forms and emits pulsating rays of ultraviolet light:

The photomultiplier tube is a special vacuum tube operating on the principle of the *photoelectric effect*, whereby an incident photon (light particle) of sufficient energy ejects an electron upon striking a metal surface. Light entering a transparent glass window on the photomultiplier tube causes electrons to be emitted from an electrically-charged metal plate called the *photocathode*. Following the photocathode plate are a series of more metal plates called *dynodes*. Each time electrons strike the a dynode plate at high energy, even more electrons are emitted in a process called *secondary emission*. Successively positive electrical potentials applied to each dynode motivate secondary emission as the electrons cascade from dynode to dynode, culminating in a flood of electrons reaching a metal plate at the end of the electrons' path in the tube called an *anode*. A pulse of current measured at the anode signals the tube's reception of each photon:



A simplified photomultiplier tube and power supply circuit are shown here:



In a real instrument, the micro-ammeter would be replaced by an amplifier circuit, producing a strong electrical signal in direct response to received light intensity. In the case of a fluorescence analyzer, the amplifier's output signal becomes a representation of SO_2 molecule concentration inside the chamber.

Like any other type of analyzer technology, we must be aware of potential interfering substances as we use fluorescence to detect the concentration of the species of interest. Not only does sulfur dioxide fluoresce when exposed to ultraviolet light, but so does nitric oxide (NO) and many hydrocarbon components, especially those large hydrocarbon compounds classified as *polynuclear aromatic hydrocarbons* or *PAH*. Unfortunately, both nitric oxide and PAH compounds are produced in some industries where sulfur dioxide is an environmental concern. In order for a fluorescence-based SO_2 analyzer to *only* measure the concentration of sulfur dioxide in a gas stream possibly containing NO and/or PAH compounds, special care must be taken to eliminate the interference.

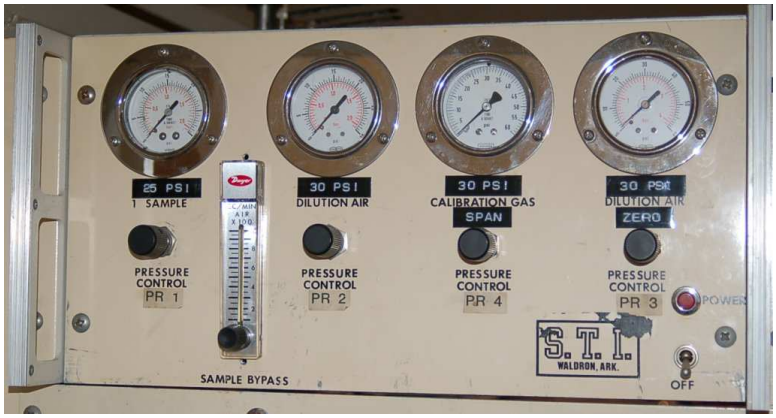
Fortunately for us, nitric oxide happens to fluoresce at a different wavelength than sulfur dioxide gas. This gives us the ability to de-sensitize the instrument to nitric oxide by placing an appropriate optical filter in front of the photomultiplier tube. This filter blocks light wavelengths emitted by the fluorescence of nitric oxide, allowing the photomultiplier tube to only "see" the light emitted by the fluorescence of sulfur dioxide.

The light from hydrocarbon compound fluorescence is not as easy to eliminate with optical filtering, and so this analyzer takes care of the PAH interference problem by *physically* filtering out hydrocarbon gas molecules prior to the sample entering the fluorescence chamber using a device called a *kicker*. The "kicker" is a form of molecular sieve, separating the hydrocarbon molecules from the other molecules in the sample stream.

After processing by the electronic circuits of the analyzer, the photomultiplier tube's output signal becomes a representation of SO_2 concentration, displayed on an analog meter movement:

As indicated by the selector switch below the meter face, this instrument has three different display ranges: 0 to 0.5 ppm (*parts per million*), 0 to 1.0 ppm, and 0 to 5.0 ppm. A different selector switch on the left-hand side of the control panel operates solenoid valves allowing either the process sample gas or one of two different calibration gases to enter the analyzer. The "zero" calibration gas contains no sulfur dioxide at all, thus providing a base-line reference for adjusting the 0% point of the analyzer. The "span" calibration gas contains a precise mixture of sulfur dioxide and some non-fluorescing carrier gas, to serve as a chemical reference for some point near the analyzer's upper range limit. These calibration gases are commercially available from chemical laboratories, with instrument technicians commonly referring to them as *zero gas* and *span gas*. Of course, the composition of any "zero" or "span" gas depends entirely on the type of analytical instrument. What may suffice as a span gas for this sulfur dioxide analyzer would certainly not suffice as a span gas for a multi-component chromatograph or for an NDIR analyzer configured to measure carbon monoxide.

Pressure regulators ensure proper gas flow conditioning in and out of the analyzer. A vacuum pump (not shown in any of the photographs) draws sample gas through the analyzer and provides the necessary differential pressure for the hydrocarbon "kicker" to work:



22.4.4 Chemiluminescence

Recall that an *exothermic* chemical reaction is one that releases a net sum of energy, as opposed to an *endothermic* reaction which requires a greater input of energy than it releases. Combustion is a common class of exothermic reactions, with the released energy being very obviously in the forms of heat and light, with heat being the predominant form.

Some exothermic reactions release energy primarily in the form of light rather than heat. The general term for this effect is *chemiluminescence*. A striking example of this reaction found in nature is the "cold" light emitted by North American species of firefly. In this small insect, a chemical reaction intermittently takes place emitting significant amounts of light but insignificant amounts of heat.

Certain industrial compounds engage in chemiluminescent reactions, and this phenomenon may be used to measure the concentration of those compounds. One such compound is nitric oxide (NO), an atmospheric pollutant formed by high-temperature combustion with air as the oxidizer[34].

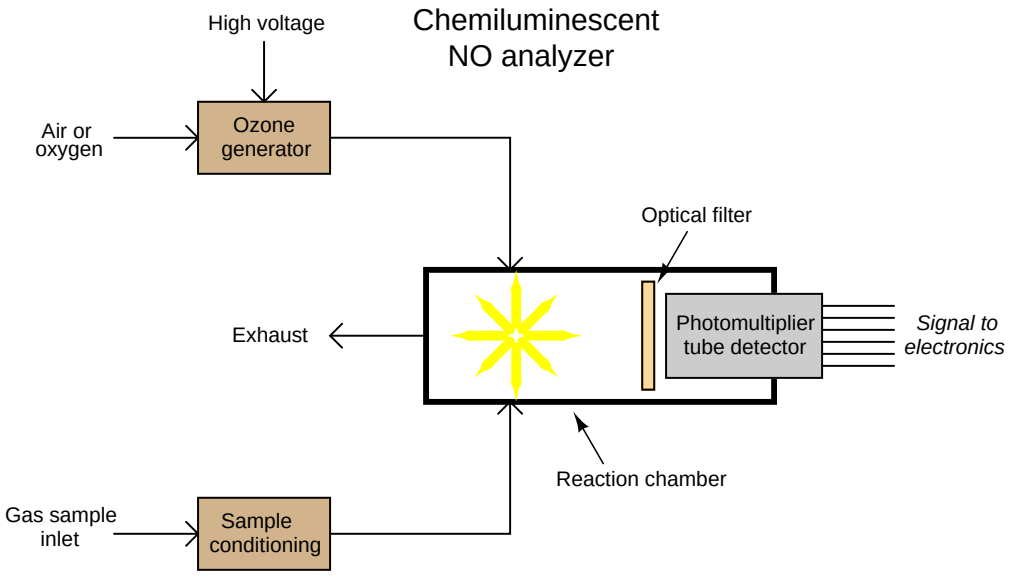
A spontaneous chemical reaction between nitric oxide and *ozone* (an unstable molecule formed of three oxygen atoms: O_3) is known to produce chemiluminescence:

$$NO + O_3 \rightarrow NO_2 + O_2 + \text{ light}$$

Although this process of generating light is quite inefficient (only a small fraction of the NO_2 molecules formed by this reaction will emit light), it is predictable enough to be used as a quantitative measurement method for nitric oxide gas. Ozone gas is very easy to produce on demand, by discharging an electric arc in the presence of oxygen.

[34] Combustion is primarily a reaction between carbon and/or hydrogen atoms in fuel, and oxygen atoms in air. However, about 78% of the air (by volume) is nitrogen, and only about 20.9% is oxygen, which means a lot of nitrogen gets pulled in with the oxygen during combustion. Some of these nitrogen atoms combine with oxygen atoms under the high temperature of combustion to form various oxides of nitrogen.

A simplified diagram for a chemiluminescent nitric oxide gas analyzer appears here:

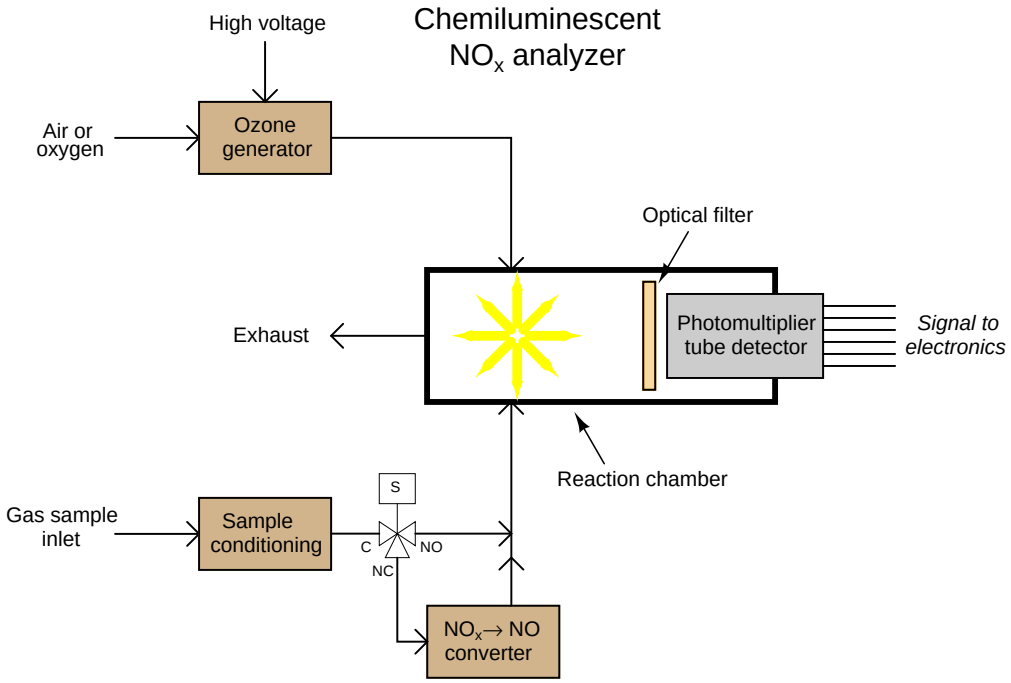


As with many optical analyzers, a photomultiplier tube serves as the light-detecting sensor, generating an electrical signal in proportion to the amount of light observed inside the reaction chamber. The higher the concentration of NO molecules in the sample gas stream, the more light will be emitted inside the reaction chamber, resulting in a stronger electrical signal produced by the photomultiplier tube.

Although this instrument readily measures the concentration of nitric oxide (NO), it is insensitive to other oxides of nitrogen (NO_2, NO_3, etc.). Normally, we would consider this selectivity to be a good thing, because it would eliminate interference problems from these other gases. However, as it so happens, these other oxides of nitrogen are every bit as significant as nitric oxide (NO) from the perspective of air pollution, and when we measure nitric oxide for pollution monitoring purposes, we usually *also* wish to measure these other oxides[35] in combination.

[35] The measures used to mitigate nitric oxide emissions are the same measures used to mitigate the other oxides of nitrogen: reduce combustion temperature, and/or reduce the NO_x compounds to elemental nitrogen by mixing the combustion exhaust gases with ammonia (NH_3) in the presence of a catalyst. So here we have a case where we really don't care to distinguish NO from NO_x: we want to measure it *all*.

In order to use chemiluminescence to measure *all* oxides of nitrogen, we must chemically convert the other oxides into nitric oxide (NO) before the sample enters the reaction chamber. This is done in a special module of the analyzer called a *converter*:



A three-way solenoid valve is shown in this diagram, providing a means to bypass the converter so the analyzer only measures nitric oxide content in the sample gas. With the solenoid valve passing all the sample through the converter, the analyzer responds to *all* oxides of nitrogen (NO_x) and not just nitric oxide (NO).

One simple way to achieve the $NO_x \rightarrow NO$ chemical conversion is to simply heat the gas to a high temperature, around 1300 ^{o}F. At this temperature, the molecular structure of NO is favored over more complex oxides. A disadvantage of this technique is that those same high temperatures also have a tendency to convert other compounds of nitrogen such as ammonia (NH_3) into nitric oxide, thereby creating an unintended interference species[36].

[36]This particular interference compound is especially problematic if we are using the analyzer to *control* the NO_x concentration in the exhaust of a combustion process, and the manipulated variable for the NO_x control loop is pure ammonia injected into the exhaust. Un-reacted ammonia sampled by the analyzer will be falsely interpreted as NO_x, rendering the measurement meaningless, and therefore making control virtually impossible.

An alternative $NO_x \rightarrow NO$ conversion technique is to use a metallic reactant in the converter to remove the extra oxygen atoms from the NO_2 molecules. One such metal that works well for this purpose is molybdenum (Mo) heated to the comparatively low temperature of 750 ^oF, which is too low to convert ammonia into nitric oxide. The reaction of NO_2 converting to NO is as follows:

$$3NO_2 + Mo \rightarrow MoO_3 + 3NO$$

Other oxides (such as NO_3) convert in similar fashion, leaving their excess oxygen atoms bound to molybdenum atoms and becoming nitric oxide (NO). The only difference between these reactions and the one shown for NO_2 is the proportional (stoichiometric) ratios between molecules.

As you can see from the reaction, the molybdenum metal is converted into the compound molybdenum trioxide over time, requiring periodic replacement. The rate at which the molybdenum metal depletes inside the converter depends on the sample flow rate and the concentration of NO_2.

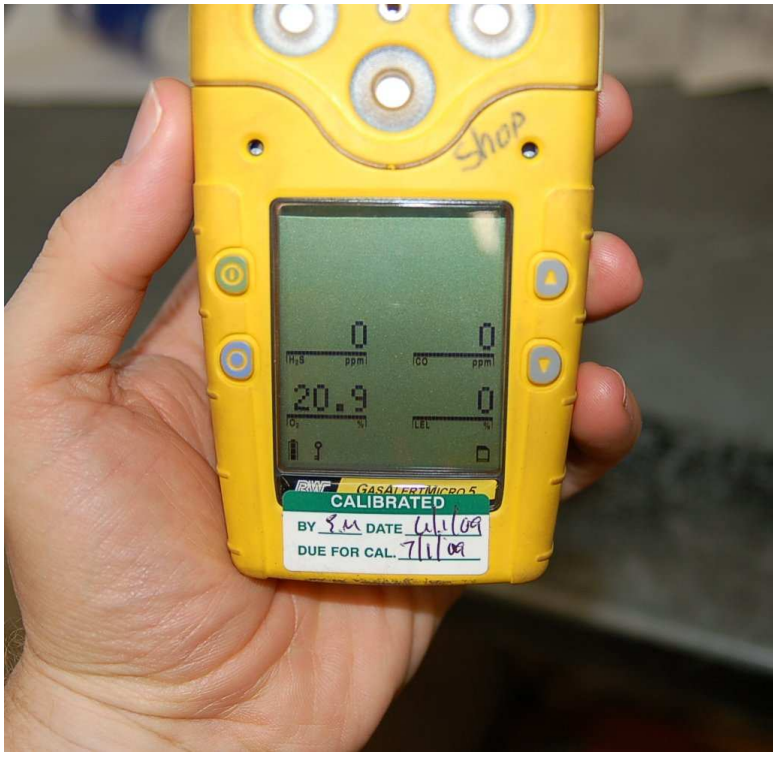
22.5 Safety gas analyzers

Process analyzers measure the concentration of specific substances for the purpose of measuring and/or controlling those concentrations in a process stream. Safety analyzers detect the presence of dangerous concentrations of specific substances to warn personnel of threats to life or health. While there is virtually no end to the different types of process analyzers in existence, with new analyzer types invented to meet the needs of process industries, safety analyzers are rather restricted to those substances known to pose health hazards to human beings.

Safety analyzers are designed for fast response, rugged construction, and ease of portability. As such, they are usually not as accurate or as sensitive to slight changes in concentration as process analyzers. The sensing technologies used in safety analyzers are often very different from those used in process analyzers. You will never, for example, see an NDIR instrument with a Luft-style detector used for portable safety applications[37]. The high accuracy of a Luft-style NDIR instrument is not necessary for safety, and the bulk (and fragility) of such an instrument makes it completely impractical as a portable device.

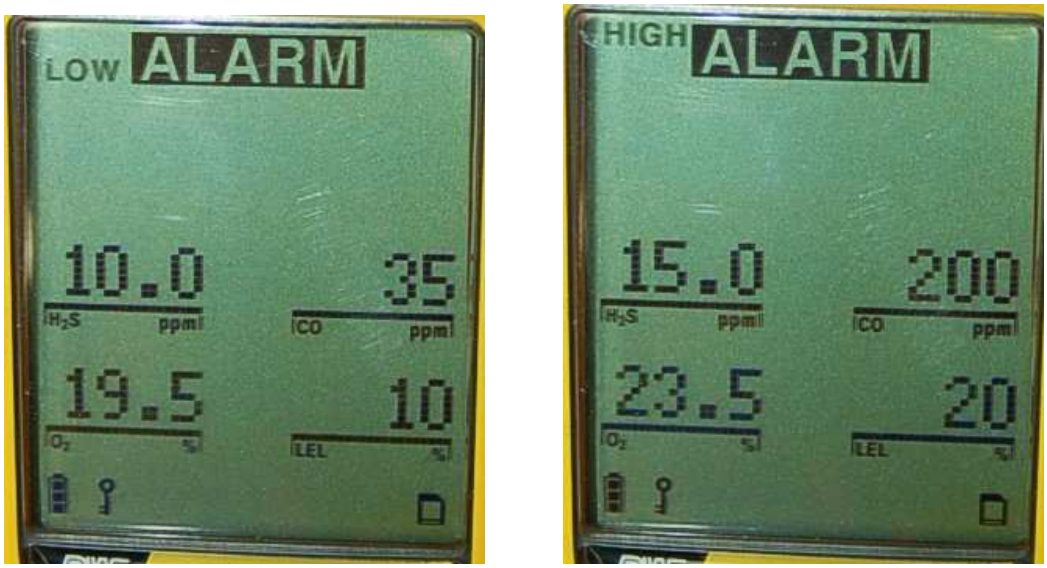
[37] Interestingly, there is a documented case of an NDIR "Luft" analyzer being used as a safety monitor for carbon monoxide, ranged 0 to 0.1% (0-1000 ppm), at one of I.G. Farbenindustrie's chemical plants in Germany during the 1940's. This was definitely not a *portable* analyzer, but rather stationary-mounted in a process unit where high concentrations of carbon monoxide gas existed in the pipes and reaction vessels. The relatively fast response and high selectivity of the NDIR technology made it an ideal match for the application, considering the other (more primitive) methods of carbon monoxide gas detection which could be "fooled" by hydrogen, methane, and other gases.

This photograph shows a hand-held safety gas monitor, used to detect four different gases with known hazard levels (oxygen, carbon monoxide, combustibles, and hydrogen sulfide). A technician working in hazardous environments would wear one of these at all times, listening for the audible warning tone generated by the device if any of its pre-set limits is exceeded:



Most portable gas analyzers such as this employ electrochemical sensing "cells" generating a small electric current when exposed to the gas of interest. Such technologies may not always be the most accurate or the most sensitive, but their characteristics are well suited for portable applications where they will be exposed to vibration and must operate on battery power.

A close-up photograph taken of the monitor's alarm thresholds reveals the relative concentrations of four gases monitored by the device. Two distinct alarm levels – "low" (alert) and "high" (danger) – exist to warn the user of threats:



The two most dangerous gases detected by this device – hydrogen sulfide (H_2S) and carbon monoxide (CO) – are measured in units of *parts per million*, or *ppm*. The next two detected gases – oxygen (O_2) and combustibles ("LEL") – are measured in units of percent (which may be thought of as *parts per hundred*).

Like all safety-related devices, portable gas analyzer require regular "proof-testing" and calibration. In order to accurately check the response of a gas analyzer, it must be exposed to gases of known concentrations. Special mixtures of "test gas" for safety analyzers are available from chemical supply companies, the next photograph showing the certified concentrations of different gases contained inside the pressurized test cylinder:

22.5.1 Oxygen gas

Most living things require oxygen to survive. The oxygen you breathe combines with nutrients from the food you eat to produce energy in a form usable by your body. If you are deprived of oxygen, your body very quickly shuts down, much like a fire dies when starved of oxygen (and for approximately the same reason). Ambient air is approximately 20.9% oxygen by volume, the majority of air (about 78% by volume) being nitrogen.

The oxygen content of air may be reduced by combustion (which combines oxygen with flammable substances to produce carbon dioxide and water vapor) or by displacement by a denser gas (such as propane) in a low-lying area or by any gas in sufficient quantity filling an enclosed area.

A modern oxygen sensor technology for safety applications is the *micro fuel cell*, generating a measurable electric current in the presence of oxygen by the oxidation of a self-contained fuel source. In many sensors, the fuel is pure lead (Pb), with the resulting chemical reaction producing lead oxide (PbO):

$$O_2 + 2Pb \rightarrow 2PbO$$

Fuel cell sensors are relatively rugged, accurate, and self-powering, enabling their use in portable oxygen analyzers. Due to their principle of operation, where an internal fuel is slowly oxidized over time, these sensors have a rather limited life and therefore must be periodically replaced.

An interesting and useful technique for testing the operation of an oxygen safety sensor is to exhale on the sensor, watching for a decrease in oxygen content to 15% or below. This testing technique makes use of the fact that your body extracts oxygen from the air, such that your exhaled breath contains less oxygen than it did when inhaled. Therefore, your own body acts as a crude "calibration gas" source for an oxygen safety analyzer.

22.5.2 Lower explosive limit (LEL)

The minimum concentration of a flammable gas in air capable of igniting is called the *Lower Explosive Limit*, or *LEL*. This limit varies with the type of gas and with the oxygen concentration of the air in which the flammable gas is mixed. Sensors designed to detect the dangerous presence of combustible gases are therefore called "LEL sensors."

LEL monitors are used whenever there is a high probability of explosive gases present in the air. These areas are referred to as *classified* areas in industry, and are precisely defined for safety engineering purposes. Classified areas harboring explosive gases or vapors are deemed *Class I* areas, with different "Group" categories delineating the specific gas or vapor types involved. For more information on classified areas, refer to section 29.1.1, beginning on page 1646.

Gases and vapors are not the only substances with the potential to explode in sufficient concentration. Certain dusts (such as grain) and fibers (such as cotton) may also present explosion hazards if present in sufficient quantity. Unfortunately, the majority of analytical technologies used to monitor lower explosive limits for safety purposes only function with gases and vapors (Class I), not dusts or fibers (Class II and Class III, respectively).

Popular sensor technologies used to detect the presence of combustibles in air include the following:

- Catalytic bead
- Thermocouple
- infrared
- Flame ionization

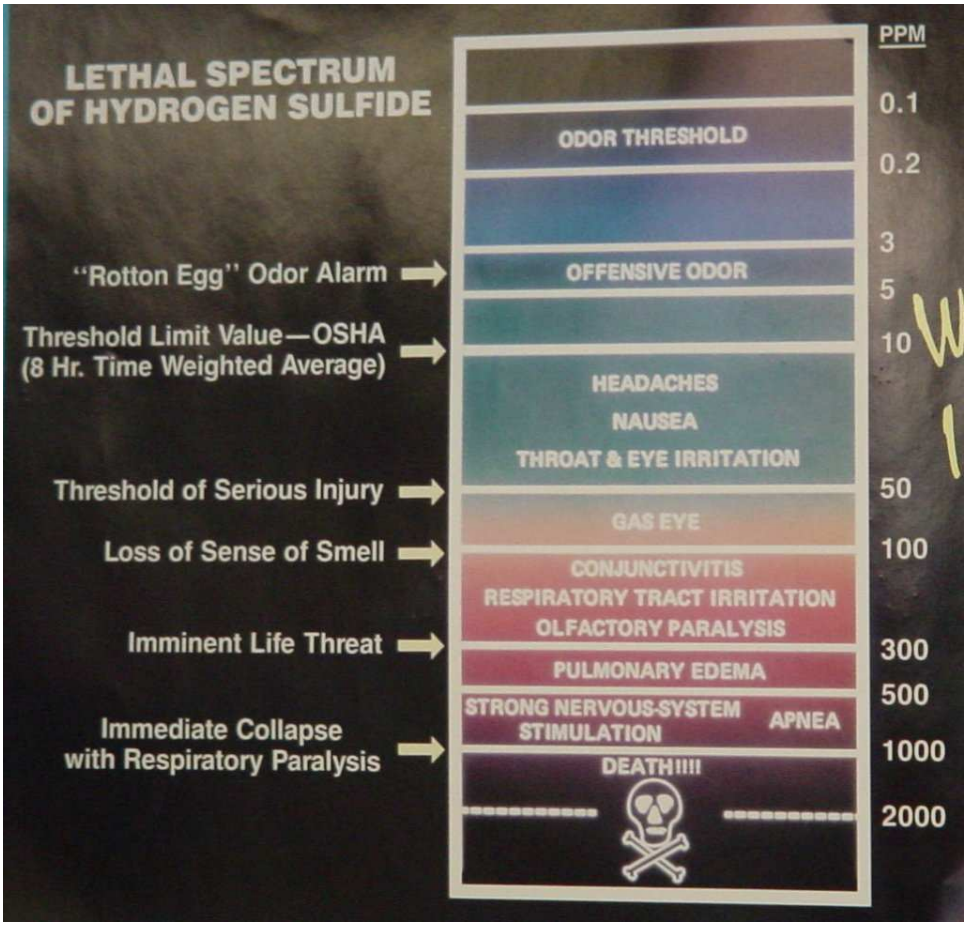
Catalytic bead and thermocouple sensors both function on the principle of heat generated during combustion. Air potentially containing a concentration of flammable gases or vapors is passed near a heated element, and any combustion occurring at that point will cause the local temperature to immediately rise. These sensors must be designed in such a way they will *not* initiate an explosion, but merely combust the sample in a safe and measurable manner. Like micro fuel cell oxygen sensors, these sensors may be manufactured in sufficiently small and rugged packages to enable their use as portable LEL sensors.

infrared analyzers exploit the phenomenon of infrared (IR) light absorption by certain types of flammable gases and vapors. A beam of infrared light passed across a sample of air will diminish in intensity if significant concentrations of the combustible substance exist in that sample. Measuring this attenuation provides an indirect measurement of explosive potential. A major disadvantage of this technique is that many non-flammable gases and vapors also absorb IR light, including carbon dioxide and water vapor. In order to successfully reject these non-flammable substances, the analyzer must use very specific wavelengths of IR light, tuned to the specific substances of interest (and/or wavelengths tuned specifically to the substances of non-interest, as a compensating reference signal for the wavelengths captured by both the substances of interest and the substances of non-interest).

Flame ionization sensors work on the same principle as FIDs for chromatographs: a non-ionizing flame (usually fueled by hydrogen gas) will generate detectable ions only in the presence of air samples containing an ionizing fuel (such as a hydrocarbon gas). Of course, this form of LEL sensor is useless to detect hydrogen gas.

22.5.3 Hydrogen sulfide gas

Hydrogen Sulfide (H_2S) is a highly toxic gas, with a pungent "rotten eggs" odor at low concentrations but no visible color. At higher concentrations, the gas acts as a nerve agent to de-sensitize human smell, so that it seems odorless. Its paralytic effect on smell extends to more important bodily functions such as breathing, causing rapid loss of consciousness and asphyxiation. A photograph of a safety chart (taken at a wastewater treatment facility) shows just how toxic hydrogen sulfide gas is:



Note how concentrations in the *parts per million* range are hazardous, and how little H_2S concentration is required to paralyze one's sense of smell. Hydrogen sulfide also happens to be flammable, its LEL value in air being 4.3%. However, the toxic properties of the gas are generally the more pressing concern when released into the atmosphere. Another hazardous property of hydrogen sulfide is its density: 1.18 times that of air. This means H_2S gas will tend to collect in low areas such as pits, electrical vaults, and empty underground storage vessels.

The principal source of hydrogen sulfide gas is anaerobic (oxygen-less) decomposition of organic matter. Sewage treatment facilities, pulp mills, and oil refineries generate H_2S gas in significant

quantity, and so employees at such facilities must be continually aware of the associated hazards.

One of the most popular analytical sensing technologies for H_2S gas appropriate for portable monitoring includes an electro-chemical reaction cell similar in principle to the micro fuel cells used to detect oxygen concentrations. Hydrogen sulfide gas entering such a cell engages in a specific chemical reaction, creating a small electrical current proportional to the gas concentration. Like oxygen-sensing fuel cells, these chemical cells also have limited lives and must be routinely replaced.

22.5.4 Carbon monoxide gas

Carbon monoxide (CO) gas is a colorless, odorless, and toxic gas principally generated by the incomplete combustion of carbon-based fuels. The mechanism of its toxicity to people and animals is the preferential binding of CO gas molecules to the hemoglobin in blood. At significant concentrations of carbon monoxide gas in air, the hemoglobin in your blood latches on to CO molecules instead of oxygen (O_2) molecules, and remains bound to the hemoglobin, preventing it from transporting oxygen. The result is that your blood rapidly loses its oxygen-carrying capacity, and your body asphyxiates from within. Like hydrogen sulfide, carbon monoxide is also flammable (LEL = 4%), but its toxic properties are generally the larger concern when released into the atmosphere.

Carbon monoxide is not to be confused with carbon *dioxide* (CO_2) gas, which is almost completely inert to the human body. Carbon dioxide is principally produced by *complete* combustion of carbon-based fuels. Its only safety hazard potential is the capacity to displace breathable air in an enclosed area if rapidly released in large volumes.

Combustion burners operating on carbon-based fuels may produce excess carbon monoxide if operating at too rich an air/fuel mixture. Even when adjusted optimally, there will always be some carbon monoxide present in the exhaust. This makes high CO concentrations possible where burners operate in enclosed areas.

Some industrial processes such as catalytic cracking in the petroleum refining industry generate huge amounts of carbon monoxide, but these extremely high concentrations are normally present only within the process piping and vessels, not released to atmosphere. Nevertheless, personnel working near such processes must wear portable CO gas safety monitors at all times to warn of leaks.

Carbon monoxide may be sensed by an electrochemical cell, using iodine pentoxide as the reacting compound. The balanced chemical reaction is as follows:

$$5CO + I_2O_5 \rightarrow 5CO_2 + I_2$$

The strength of the electric current produced by the cell indicates the concentration of carbon monoxide gas.

22.5.5 Chlorine gas

Chlorine (Cl_2) gas is a strong-odored, toxic gas used as a biological disinfectant, bleaching agent, and as an oxidizer in many industrial processes. Colorless in low concentrations, it may appear green in color when mixed in very high concentrations with air. Chlorine is highly reactive, presenting a distinct hazard to mucous membranes (eyes, nose, throat, lungs) by creating hypochlorous acid (HOCl) and hydrochloric acid (HCl) upon contact with water:

$$Cl_2 + H_2O \rightarrow HOCl + HCl$$

The following table correlates levels of chlorine gas concentration in ambient air with degree of hazard. Note the unit of measurement for chlorine concentration in air – *parts per million* (ppm). Bear in mind that one part per million is equivalent to just 0.0001 percent:

Concentration in air	**Hazard**
1 ppm to 3 ppm	Mild mucous membrane irritation
5 ppm to 15 ppm	Upper respiratory tract irritation
30 ppm	Immediate chest pain, cough, and difficulty breathing
40 ppm to 60 ppm	Toxic pneumonitis and pulmonary edema
430 ppm	Lethal over 30 minutes
1000 ppm (0.1%)	Lethal within a few minutes

Water and wastewater treatment operations frequently[38] use chlorine for disinfection of water. Pulp mills use either chlorine or chlorine compounds as a bleaching agent to whiten wood pulp.

[38] Some water treatment facilities use powerful ultraviolet lamps to disinfect water without the use of chemicals. Some potable (drinking) water treatment plants use ozone gas (O_3) as a disinfectant, which is generated on-site from atmospheric oxygen. A disadvantage to both chlorine-free approaches for drinking water is that neither one provides lasting disinfection throughout the distribution and storage system to the same degree that chlorine does.

Chlorine may be generated on site by the electrolytic decomposition of salt (sodium chloride – NaCl), or delivered in cylindrical pressure vessels in liquid form as shown here at a large wastewater treatment facility:

References

"Automated Measuring System Technologies", Best Practice brochure, Cleaner Fossil Fuels Programme, document BPB008, DTI, 2004.

Boylestad, Robert L., *Introductory Circuit Analysis*, 9th Edition, Prentice Hall, Upper Saddle River, NJ, 2000.

Carroll, Grady C., *Industrial Process Measuring Instruments*, McGraw-Hill Book Company, Inc., New York, NY, 1962.

Chu, P.M.; Guenther, F.R.; Rhoderick, G.C.; Lafferty, W.J.; "The NIST Quantitative Infrared Database", Journal of Research of the National Institute of Standards and Technology, Volume 104, Number 1, Gaithersburg, MD, January-February 1999.

Fribance, Austin E., *Industrial Instrumentation Fundamentals*, McGraw-Hill Book Company, New York, NY, 1962.

Gregory, C.H.; Appleton, H.B.; Lowes, A.P.; Whalen, F.C.; *Instrumentation and Control in the German Chemical Industry*, Mapleton House, Brooklyn, NY, 1947.

"Investigation Report – Chlorine Release", Report number 2002-04-I-MO, U.S. Chemical Safety and Hazard Investigation Board, Washington DC, 2003.

Jernigan, J. Ron, "Chemiluminescence NO_x and GFC NDIR CO Analyzers For Low Level Source Monitoring", Thermo Environmental Instruments, Franklin, MA.

Kohlmann, Frederick J., "What Is pH, And How Is It Measured?", Hach Company, 2003.

Kume, Hidehiro, *Photomultiplier Tube – Principle to Application*, Hamamatsu Photonics K.K., 1994.

Lavigne, John R., *Instrumentation Applications for the Pulp and Paper Industry*, Miller Freeman Publications, Foxboro, MA, 1979.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Measurement and Analysis Volume I*, Fourth Edition, CRC Press, New York, NY, 2003.

Novak, Joe, *What Is Conductivity, And How Is It Measured?*, Hach Company, 2003.

Pauling, Linus, *General Chemistry*, Dover Publications, Inc., Mineola, NY, 1988.

Scott, Raymond P.W., *Gas Chromatography*, Library4Science, LLC, 2003.

Scott, Raymond P.W., *Gas Chromatography Detectors*, Library4Science, LLC, 2003.

Scott, Raymond P.W., *Liquid Chromatography*, Library4Science, LLC, 2003.

Scott, Raymond P.W., *Liquid Chromatography Detectors*, Library4Science, LLC, 2003.

Scott, Raymond P.W., *Principles and Practice of Chromatography*, Library4Science, LLC, 2003.

Sherman, R.E.; Rhodes, L.J., *Analytical Instrumentation: practical guides for measurement and control*, ISA, Research Triangle Park, NC, 1996.

Shinskey, Francis G., *pH and pION Control in Process and Waste Streams*, John Wiley & Sons, New York, NY, 1973.

"Standard Operating Procedures – Thermo Environmental Instruments Model 43C Trace Level Pulsed Fluorescence Sulfur Dioxide Analyzer", version 2.0, Environmental Protection Agency, Research Triangle Park, NC, 2009.

Theory and Practice of pH Measurement, PN 44-6033, Rosemount Analytical, 1999.

"XSTREAM Gas Analyzer Series Instruction Manual", document HASX2E-IM-HS, Rosemount Analytical, 2009.

Chapter 23

Machine vibration measurement

Unlike most process measurements, the measurement of a rotating machine's *vibration* is primarily for the benefit of the process equipment rather than the process itself. Vibration monitoring on an ammonia vapor compressor, for instance, may very well be useful in extending the operating life of the compressor, but it offers little benefit to the control of the ammonia vapor.

Nevertheless, the prevalence of machine vibration measurement technology is so widespread in the process industries that it cannot be overlooked by the instrument technician. Rotating machinery equipped with vibration sensors are often controlled by *protection* equipment designed to automatically shut down the machine in the event of excessive vibration. The configuration and maintenance of this protection equipment, and the sensors feeding vibration data to it, is often the domain of instrument technicians.

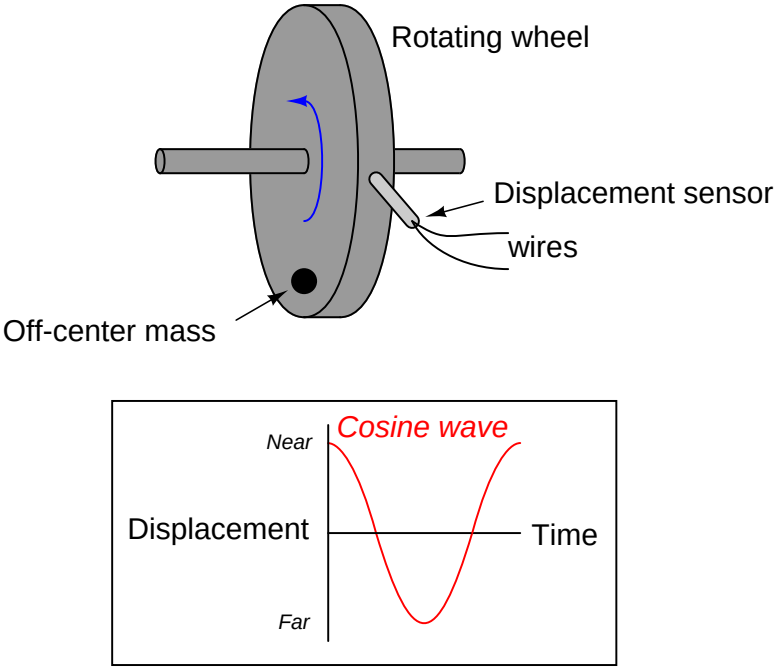
23.1 Vibration physics

One very convenient feature of waves is that their properties are universal. Waves of water in the ocean, sound waves in air, electronic signal waveforms, and even waves of mechanical vibration may all be expressed in mathematical form using the trigonometric *sine* and *cosine* functions. This means the same tools (both mathematical and technological) may be applied to the analysis of different kinds of waves. A strong example of this is the *Fourier Transform*, used to determine the frequency spectrum of a waveform, which may be applied with equal validity to any kind of wave[1].

[1] The "spectrum analyzer" display often seen on high-quality audio reproduction equipment such as stereo equalizers and amplifiers is an example of the Fourier Transform applied to music. This exact same technology may be applied to the analysis of a machine's vibration to indicate sources of vibration, since different components of a machine tend to generate vibratory waves of differing frequencies.

23.1.1 Sinusoidal vibrations

If a rotating wheel is unbalanced by the presence of an off-center mass, the resulting vibration will take the form of a cosine wave as measured by a displacement (position) sensor near the periphery of the object (assuming an angle of zero is defined by the position of the displacement sensor). The displacement sensor measures the air gap between the sensor tip and the rim of the spinning wheel, generating an electronic signal (most likely a voltage) directly proportional to that gap:



Since the wheel's shaft "bows" in the direction of the off-center mass as it spins, the gap between the wheel and the sensor will be at a minimum at 0^o, and maximum at 180^o.

We may begin to express this phenomenon mathematically using the cosine function:

$$x = D\cos\omega t + b$$

Where,

x = Displacement as measured by sensor at time t
D = Peak displacement amplitude
ω = Angular velocity (typically expressed in units of radians per second)
b = "Bias" air gap measured with no vibration
t = Time (seconds)

Since the cosine function alternates between extreme values of +1 and -1, the constant D is necessary in the formula as a coefficient relating the cosine function to peak displacement. The cosine function's argument (i.e. the angle given to it) deserves some explanation as well: the product ωt is the multiple of angular velocity and time, angular velocity typically measured in radians per second and time typically measured in seconds. The product ωt, then, has a unit of radians. At time=0 (when the mass is aligned with the sensor), the product ωt is zero and the cosine's value is +1.

For a wheel spinning at 1720 RPM (approximately 180.1 radians per second), the angle between the off-center mass and the sensor will be as follows:

Time	Angle (radians)	Angle (degrees)	$\cos\omega t$
0 ms	0 rad	0^o	+1
8.721 ms	$\frac{\pi}{2}$ rad	90^o	0
17.44 ms	π rad	180^o	-1
26.16 ms	$\frac{3\pi}{2}$ rad	270^o	0
34.88 ms	0 rad	360^o or 0^o	+1

We know from physics that *velocity* is the time-derivative of displacement. That is, velocity is defined as the rate at which displacement changes over time. Mathematically, we may express this relationship using the calculus notation of the derivative:

$$v = \frac{dx}{dt} \qquad \text{or} \qquad v = \frac{d}{dt}(x)$$

Where,

v = Velocity of an object
x = Displacement (position) of an object
t = Time

Since we happen to know the equation describing displacement (x) in this system, we may differentiate this equation to arrive at an equation for velocity:

$$v = \frac{dx}{dt} = \frac{d}{dt}(D\cos\omega t + b)$$

Applying the differentiation rule that the derivative of a sum is the sum of the derivatives:

$$v = \frac{d}{dt}(D\cos\omega t) + \frac{d}{dt}b$$

Recall that D, ω, and b are all constants in this equation. The only variable here is t, which we are differentiating with respect to. We know from calculus that the derivative of a simple cosine function is a negative sine ($\frac{d}{dx}\cos x = -\sin x$), and that the presence of a constant multiplier in the cosine's argument results in that multiplier applied to the entire derivative[2] ($\frac{d}{dx}\cos ax = -a\sin ax$). We also know that the derivative of any constant is simply zero ($\frac{d}{dx}C = 0$), which eliminates the b term:

$$v = -\omega D\sin\omega t$$

What this equation tells us is that for any given amount of peak displacement (D), the velocity of the wheel's "wobble" increases linearly with speed (ω). This should not surprise us, since we know an increase in rotational speed would mean the wheel displaces the same vibrating distance in less time, which would necessitate a higher velocity of vibration.

We may take the process one step further by differentiating the equation again with respect to time in order to arrive at an equation describing the vibrational *acceleration* of the wheel's rim, since we know acceleration is the time-derivative of velocity ($a = \frac{dv}{dt}$):

$$a = \frac{dv}{dt} = \frac{d}{dt}(-\omega D\sin\omega t)$$

From calculus, we know that the derivative of a sine function is a cosine function ($\frac{d}{dx}\sin x = \cos x$), and the same rule regarding constant multipliers in the function's argument applies here as well ($\frac{d}{dx}\sin ax = a\cos ax$):

$$a = -\omega^2 D\cos\omega t$$

What this equation tells us is that for any given amount of peak displacement (D), the acceleration of the wheel's "wobble" increases with the *square* of the speed (ω). This is of great importance to us, since we know the lateral force imparted to the wheel (and shaft) is proportional to the lateral acceleration and also the mass of the wheel, from Newton's Second Law of Motion:

$$F = ma$$

[2] This rule makes intuitive sense as well: if a sine or cosine wave increases frequency while maintaining a constant peak-to-peak amplitude, the rate of its rise and fall *must* increase as well, since the higher frequency represents less time (shorter period) for the wave to travel the same amplitude. Since the derivative is the *rate of change* of the waveform, this means the derivative of a waveform must increase with that waveform's frequency.

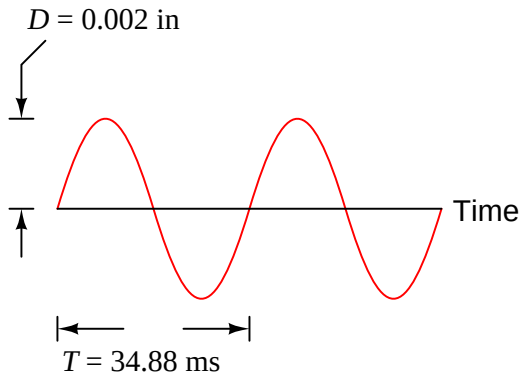
Therefore, the vibrational force experienced by this wheel grows rapidly as rotational speed increases:

$$F = ma = -m\omega^2 D \cos \omega t$$

This is why vibration can be so terribly destructive to rotating machinery. Even a small amount of lateral displacement caused by a mass imbalance or other effect may generate enormous forces on the rotating part(s), and these forces grow with the square of the rotating speed (e.g. doubling the speed quadruples the force; tripling the speed increases force by *9 times*). Worse yet, these calculations assume a constant displacement (D), which we know will also increase with speed owing to the increased centrifugal force pulling the off-center mass away from the shaft centerline. In practice, doubling or tripling an imbalanced machine's speed may multiply vibrational forces well in excess of four or nine times, respectively.

In the United States, it is customary to measure vibrational displacement (D) in units of *mils*, with one "mil" being $\frac{1}{1000}$ of an inch (0.001 inch). Vibrational velocity is measured in inches per second, following the displacement unit of the inch. Acceleration, although it could be expressed in units of inches per second squared, is more often represented in the unit of the G: a multiple of Earth's own gravitational acceleration.

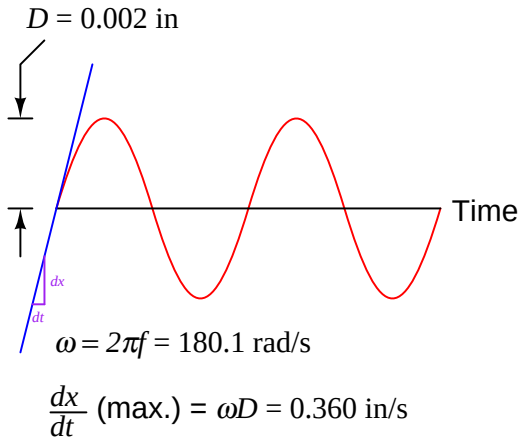
To give perspective to these units, it is helpful to consider a real application. Suppose we have a rotating machine vibrating in a sinusoidal (sine- or cosine-shaped) manner with a peak displacement (D) of 2 mils (0.002 inch) at a rotating speed of 1720 RPM (revolutions per minute). The frequency of this rotation is 28.667 Hz (revolutions per *second*), or 180.1 radians per second:



f = 1720 RPM = 28.667 Hz

$\omega = 2\pi f$ = 180.1 rad/s

If D is the peak displacement of the sinusoid, then ωD must be the peak velocity (maximum rate-of-change over time) of the sinusoid[3]. This yields a peak velocity of 0.360 inches per second:



We may apply differentiation once more to obtain the acceleration of this machine's rotating element. If D is the peak displacement of the sinusoid, and ωD the peak velocity, then $\omega^2 D$ will be its peak acceleration.

$$D = \text{Peak displacement} = 0.002 \text{ in}$$

$$\omega D = \text{Peak velocity} = 0.360 \text{ in/s}$$

$$\omega^2 D = \text{Peak acceleration} = 64.9 \text{ in/s}^2$$

The nominal value of Earth's gravitational acceleration (g) is 32.17 feet per second squared. This equates to about 386 inches per second squared. Since our machine's peak vibrational acceleration is 64.9 inches per second squared, this may be expressed as a "G" ratio to Earth's gravity:

$$\frac{64.9 \text{ in/s}^2}{386 \text{ in/s}^2} = 0.168 \text{ G's of peak acceleration}$$

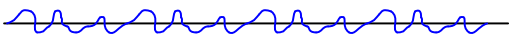
Using "G's" as a unit of acceleration makes it very easy to calculate forces imparted to the rotating element. If the machine's rotating piece weighs 1200 pounds (in 1 "G" of Earth gravity), then the force imparted to this piece by the vibrational acceleration of 0.168 G's will be 16.8% of its weight, or 201.7 pounds.

[3]Recall that the derivative of the sinusoidal function $\sin \omega t$ is equal to $\omega \cos \omega t$, and that the second derivative of $\sin \omega t$ is equal to $-\omega^2 \sin \omega t$. With each differentiation, the constant of angular velocity (ω) is applied as a multiplier to the entire function.

23.1.2 Non-sinusoidal vibrations

Normal machine vibrations rarely take the form of perfect sinusoidal waves. Although typical vibration waveforms are periodic (i.e. they repeat a pattern over time), they usually do not resemble sine or cosine waves in their shape:

A periodic, non-sinusoidal waveform



An unfortunate quality of non-sinusoidal waveforms is that they do not lend themselves as readily to mathematical analysis as sinusoidal waves. From the previous discussion on sinusoidal vibrations, we saw how simple it was to take the derivative of a sinusoidal waveform ($\frac{d}{dt} \sin \omega t = \omega \cos \omega t$), and how well this worked to predict velocity and acceleration from a function describing displacement. Most non-sinusoidal waveforms cannot be expressed as simply and neatly as $\sin \omega t$, however, and as such are not as easy to mathematically analyze.

Fortunately, though, there is a way to represent non-sinusoidal waveforms as combinations of sinusoidal waveforms. The French mathematician and physicist Jean Baptiste Joseph Fourier (1768-1830) proved mathematically that *any* periodic waveform, no matter how strange or asymmetrical its shape may be, may be replicated by a specific sum of sine and cosine waveforms of integer-multiple frequencies. That is, any periodic waveform (a periodic function of time, $f(\omega t)$ being the standard mathematical expression) is equivalent to a series of the following form[4]:

$$f(\omega t) = A_1 \cos \omega t + B_1 \sin \omega t + A_2 \cos 2\omega t + B_2 \sin 2\omega t + \cdots A_n \cos n\omega t + B_n \sin n\omega t$$

Here, ω represents the *fundamental* frequency of the waveform, while multiples of ω (e.g. 2ω, 3ω, 4ω, etc.) represent *harmonic* or *overtone* frequencies of that fundamental. The A and B coefficients describe the *amplitudes* (heights) of each sinusoid. We may break down a typical Fourier series in table form, labeling each term according to frequency:

Terms	Harmonic	Overtone
$A_1 \cos \omega t + B_1 \sin \omega t$	1st harmonic	Fundamental
$A_2 \cos 2\omega t + B_2 \sin 2\omega t$	2nd harmonic	1st overtone
$A_3 \cos 3\omega t + B_3 \sin 3\omega t$	3rd harmonic	2nd overtone
$A_4 \cos 4\omega t + B_4 \sin 4\omega t$	4th harmonic	3rd overtone
$A_n \cos n\omega t + B_n \sin n\omega t$	nth harmonic	$(n-1)$th overtone

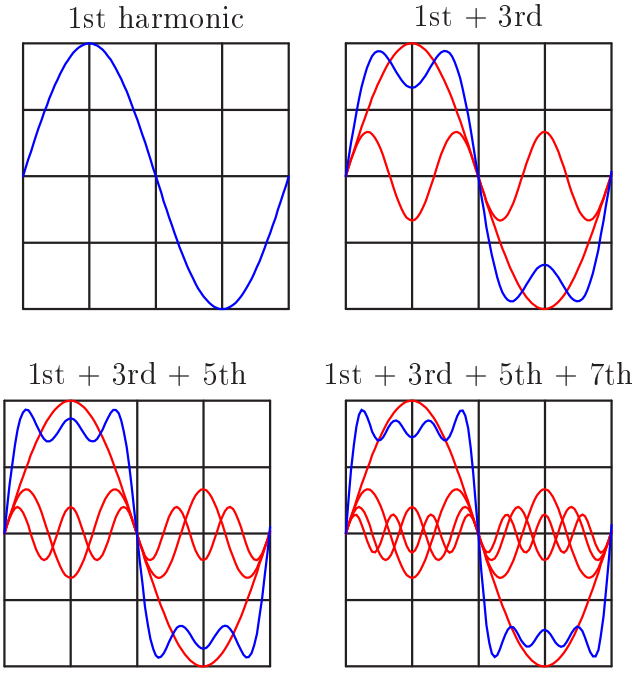
One of the most visually convincing examples of Fourier's theorem is the ability to describe a square wave as a series of sine waves. Intuition would suggest it is impossible to synthesize a sharp-edged waveform such as a square wave using nothing but rounded sinusoids, but it is indeed possible if one combines an *infinite* series of sinusoids of successively higher harmonic frequencies, given just the right combination of harmonic frequencies and amplitudes.

[4] There is an additional term missing in this Fourier series, and that is the "DC" or "bias" term A_0. Many non-sinusoidal waveforms having peak values centered about zero on a graph or oscilloscope display actually have *average* values that are non-zero, and the A_0 term accounts for this. However, this is usually not relevant in discussions of machine vibration, which is why I have opted to present the simplified Fourier series here.

The Fourier series for a square wave is as follows:

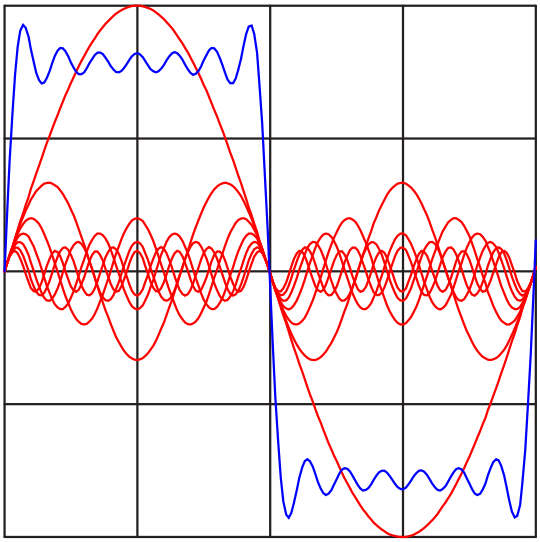
$$\text{Square wave} = 1 \sin \omega t + \frac{1}{3} \sin 3\omega t + \frac{1}{5} \sin 5\omega t + \frac{1}{7} \sin 7\omega t + \cdots$$

Such a series would be impossible to numerically calculate, but we may approximate it by adding several of the first (largest) harmonics together to see the resulting shape. In each of the following plots, we see the individual harmonic waveforms plotted in red, with the sum plotted in blue:



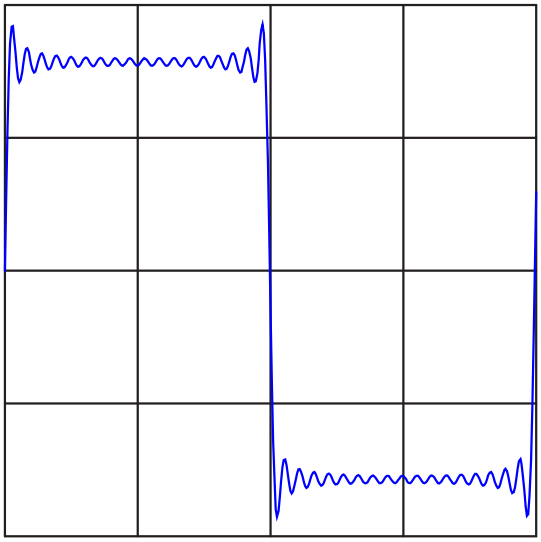
If we continue this pattern up to the 13th harmonic (following the same pattern of diminishing reciprocal amplitudes shown in the Fourier series for a square wave), we see the resultant sum looking more like a square wave:

1st + 3rd + 5th + 7th + 9th + 11th + 13th



Continuing on to the 35th harmonic, the resultant sum looks like a square wave with ripples at each rising and falling edge:

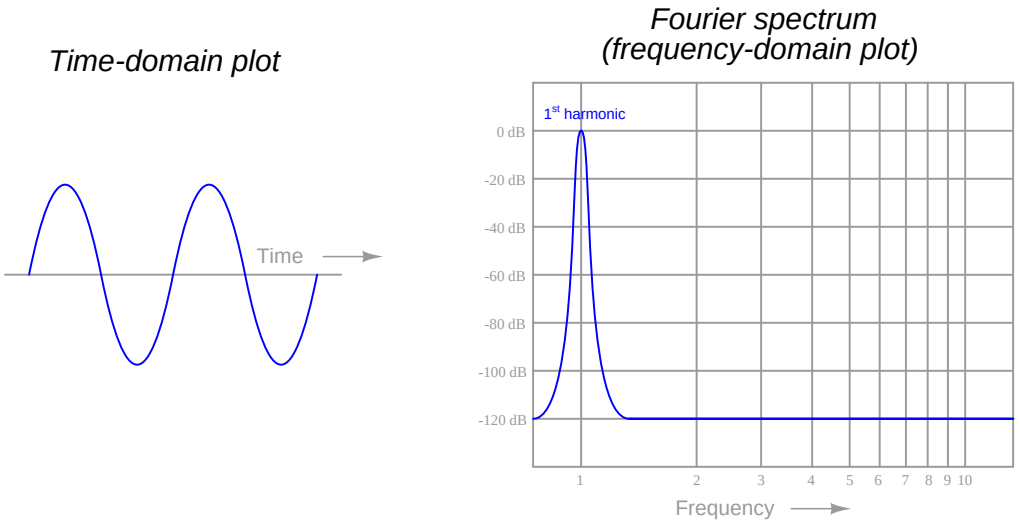
All odd-numbered harmonics up to the 35th



If we were to continue adding successive terms in this infinite series, the resulting superposition of sinusoids would look more and more like a perfect square wave.

The only real question in any practical application is, "What are the A, B, and ω coefficient values necessary to describe a particular non-periodic waveform using a Fourier series?" Fourier's theorem tells us we should be able to represent *any* periodic waveform – no matter what its shape – by summing together a particular series of sinusoids of just the right amplitudes and frequencies, but actually determining those amplitudes and frequencies is a another matter entirely. Fortunately, modern computational techniques such as the *Fast Fourier Transform* (or *FFT*) algorithm make it very easy to sample any periodic waveform and have a digital computer calculate the relative amplitudes and frequencies of its constituent harmonics. The result of a FFT analysis is a summary of the amplitudes, frequencies, and (in some cases) the phase angle of each harmonic.

To illustrate the relationship between a waveform plotted with respect to time versus a Fourier analysis showing component frequencies, I will show a pair of Fourier spectrum plots for two waveforms – one a perfect sinusoid and the other a non-sinusoidal waveform. First, the perfect sinusoid:

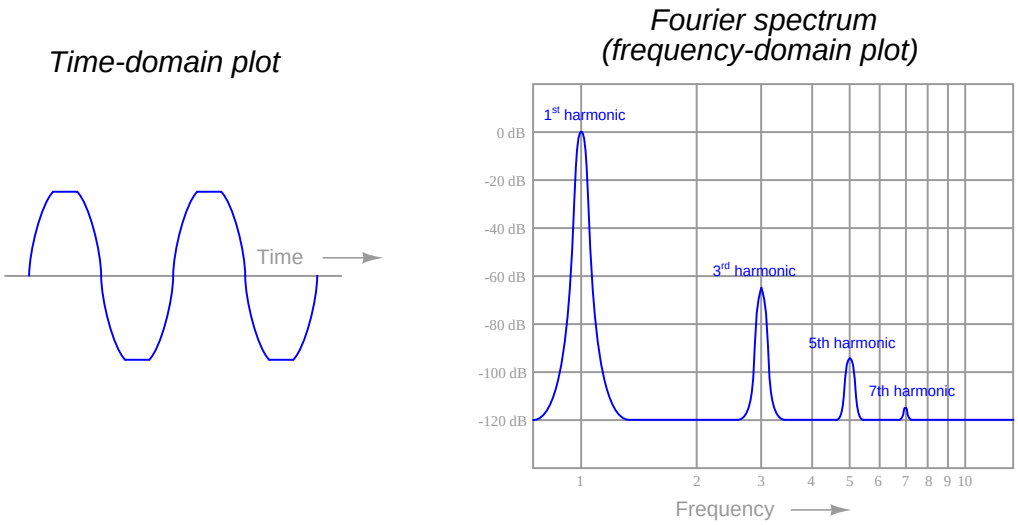


Fourier spectra are often referred to as *frequency-domain* plots because the x-axis (the "domain" in mathematical lingo) is frequency. A standard oscilloscope-type plot is called a *time-domain* plot because the x-axis is time. In this first set of plots, we see a perfect sine wave reduced to a single peak on the Fourier spectrum, showing a signal with only one frequency (the fundamental, or 1st harmonic). Here, the Fourier spectrum is very plain because there is only one frequency to display. In other words, the Fourier series for this perfect sinusoid would be:

$$f(\omega t) = 0\cos\omega t + 1\sin\omega t + 0\cos 2\omega t + 0\sin 2\omega t + \cdots 0\cos n\omega t + 0\sin n\omega t$$

Only the B_1 coefficient has a non-zero value. All other coefficients are zero because it only takes one sinusoid to perfectly represent this waveform.

Next, we will examine the Fourier analysis of a non-sinusoidal waveform:



In this second set of plots, we see the waveform is similar to a sine wave, except that it appears "clipped" at the peaks. This waveform is obviously not a perfect sinusoid, and therefore cannot be described by just one of the terms ($\sin \omega t$) in a Fourier series. It can, however, be described as equivalent to a *series* of perfect sinusoids summed together. In this case, the Fourier spectrum shows one sinusoid at the fundamental frequency, plus another (smaller) sinusoid at three times the fundamental frequency (3ω), plus another (yet smaller) sinusoid at the 5th harmonic and another (smaller still!) at the 7th: a series of *odd-numbered* harmonics.

If each of these harmonics is in phase with each other[5], we could write the Fourier series as a set of sine terms:

$$f(\omega t) = (0 \text{ dB}) \sin \omega t + (-65 \text{ dB}) \sin 3\omega t + (-95 \text{ dB}) \sin 5\omega t + (-115 \text{ dB}) \sin 7\omega t$$

Translating the decibel amplitude values into simple coefficients, we can see just how small these harmonic sinusoids are in comparison to the fundamental:

$$f(\omega t) = 1 \sin \omega t + 0.000562 \sin 3\omega t + 0.0000178 \sin 5\omega t + 0.00000178 \sin 7\omega t$$

If the waveform deviated even further from a perfect sinusoid, we would see a Fourier spectrum with taller harmonic peaks, and perhaps more of them (possibly including some even-numbered harmonics, not just odd-numbered), representing a harmonically "richer" spectrum.

Within the technical discipline of machine vibration analysis, harmonic vibrations are often referred to by labels such as *1X*, *2X*, and *3X*, the integer number corresponding to the harmonic order of the vibration. The fundamental, or first harmonic, frequency of a vibration would be represented by "1X" while "2X" and "3X" represent the second- and third-order harmonic frequencies, respectively.

[5] We have no way of knowing this from the Fourier spectrum plot, since that only shows us amplitude (height) and frequency (position on the x-axis).

On a practical note, the Fourier analysis of a machine's vibration waveform holds clues to the successful balancing of that machine. A first-harmonic vibration may be countered by placing an off-center mass on the rotating element 180 degrees out of phase with the offending sinusoid. Given the proper phase (180^o – exactly opposed) and magnitude, any harmonic may be counterbalanced by an off-center mass rotating at the same frequency. In other words, we may cancel any particular harmonic vibration with an equal and opposite harmonic vibration.

If you examine the "crankshaft" of a piston engine, for example, you will notice counterweights with blind holes drilled in specific locations for balancing. These precisely-trimmed counterweights compensate for first-harmonic (fundamental) frequency vibrations resulting from the up-and-down oscillations of the pistons within the cylinders. However, in some engine designs such as inline 4-cylinder arrangements, there are significant harmonic vibrations of greater order than the fundamental, which *cannot* be counterbalanced by any amount of weight, in any location, on the rotating crankshaft. The reciprocating motion of the pistons and connecting rods produce periodic vibrations that are non-sinusoidal, and these vibrations (like all periodic, non-sinusoidal waveforms) are equivalent to a series of harmonically-related sinusoidal vibrations.

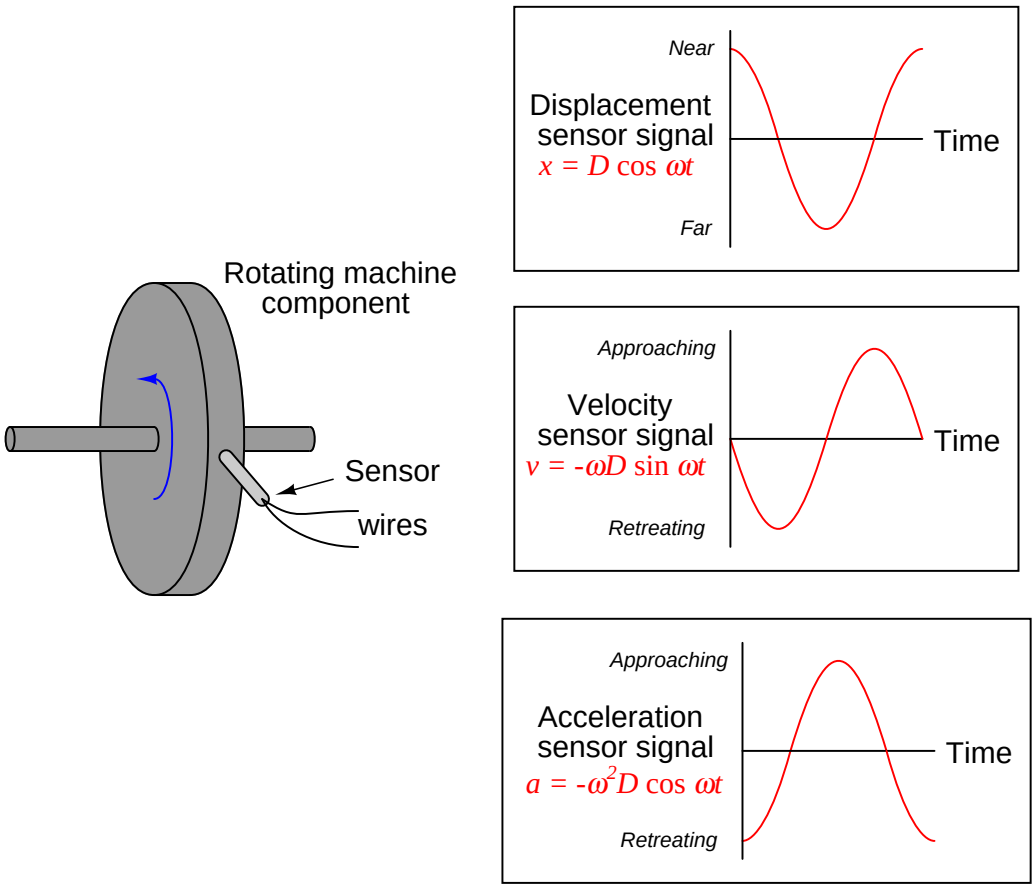
Any weight attached to the crankshaft will produce a first-order (fundamental) sinusoidal vibration, and that is all. In order to counteract harmonic vibrations of higher order, the engine requires counterbalance shafts spinning at speeds corresponding to those higher orders. This is why many high-performance inline 4-cylinder engines employ counterbalance shafts spinning at *twice* the crankshaft speed: to counteract the second-harmonic vibrations created by the reciprocating parts. If an engine designer were so inclined, he or she could include several counterbalance shafts, each one spinning at a different multiple of the crankshaft speed, to counteract as many harmonics as possible. At some point, however, the inclusion of all these shafts and the gearing necessary to ensure their precise speeds and phase shifts would interfere with the more basic design features of the engine, which is why you do not typically see an engine with multiple counterbalance shafts.

The harmonic content of a machine's vibration signal in and of itself tells us little about the health or balance of that machine. It may be perfectly normal for a machine to have a very "rich" harmonic signature due to convoluted motions of its parts[6]. However, Fourier analysis provides a simple way to quantify complex vibrations and to archive them for future reference. For example, we might gather vibration data on a new machine immediately after installation (including its Fourier spectra on all vibration measurement points) and store this data for safe keeping in the maintenance archives. Later, if and when we suspect a vibration-related problem with this machine, we may gather new vibration data and compare it against the original "signature" spectra to see if anything substantial has changed. Changes in harmonic amplitudes and/or the appearance of new harmonics may point to specific problems inside the machine. Expert knowledge is usually required to interpret the spectral changes and discern what those specific problem(s) might be, but at least this technique does have diagnostic value in the right hands.

[6] Machines with reciprocating components, such as pistons, cam followers, poppet valves, and such are notorious for generating vibration signatures which are anything but sinusoidal even under normal operating circumstances!

23.2 Vibration sensors

Sensors used to measure vibration come in three basic types: *displacement*, *velocity*, and *acceleration*. Displacement sensors measure changes in distance between a machine's rotating element and its stationary housing (frame). Displacement sensors come in the form of a probe that threads into a hole drilled and tapped in the machine's frame, just above the surface of a rotating shaft. Velocity and acceleration sensors, by contrast, measure the velocity or acceleration of whatever element the sensor is attached to, which is usually some external part of the machine frame[7].



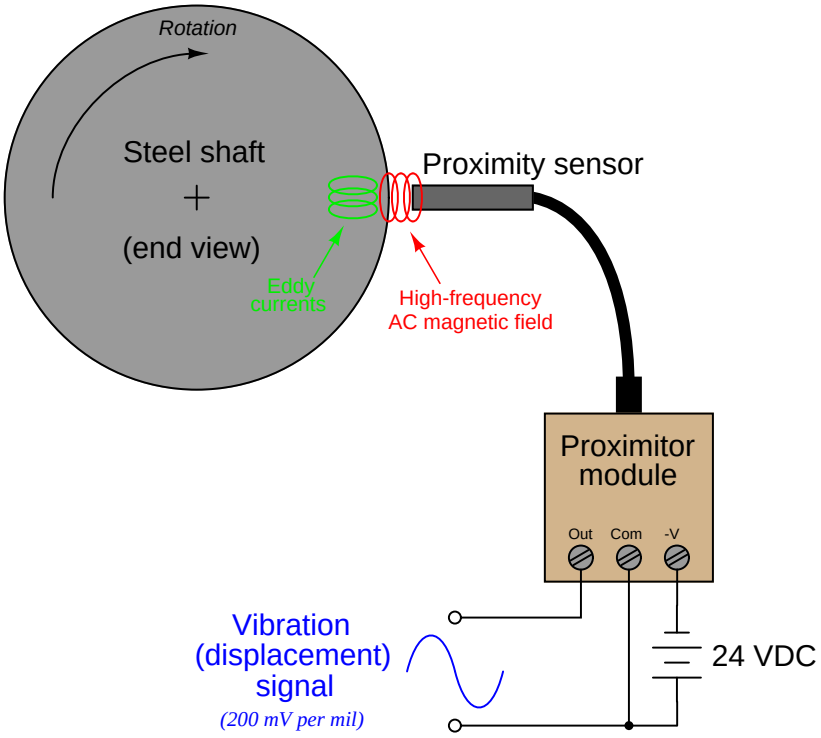
A design of displacement sensor manufactured by the Bently-Nevada corporation uses electromagnetic *eddy current* technology to sense the distance between the probe tip and the rotating machine shaft. The sensor itself is an encapsulated coil of wire, energized with high-frequency

[7] From the perspective of measurement, it would be ideal to affix a velocimeter or accelerometer sensor directly to the rotating element of the machine, but this leads to the problem of electrically connecting the (now rotating!) sensor to stationary analysis equipment. Unless the velocity or acceleration sensor is wireless, the only practical mounting location is on the stationary frame of the machine.

alternating current (AC). The magnetic field produced by the coil induces eddy currents in the metal shaft of the machine, as though the metal piece were a short-circuited secondary coil of a transformer (with the probe's coil as the transformer primary winding). The closer the shaft moves toward the sensor tip, the tighter the magnetic coupling between the shaft and the sensor coil, and the stronger the eddy currents.

The high-frequency oscillator circuit providing the sensor coil's excitation signal becomes loaded by the induced eddy currents. Therefore, the oscillator's load becomes a direct indication of how close the probe tip is to the metal shaft. This is not unlike the operation of a metal detector: measuring the proximity of a wire coil to any metal object by the degree of loading caused by eddy current induction.

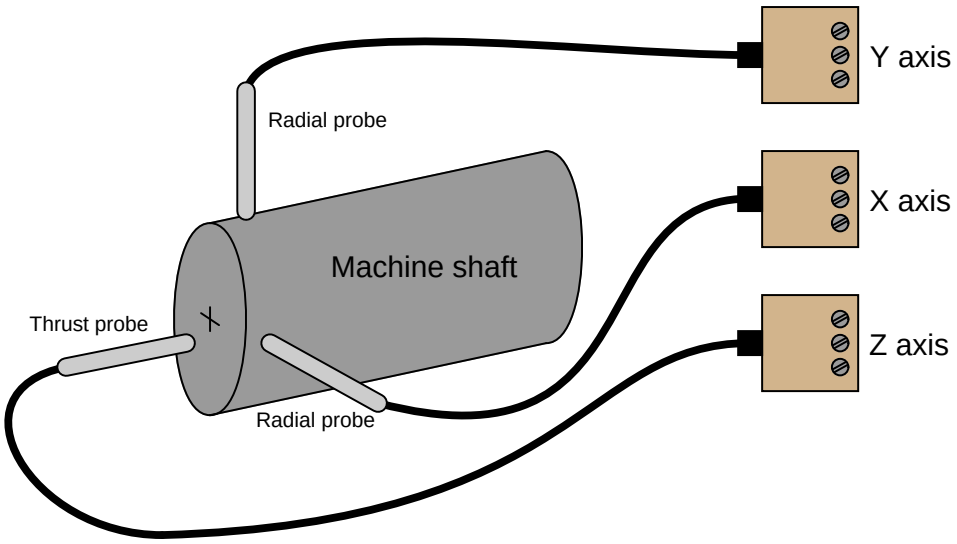
In the Bently-Nevada design, the oscillator circuit providing sensor coil excitation is called a *proximitor*. The proximitor module is powered by an external DC power source, and drives the sensor coil through a coaxial cable. Proximity to the metal shaft is represented by a DC voltage output from the proximitor module, with 200 millivolts per mil (1 mil = $\frac{1}{1000}$ inch) of motion being the standard calibration.



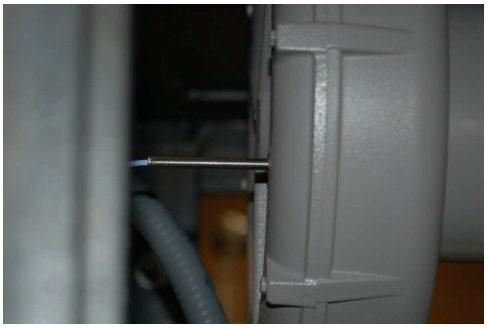
Since the proximitor's output voltage is a direct representation of distance between the probe's tip and the shaft's surface, a "quiet" signal (no vibration) will be a pure DC voltage. The probe is adjusted by a technician such that this quiescent voltage will lie between the proximitor's output voltage range limits. Any vibration of the shaft will cause the proximitor's output voltage to vary in precise step. A shaft vibration of 28.67 Hz, for instance, will cause the proximitor output signal to be a 28.67 Hz waveform superimposed on the DC "bias" voltage set by the initial probe/shaft gap.

An oscilloscope connected to this output signal will show a direct representation of shaft vibration, as measured in the axis of the probe. In fact, *any* electronic test equipment capable of analyzing the voltage signal output by the proximitor may be used to analyze the machine's vibration: oscilloscopes, spectrum analyzers, peak-indicating voltmeters, RMS-indicating voltmeters, etc.

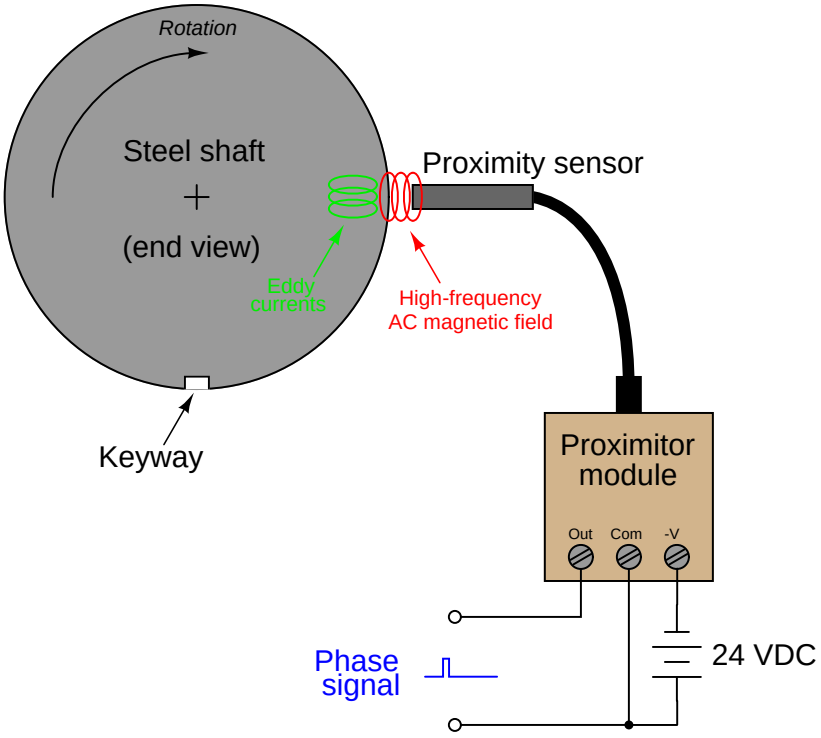
It is customary to arrange a set of three displacement probes at the end of a machine shaft to measure vibration: two *radial* probes and one *axial* (or *thrust*) probe. The purpose of this *triaxial* probe configuration is to measure shaft vibration (and/or shaft displacement) in all three dimensions:



Photographs of a Bently-Nevada displacement sensor (sensing axial vibration on a "ring" style air compressor) and two proximitor modules are shown here:



It is also common to see one *phase reference* probe installed on the machine shaft, positioned in such a way that it detects the periodic passing of a keyway or other irregular feature on the shaft. The "keyphasor" signal will consist of one large pulse per revolution:



The purpose of a keyphasor signal is two-fold: to provide a reference point in the machine's rotation to correlate other vibration signals against, and to provide a simple means of measuring shaft speed. The location in time of the pulse represents shaft position, while the frequency of that pulse signal represents shaft speed.

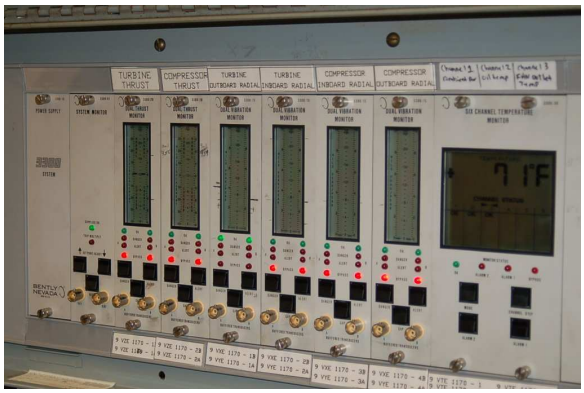
For instance, if one of the radial displacement sensors indicates a high vibration at the same frequency as the shaft rotation (i.e. the shaft is bowed in one direction, like a banana spinning on its long axis), the phase shift between the vibration's sinusoidal peak and the phase reference pulse will indicate to maintenance machinists where the machine is out of balance. This is not unlike automatic tire-balancing machines designed to measure imbalance in automobile tire and wheel assemblies: the machine must have some way of indicating to the human operator *where* a balancing weight should be placed, not just how far out of balance the tire is. In the case of machine vibration monitoring equipment, the keyphasor signal and one of the axial displacement signals may be simultaneously plotted on a dual-trace oscilloscope for the purposes of determining the position of the imbalance on the machine shaft.

23.3 Monitoring hardware

The following photograph shows a large air blower in a wastewater treatment facility equipped with a Bently-Nevada model 3300 vibration monitoring rack (located left-center on the foreground panel):

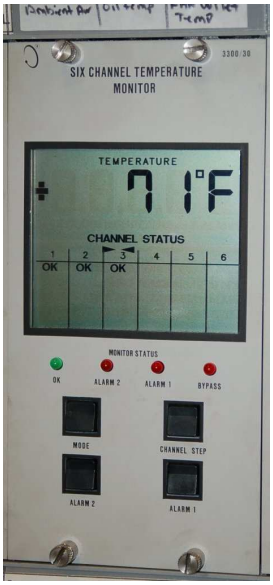
Five vibration measurement and display cards are installed in this rack, each card capable of processing up to two displacement sensor signals. A six-channel temperature monitor card is also installed in the rack, used to display bearing and other machine component temperatures. Like the vibration cards, the temperature card is capable of generating both "alert" and "trip" signals, monitoring the presence of slightly abnormal conditions and taking automatic shut-down action in the event of excessively abnormal conditions, respectively.

A closer view of a different Bently-Nevada model 3300 vibration monitoring rack is shown in this photograph:



Each "card" inserted into this rack performs a different measurement function.

The following photographs show even closer views of the cards, revealing the display bargraphs and the units of measurement. From left to right; thrust measurement, vibration measurement, temperature measurement (6 channels), and speed measurement:



BNC-style cable connectors on the front of the cards provide convenient connection points for electronic test equipment such as oscilloscopes or spectrum analyzers. This eliminates the need to un-do wire connections on the proximitor units in order to take diagnostic measurements. Each card also provides "alert" and "danger" levels for their respective measurements, generating a contact-closure signal which may be connected to an automatic shutdown ("trip") system to take protective action if vibration or thrust displacement ever exceeds pre-set limits.

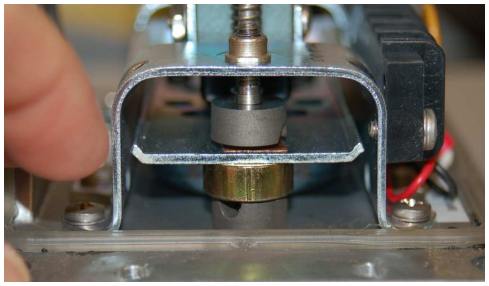
Another variety of vibration monitoring hardware is the Bently-Nevada 1701 FieldMonitor. This hardware lacks the convenient front-panel displays of the model 3300, opting instead to communicate vibration data in digital form to an Allen-Bradley programmable logic controller (PLC). Not only does this make it possibly to display the vibration data remotely through HMI (Human-Machine Interface) panels, but it also enables vibration data to engage automatic "trip" logic programming in the PLC to shut the machine down in the event of excessive vibration. This next photograph shows several FieldMonitor modules plugged into a rack, acquiring displacement data from eight proximity probes (X and Y axis radial measurements at three machine bearing locations, plus one axial (thrust) measurement and one phase reference measurement):

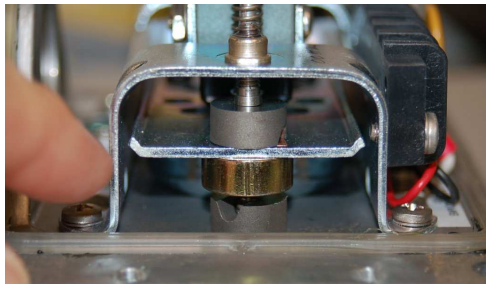
23.4 Mechanical vibration switches

A much simpler alternative to continuous vibration sensors (displacement or acceleration) and monitoring equipment suitable for less critical applications is a simple mechanical switch actuated by a machine's vibration. These switches cannot, of course, quantitatively analyze machine vibrations, but they do serve as qualitative indicators of gross vibration.

The following photograph shows a Robertshaw "Vibraswitch" unit:

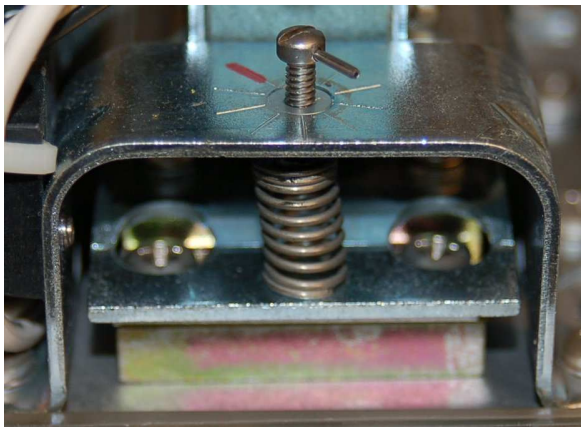
This switch works on the principle of a weighted lever generating a force when shaken. A pair of magnets located at the weighted end of the lever hold it in either the "reset" (normal) or "tripped" position:





When reset, the lever is pre-loaded by spring tension to flip to the "tripped" position. All it needs to make that transition is enough acceleration to generate the "breakaway" force necessary to pull away from the holding magnet. Once the acceleration force exceeds that threshold, the lever moves toward the other magnet, which holds it securely in position so that switch will not "reset" itself with additional vibration.

This pre-loading spring is adjustable by a small screw, making it possible to easily vary the sensitivity of the switch:



References

Kaplan, Wilfred, *Advanced Mathematics for Engineers*, Addison-Wesley Publishing Company, Reading, MA, 1981.

Smith, Steven W., *The Scientist and Engineer's Guide to Digital Signal Processing*, California Technical Publishing, San Diego, CA, 1997.

White, Glenn D., *Introduction to Machine Vibration*, version 1.76, part number 8569, DLI Engineering Corp., Bainbridge Island, WA, 1995.

Chapter 24

Signal characterization

Mathematics is full of complementary principles and symmetry. Perhaps nowhere is this more evident than with *inverse functions*: functions that "un-do" one another when put together. A few examples of inverse functions are shown in the following table:

$f(x)$	$f^{-1}(x)$
Addition	Subtraction
Multiplication	Division
Power	Root
Exponential	Logarithm
Derivative	Integral

Inverse functions are vital to master if one hopes to be able to manipulate algebraic (literal) expressions. For example, to solve for time (t) in this exponential formula, you must know that the natural logarithm function directly "un-does" the exponential e^x. This is the only way to "unravel" the equation and get t isolated by itself on one side of the equals sign:

$$V = 12e^{-t}$$

Divide both sides by 12

$$\frac{V}{12} = e^{-t}$$

Take the natural logarithm of both sides

$$\ln\left(\frac{V}{12}\right) = \ln\left(e^{-t}\right)$$

The natural logarithm "cancels out" the exponential

$$\ln\left(\frac{V}{12}\right) = -t$$

Multiply both sides by negative one

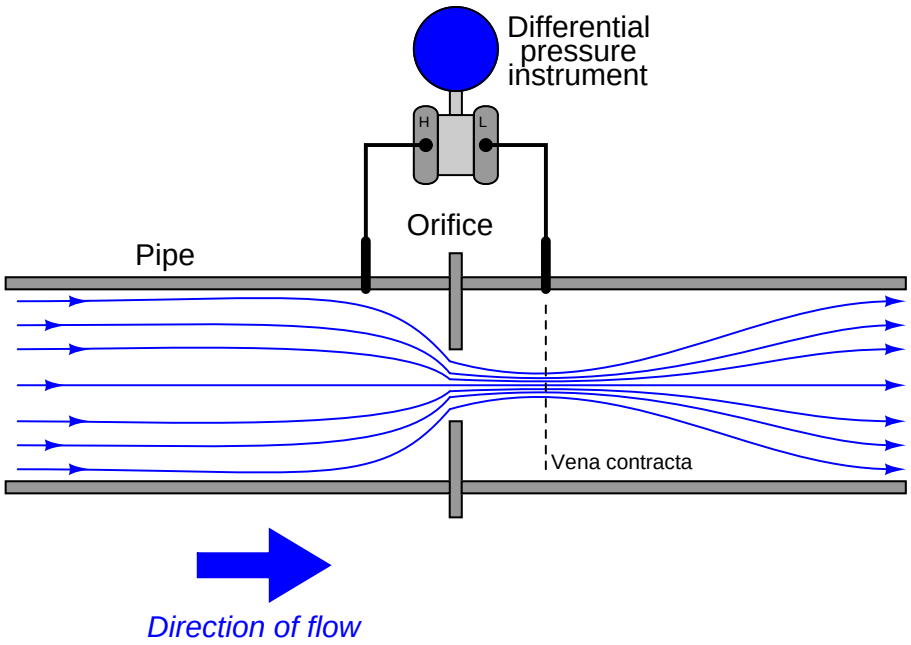
$$-\ln\left(\frac{V}{12}\right) = t$$

In industry there exist a great many practical problems where inverse functions play a similar role. Just as inverse functions are useful for manipulating literal expressions in algebra, they are also useful in inferring measurements of things we cannot directly measure. Many continuous industrial measurements are *inferential* in nature, meaning that we actually measure some other variable in order to quantify the variable of interest. More often than not, the relationship between the primary variable and the inferred variable is nonlinear, necessitating some form of mathematical processing to complete the inferential measurement.

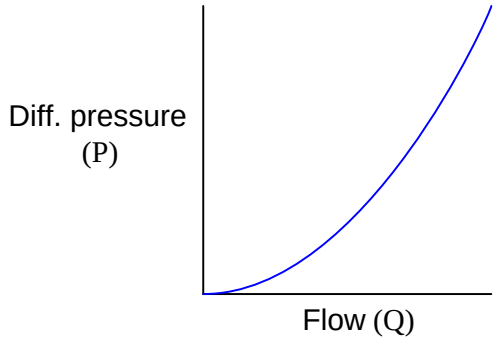
Take for instance the problem of measuring fluid flow through a pipe. To the layperson, this may seem to be a trivial problem. However there is no practical way to *directly* and continuously measure the flow rate of a fluid, especially when we cannot allow the fluid in question to become exposed to the atmosphere (e.g. when the liquid or gas in question is toxic, flammable, under high pressure, or any combination thereof).

One standard way to measure the flow rate of a fluid through a pipe is to intentionally place a restriction in the path of the fluid, and measure the pressure drop across that restriction. The most common form of intentional restriction used for this purpose is a thin plate of metal with a hole precisely machined in the center, called an *orifice plate*.

A side view of the orifice plate assembly and pressure-measuring instrument looks like this:



This approach should make intuitive sense: the faster the flow rate of the fluid, the greater the pressure difference developed across the orifice. The actual physics of this process has to do with energy exchanging between potential and kinetic forms, but that is incidental to this discussion. The mathematically interesting characteristic of this flow measurement technique is its nonlinearity. Pressure does not rise linearly with flow rate; rather, it increases with the *square* of the flow rate:

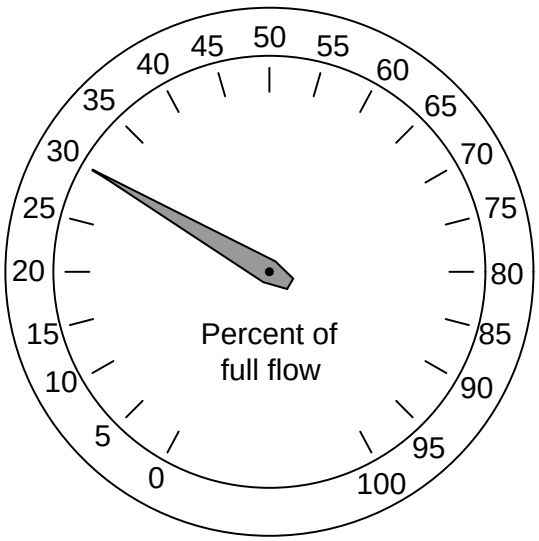


To write this as a proportionality, we relate flow rate (Q) to pressure (P) as follows (the constant k accounts for unit conversions and the geometries of the orifice plate and pipe):

$$P = kQ^2$$

This is a practical problem for us because our intent is to use pressure measurement (P) as an indirect (inferred) indication of flow rate (Q). If the two variables are not *directly* related to one another, we will not be able to regard one as being directly representative of the other. To make this problem more clear to see, imagine a pressure gauge connected across the restriction, with the face of the gauge labeled in percent:

Face of pressure gauge, calibrated to read in percent of full flow rate

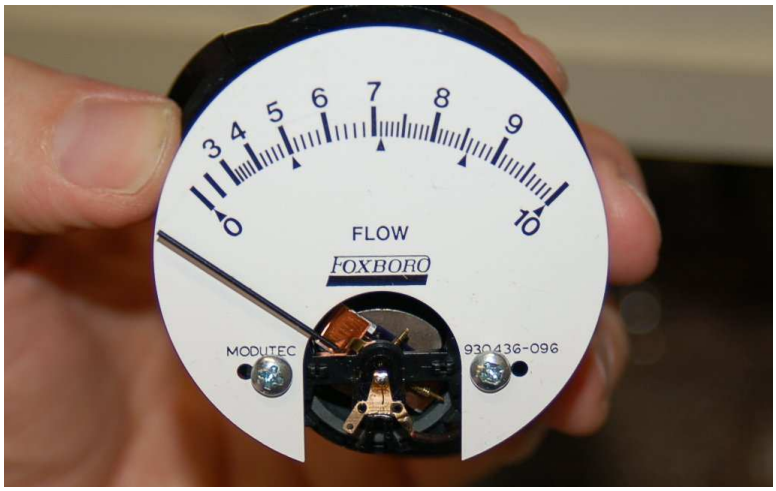


Consider a pressure gauge such as the one shown above, registering 20 percent on a linear scale at some amount of flow through the pipe. What will happen if the flow rate through that pipe suddenly doubles? An operator or technician looking at the gauge *ought* to see a new reading of 40 percent, if indeed the gauge is supposed to indicate flow rate. However, this will not happen. Since the pressure dropped across the orifice in the pipe increases with the square of flow rate, a doubling of flow rate will actually cause the pressure gauge reading to *quadruple!* In other words, it will go from reading 20% to reading 80%, which is definitely not an accurate indication of the flow increase.

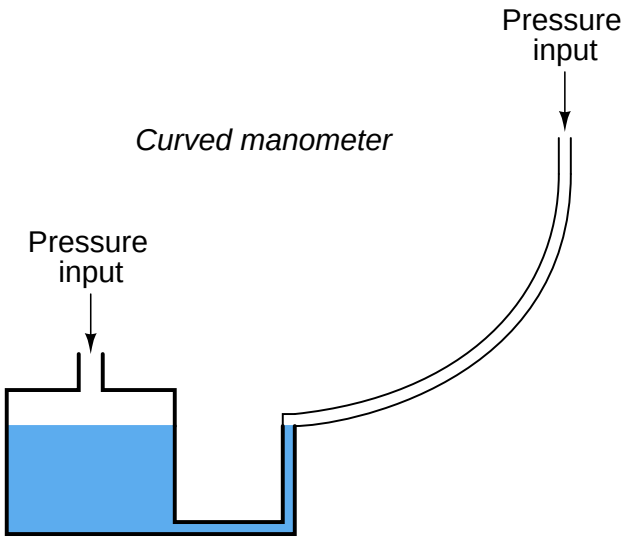
A couple of simple solutions exist for addressing this problem. One is to re-label the pressure gauge with a "square root" scale. Examine this photograph of a 3-15 PSI receiver gauge:

Now, a doubling of fluid flow rate still results in a quadrupling of needle motion, but due to the nonlinear (inner) scale on this gauge this needle motion translates into a simple doubling of indicated flow, which is precisely what we need for this to function as an accurate flow indicator.

If the differential pressure instrument outputs a 4-20 mA analog electronic signal instead of a 3-15 PSI pneumatic signal, we may apply the same "nonlinear scale" treatment to any current meter and achieve the same result:



Another simple solution is to use a *nonlinear manometer*, with a curved viewing tube[1]:



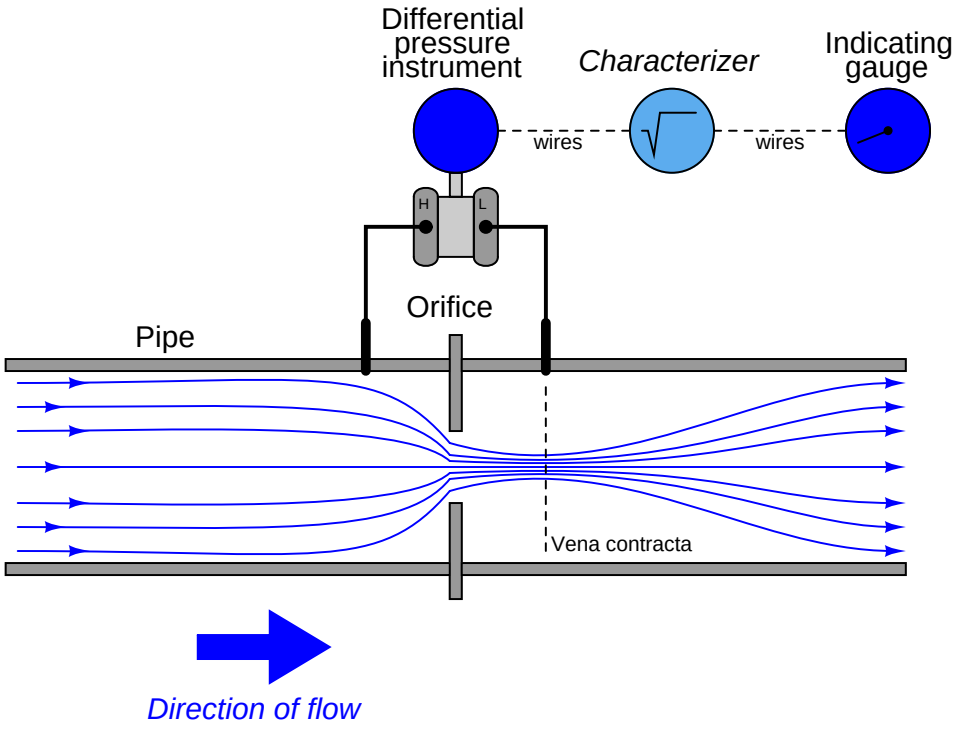
The scale positioned alongside the curved viewing tube will be linear, with equal spacings between division marks along its entire length. The vertical height of the liquid column translates pressure into varying degrees of movement along the axis of the tube by the tube's curvature. Literally, any

[1] This solution works best for measuring the flow rate of gases, not liquids, since the manometer obviously must use a liquid of its own to indicate pressure, and mixing or other interference between the process liquid and the manometer liquid could be problematic.

inverse function desired may be "encoded" into this manometer by fashioning the viewing tube into the desired (custom) shape without any need to print a nonlinear scale.

Shown here is a photograph of an actual curved-tube manometer. This particular specimen does not have a scale reading in units of flow, but it certainly could if it had the correct curve for a square-root characterization:

A more sophisticated solution to the "square root problem" is to use a computer to manipulate the signal coming from the differential pressure instrument so the characterized signal becomes a direct, linear representation of flow. In other words, the computer *square-roots* the pressure sensor's signal in order that the final signal becomes a direct representation of fluid flow rate:



Both solutions achieve their goal by mathematically "un-doing" the nonlinear (square) function intrinsic to the physics of the orifice plate with a complementary (inverse) function. This intentional compounding of inverse functions is sometimes called *linearization*, because it has the overall effect of making the output of the instrument system a direct proportion of the input:

$$\text{Output} = k(\text{Input})$$

Fluid flow rate measurement in pipes is not the only application where we find nonlinearities complicating the task of measurement. Several other applications exhibit similar challenges:

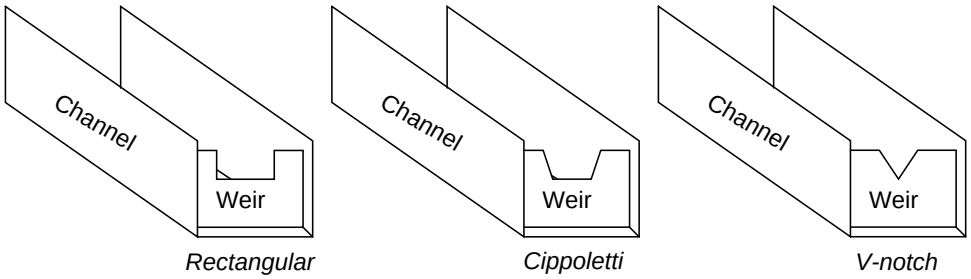
- Liquid flow measurement in open channels (over weirs)
- Liquid level measurement in non-cylindrical vessels
- Temperature measurement by radiated energy
- Chemical composition measurement

The following sections will describe the mathematics behind each of these measurement applications.

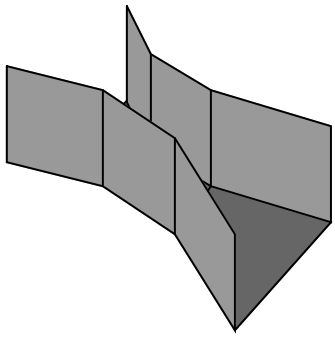
24.1 Flow measurement in open channels

Measuring the flow rate of liquid through an open channel is not unlike measuring the flow rate of a liquid through a closed pipe: one of the more common methods for doing so is to place a restriction in the path of the liquid flow and then measure the "pressure" dropped across that restriction. The easiest way to do this is to install a low "dam" in the middle of the channel, then measure the height of the liquid upstream of the dam as a way to infer flow rate. This dam is technically referred to as a *weir*, and three styles of weir are commonly used:

Different styles of weirs for measuring open-channel liquid flow



Another type of open-channel restriction used to measure liquid flow is called a *flume*. An illustration of a *Parshall flume* is shown here:



Weirs and *flumes* may be thought of being somewhat like "orifice plates" and "venturi tubes," respectively, for open-channel liquid flow. Like an orifice plate, a weir or a flume generates a differential pressure that varies with the flow rate through it. However, this is where the similarities end. Exposing the fluid stream to atmospheric pressure means the differential pressure caused by the flow rate manifests itself as a difference in liquid height at different points in the channel. Thus, weirs and flumes allow the indirect measurement of liquid flow by sensing liquid height. An interesting feature of weirs and flumes is that although they are nonlinear primary sensing elements, their nonlinearity is quite different from that of an orifice.

Note the following transfer functions for different weirs and flumes, relating the rate of liquid flow through the device (Q) to the level of liquid rise upstream of the device (called "head", or H):

$$Q = 2.48 \left(\tan \frac{\theta}{2} \right) H^{\frac{5}{2}} \qquad \text{V-notch weir}$$

$$Q = 3.367 L H^{\frac{3}{2}} \qquad \text{Cippoletti weir}$$

$$Q = 0.992 H^{1.547} \qquad \text{3-inch wide throat Parshall flume}$$

$$Q = 3.07 H^{1.53} \qquad \text{9-inch wide throat Parshall flume}$$

Where,
Q = Volumetric flow rate (cubic feet per second – CFS)
L = Width of notch crest or throat width (feet)
θ = V-notch angle (degrees)
H = Head (feet)

It is important to note these functions provide answers for flow rate (Q) with head (H) being the independent variable. In other words, they will tell us how much liquid is flowing given a certain head. In the course of calibrating the head-measuring instruments that infer flow rate, however, it is important to know the inverse transfer function: how much head there will be for any given value of flow. Here, algebraic manipulation becomes important to the technician. For example, here is the solution for H in the function for a Cippoletti weir:

$$Q = 3.367 L H^{\frac{3}{2}}$$

Dividing both sides of the equation by 3.367 and L:

$$\frac{Q}{3.367L} = H^{\frac{3}{2}}$$

Taking the $\frac{3}{2}$ root of both sides:

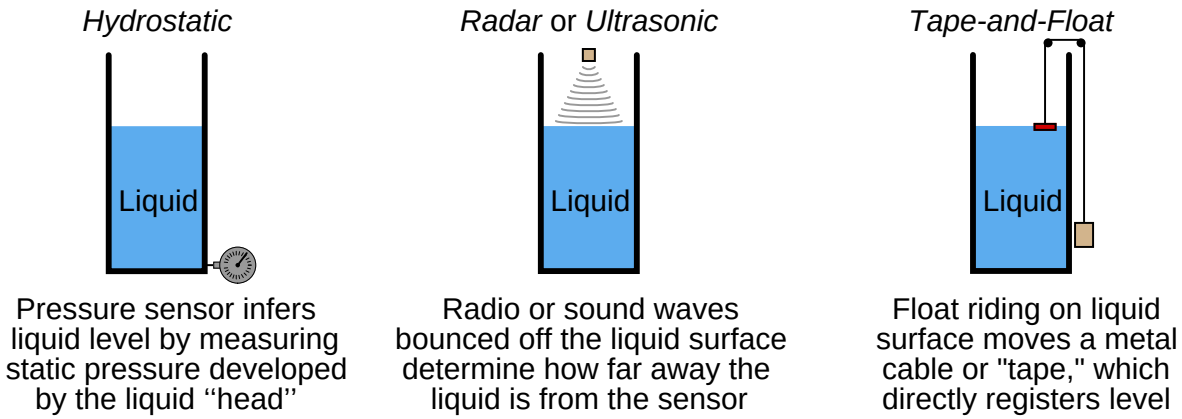
$$\sqrt[3/2]{\frac{Q}{3.367L}} = H$$

This in itself may be problematic, as some hand calculators do not have an $\sqrt[x]{y}$ function. In cases such as this, it is helpful to remember that a root is nothing more than an inverse power. Therefore, we could re-write the final form of the equation using a $\frac{2}{3}$ power instead of a $\frac{3}{2}$ root:

$$\left(\frac{Q}{3.367L}\right)^{\frac{2}{3}} = H$$

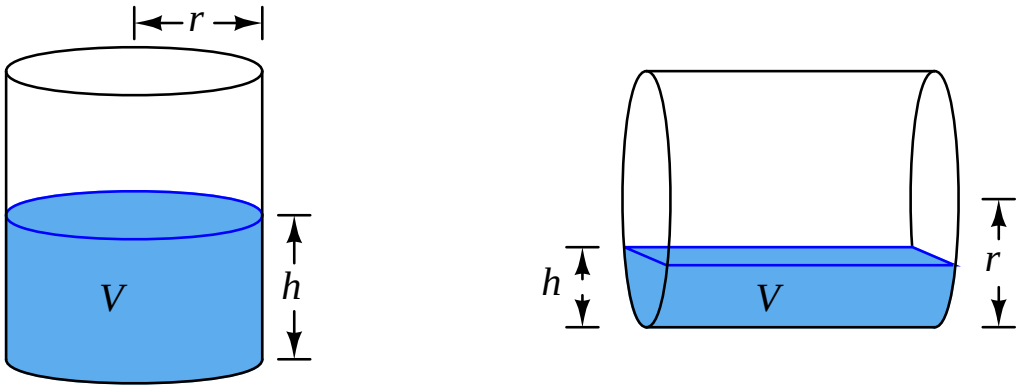
24.2 Liquid volume measurement

A variety of technologies exist to measure the quantity of stored liquid in a vessel. Hydrostatic pressure, radar, ultrasonic, and tape-and-float are just a few of the more common technologies:



These liquid measuring technologies share a common trait: they infer the quantity of liquid in the vessel by measuring liquid *height*. If the vessel in question has a constant cross-sectional area throughout its working height (e.g. a vertical cylinder), then liquid height will directly correspond to liquid volume. However, if the vessel in question does not have a constant cross-sectional area throughout its height, the relationship between liquid height and liquid volume will not be linear.

For example, there is a world of difference between the height/volume functions for a vertical cylinder versus a horizontal cylinder:

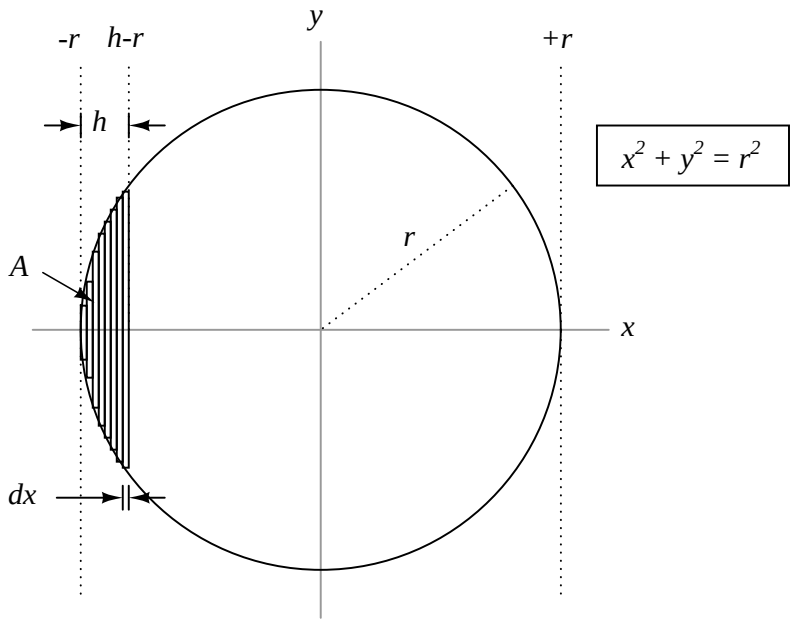


The volume function for a vertical cylinder is a simple matter of geometry – height (h) multiplied by the cylinder's cross-sectional area (πr^2):

$$V = \pi r^2 h$$

Calculating the volume of a horizontal cylinder as a function of liquid height (h) is a far more complicated matter, because the cross-sectional area is also a function of height. For this, we need to apply calculus.

First, we begin with the mathematical definition of a circle, then graphically represent a partial area of that circle as a series of very thin rectangles:



In this sketch, I show the circle "filling" from left to right rather than from bottom to top. I have done this strictly out of mathematical convention, where the x (horizontal) axis is the independent variable. No matter how the circle gets filled, the relationship of area (A) to fill distance (h) will be the same.

If $x^2 + y^2 = r^2$ (the mathematical definition of a circle), then the area of each rectangular "slice" comprising the accumulated area between $-r$ and $h - r$ is equal to $2y\,dx$. In other words, the total accumulated area between $-r$ and $h - r$ is:

$$A = \int_{-r}^{h-r} 2y\,dx$$

Now, writing y in terms of r and x ($y = \sqrt{r^2 - x^2}$) and moving the constant "2" outside the integrand:

$$A = 2\int_{-r}^{h-r} \sqrt{r^2 - x^2}\,dx$$

Consulting a table of integrals, we find this solution for the general form:

$$\int \sqrt{a^2 - u^2}\,du = \frac{u}{2}\sqrt{a^2 - u^2} + \frac{a^2}{2}\sin^{-1}\left(\frac{u}{a}\right) + C$$

Applying this solution to our particular integral . . .

$$A = 2\left[\frac{x}{2}\sqrt{r^2 - x^2} + \frac{r^2}{2}\sin^{-1}\left(\frac{x}{r}\right)\right]_{-r}^{h-r}$$

$$A = 2\left[\left(\frac{(h-r)}{2}\sqrt{r^2 - (h-r)^2} + \frac{r^2}{2}\sin^{-1}\frac{(h-r)}{r}\right) - \left(\frac{-r}{2}\sqrt{r^2 - (-r)^2} + \frac{r^2}{2}\sin^{-1}\frac{-r}{r}\right)\right]$$

$$A = 2\left[\left(\frac{(h-r)}{2}\sqrt{r^2 - (h^2 - 2hr + r^2)} + \frac{r^2}{2}\sin^{-1}\frac{(h-r)}{r}\right) - \left(\frac{-r}{2}\sqrt{0} + \frac{r^2}{2}\frac{-\pi}{2}\right)\right]$$

$$A = 2\left[\left(\frac{(h-r)}{2}\sqrt{2hr - h^2} + \frac{r^2}{2}\sin^{-1}\frac{(h-r)}{r}\right) - \left(\frac{-\pi r^2}{4}\right)\right]$$

$$A = \left[(h-r)\sqrt{2hr - h^2} + r^2\sin^{-1}\frac{(h-r)}{r} + \frac{\pi r^2}{2}\right]$$

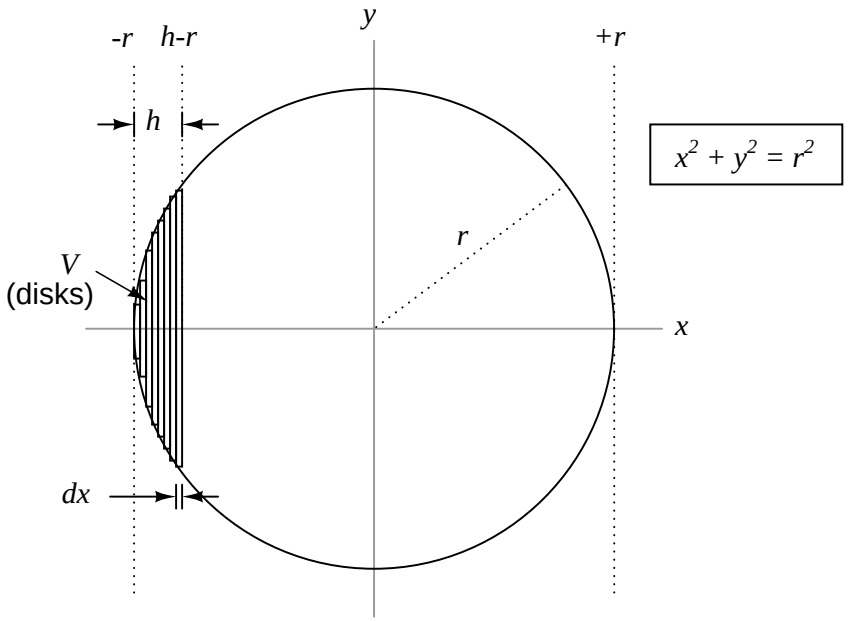
Knowing that the stored liquid volume in the horizontal tank will be this area multiplied by the constant length (L) of the tank, our formula for volume is as follows:

$$V = L\left[(h-r)\sqrt{2hr - h^2} + r^2\sin^{-1}\frac{(h-r)}{r} + \frac{\pi r^2}{2}\right]$$

As you can see, the result is far from simple. Any instrumentation system tasked with the inference of stored liquid volume by measurement of liquid height in a horizontal cylinder must somehow apply this formula on a continuous basis. This is a prime example of how digital computer technology is essential to certain continuous measurement applications!

Spherical vessels, such as those used to store liquefied natural gas (LNG) and butane, present a similar challenge. The height/volume function is nonlinear because the cross-sectional area of the vessel changes with height.

Calculus provides a way for us to derive an equation solving for stored volume (V) with height (h) as the independent variable. We begin in a similar manner to the last problem with the mathematical definition of a circle, except now we consider the filling of a sphere with a series of thin, circular disks:



If $x^2 + y^2 = r^2$ (the mathematical definition of a circle), then the volume of each circular disk comprising the accumulated volume between $-r$ and $h - r$ is equal to $\pi y^2\, dx$. In other words, the total accumulated area between $-r$ and $h - r$ is:

$$V = \int_{-r}^{h-r} \pi y^2\, dx$$

Now, writing y in terms of r and x ($y = \sqrt{r^2 - x^2}$) and moving the constant π outside the integrand:

$$V = \pi \int_{-r}^{h-r} \left(\sqrt{r^2 - x^2}\right)^2\, dx$$

Immediately we see how the square and the square-root cancel one another, leaving us with a fairly simple integrand:

$$V = \pi \int_{-r}^{h-r} r^2 - x^2\, dx$$

We may write this as the difference of two integrals:

$$V = \left(\pi \int_{-r}^{h-r} r^2 \, dx\right) - \left(\pi \int_{-r}^{h-r} x^2 \, dx\right)$$

Since r is a constant, the left-hand integral is simply $\pi r^2 x$. The right-hand integral is solvable by the power rule:

$$V = \pi r^2 \left[x\right]_{-r}^{h-r} - \left[\pi \frac{x^3}{3}\right]_{-r}^{h-r}$$

$$V = \pi r^2 \left[(h-r) - (-r)\right] - \left[\pi \frac{(h-r)^3}{3} - \pi \frac{(-r)^3}{3}\right]$$

$$V = \pi r^2 \left[h - r + r\right] - \frac{\pi}{3}\left[(h-r)^3 - (-r)^3\right]$$

$$V = \pi h r^2 - \frac{\pi}{3}\left[(h^3 - 2h^2 r + h r^2 - h^2 r + 2h r^2 - r^3) + r^3\right]$$

$$V = \pi h r^2 - \frac{\pi}{3}\left[h^3 - 3h^2 r + 3h r^2\right]$$

$$V = \pi h r^2 - \frac{\pi h^3}{3} + \pi h^2 r - \pi h r^2$$

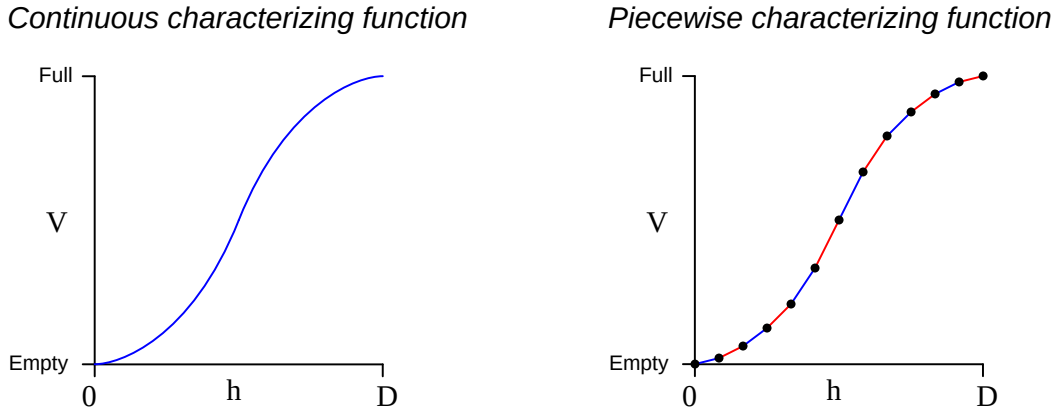
$$V = -\frac{\pi h^3}{3} + \pi h^2 r$$

$$V = \pi h^2 r - \frac{\pi h^3}{3}$$

$$V = \pi h^2 (r - \frac{h}{3})$$

This function will "un-do" the inherent height/volume nonlinearity of a spherical vessel, allowing a height measurement to translate directly into a volume measurement. A "characterizing" function such as this is typically executed in a digital computer connected to the level sensor, or sometimes in a computer chip within the sensor device itself.

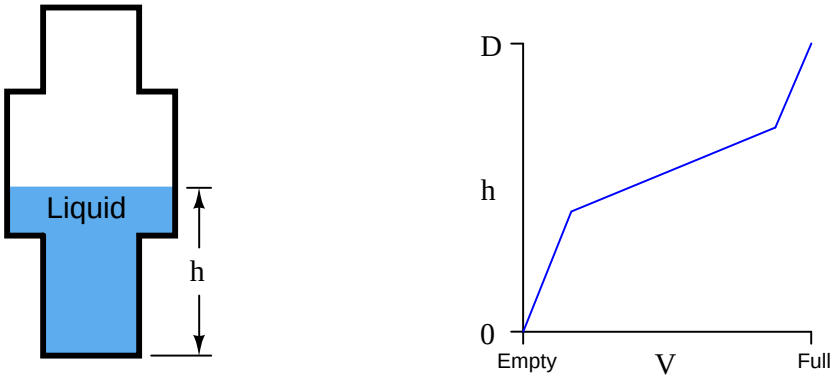
An interesting alternative to a formal equation for linearizing the level measurement signal is to use something called a *multi-segment characterizer* function, also implemented in a digital computer. This is an example of what mathematicians call a *piecewise function*: a function made up of line segments. Multi-segment characterizer functions may be programmed to emulate virtually any continuous function, with reasonable accuracy:



The computer correlates the input signal (height measurement, h) to a point on this piecewise function, linearly interpolating between the nearest pair of programmed coordinate points. The number of points available for multi-point characterizers varies between ten and one hundred[2] depending on the desired accuracy and the available computing power.

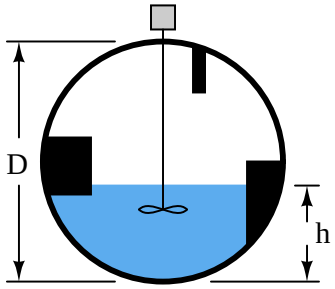
[2]There is no theoretical limit to the number of points in a digital computer's characterizer function given sufficient processing power and memory. There is, however, a limit to the patience of the human programmer who must encode all the necessary x, y data points defining this function. Most of the piecewise characterizing functions I have seen available in digital instrumentation systems provide 10 to 20 (x, y) coordinate points to define the function. Fewer than 10 coordinate points risks excessive interpolation errors, and more than 20 would just be tedious to set up.

Although true fans of math might blanch at the idea of approximating an inverse function for level measurement using a piecewise approach rather than simply implementing the correct continuous function, the multi-point characterizer technique does have certain practical advantages. For one, it is readily adaptable to any shape of vessel, no matter how strange. Take for instance this vessel, made of separate cylindrical sections welded together:



Here, the vessel's very own height/volume function is fundamentally piecewise, and so *nothing but* a piecewise characterizing function could possibly linearize the level measurement into a volume measurement!

Consider also the case of a spherical vessel with odd-shaped objects welded to the vessel walls, and/or inserted into the vessel's interior:

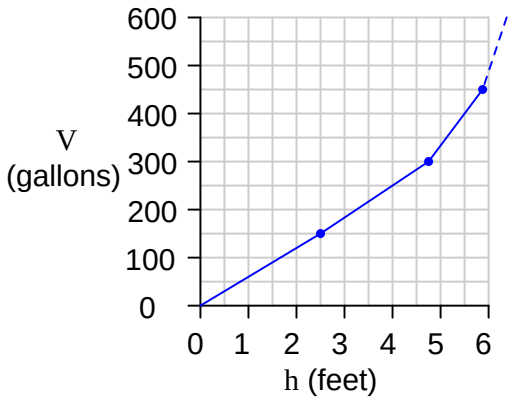


The volumetric space occupied by these structures will introduce all kinds of discontinuities into the transfer function, and so once again we have a case where a continuous characterizing function cannot properly linearize the level signal into a volume measurement. Here, only a piecewise function will suffice.

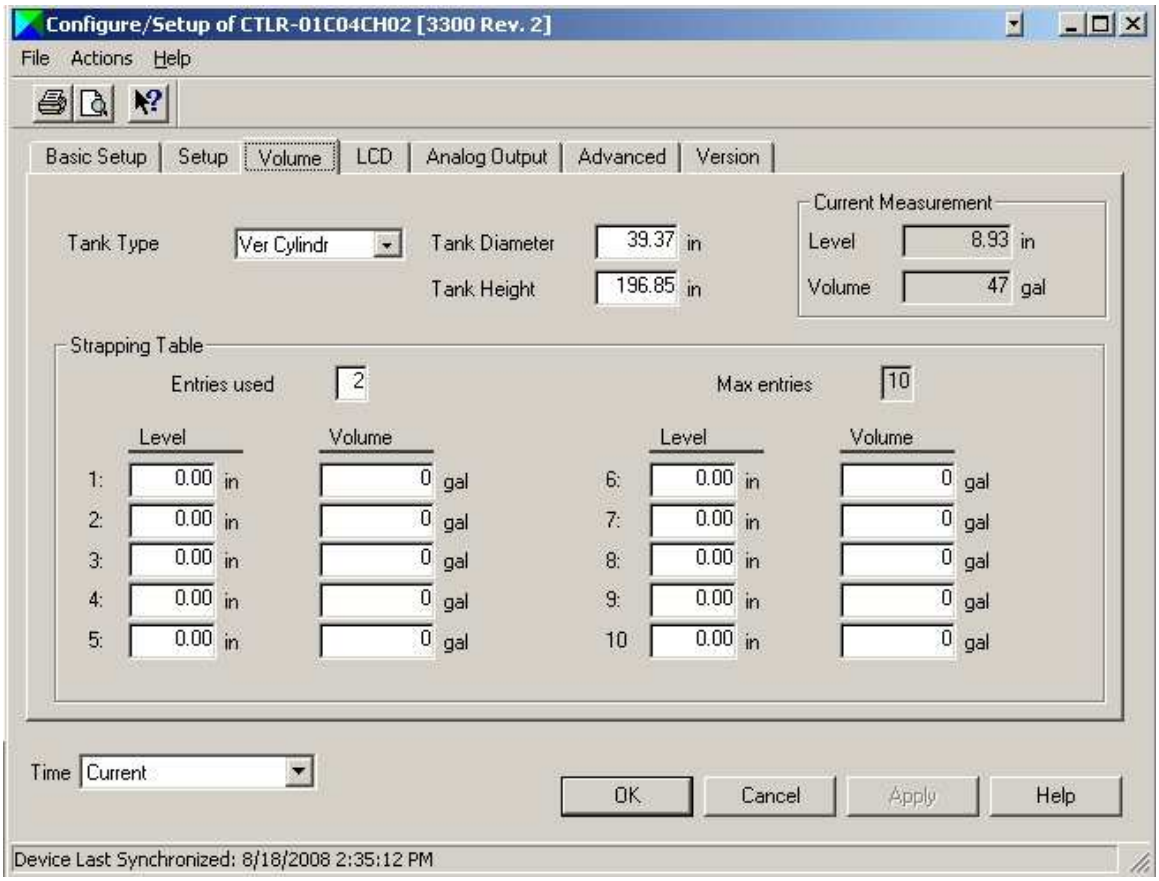
To best generate the coordinate points for a proper multi-point characterizer function, one must collect data on the storage vessel in the form of a *strapping table*. This entails emptying the vessel completely, then filling it with measured quantities of liquid, one sample at a time, and taking level readings:

Introduced liquid volume	Measured liquid level
150 gallons	2.46 feet
300 gallons	4.72 feet
450 gallons	5.8 feet
600 gallons	(etc., etc.)
750 gallons	(etc., etc.)

Each of these paired numbers would constitute the coordinates to be programmed into the characterizer function computer by the instrument technician or engineer:



Many "smart" level transmitter instruments possess enough computational power to perform the level-to-volume characterization directly, so as to transmit a signal corresponding directly to liquid volume rather than just liquid level. This eliminates the need for an external "level computer" to perform the necessary characterization. The following screenshot was taken from a personal computer running configuration software for a radar level transmitter[3], showing the strapping table data point fields where a technician or engineer would program the vessel's level-versus-volume piecewise function:



This configuration window actually shows more than just a strapping table. It also shows the option of calculating volume for different vessel shapes (*vertical cylinder* is the option selected here) including horizontal cylinder and sphere. In order to use the strapping table option, the user would have to select "Strapping Table" from the list of Tank Types. Otherwise, the level transmitter's computer will attempt to calculate volume from an ideal tank shape.

[3]The configuration software is Emerson's AMS, running on an engineering workstation in a DeltaV control system network. The radar level transmitter is a Rosemount model 3301 (guided-wave) unit.

24.3 Radiative temperature measurement

Temperature measurement devices may be classified into two broad types: *contact* and *non-contact*. Contact-type temperature sensors detect temperature by directly touching the material to be measured, and there are several varieties in this category. Non-contact temperature sensors work by detecting light emitted by hot objects.

Energy radiated in the form of electromagnetic waves (photons, or light) relates to object temperature by an equation known as the Stefan-Boltzmann equation, which tells us the rate of heat lost by radiant emission from a hot object is proportional to the *fourth power* of its absolute temperature:

$$P = e\sigma AT^4$$

Where,

P = Radiated energy power (watts)
e = Emissivity factor (unitless)
σ = Stefan-Boltzmann constant (5.67×10^{-8} W / m^2 $\cdot$ K^4)
A = Surface area (square meters)
T = Absolute temperature (Kelvin)

Solving for temperature (T) involves the use of the fourth root, to "un-do" the fourth power function inherent to the original function:

$$T = \sqrt[4]{\frac{P}{e\sigma A}}$$

Any optical temperature sensor measuring the emitted power (P) must "characterize" the power measurement using the above equation to arrive at an inferred temperature. This characterization is typically performed inside the temperature sensor by a microcomputer.

24.4 Analytical measurements

A great many chemical composition measurements may be made indirectly by means of electricity, if those measurements are related to the concentration of ions (electrically charged molecules). Such measurements include:

- pH of an aqueous solution
- Oxygen concentration in air
- Ammonia concentration in air
- Lead concentration in water

The basic principle works like this: two different chemical samples are placed in close proximity to each other, separated only by an *ion-selective membrane* able to pass the ion of interest. As the ion activity attempts to reach equilibrium through the membrane, an electrical voltage is produced across that membrane. If we measure the voltage produced, we can infer the relative activity of the ions on either side of the membrane.

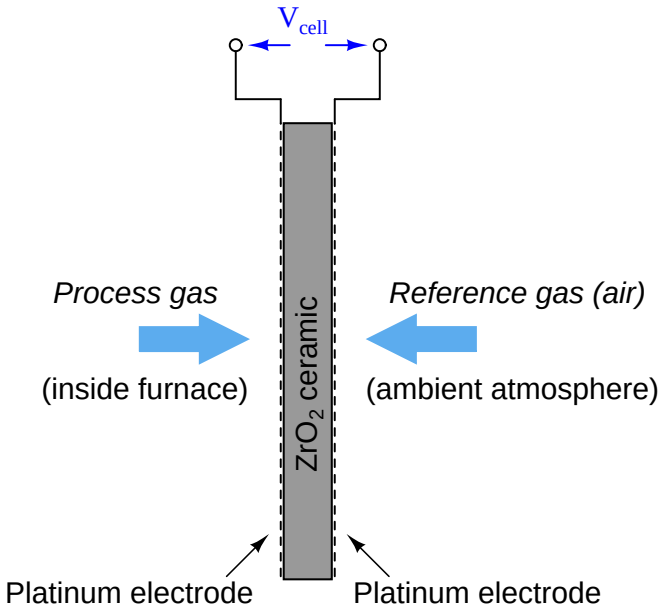
Not surprisingly, the function relating ion activity to the voltage generated is nonlinear. The standard equation describing the relationship between ionic activity on both sides of the membrane and the voltage produced is called the *Nernst equation*:

$$V = \frac{RT}{nF} \ln \left(\frac{a_1}{a_2} \right)$$

Where,
V = Electrical voltage produced across membrane due to ion exchange (volts)
R = Universal gas constant (8.315 J/mol·K)
T = Absolute temperature (Kelvin)
n = Number of electrons transferred per ion exchanged (unitless)
F = Faraday constant (96,485 coulombs per mole)
a_1 = Activity of ion in measured sample
a_2 = Activity of ion in reference sample (on other side of membrane)

A practical application for this technology is in the measurement of oxygen concentration in the flue gas of a large industrial burner, such as what might be used to heat up water to generate steam. The measurement of oxygen concentration in the exhaust of a combustion heater (or boiler) is very important both for maximizing fuel efficiency and for minimizing pollution (specifically, the production of NO_x molecules). Ideally, a burner's exhaust gas will contain no oxygen, having consumed it all in the process of combustion with a perfect stoichiometric mix of fuel and air. In practice, the exhaust gas of an efficiently-controlled burner will be somewhere near 2%, as opposed to the normal 21% of ambient air.

One way to measure the oxygen content of hot exhaust is to use a *high-temperature zirconium oxide* detector. This detector is made of a "sandwich" of platinum electrodes on either side of a solid zirconium oxide electrolyte. One side of this electrochemical cell is exposed to the exhaust gas (process), while the other side is exposed to heated air which serves as a reference:



The electrical voltage generated by this "sandwich" of zirconium and platinum is sent to an electronic amplifier circuit, and then to a microcomputer which applies an inverse function to the measured voltage in order to arrive at an inferred measurement for oxygen concentration. This type of chemical analysis is called *potentiometric*, since it measures ("metric") based on an electrical voltage ("potential").

The Nernst equation is an interesting one to unravel, to solve for ion activity in the sample (a_1) given voltage (V):

$$V = \frac{RT}{nF} \ln \left(\frac{a_1}{a_2} \right)$$

Multiplying both sides by nF:

$$nFV = RT \ln \left(\frac{a_1}{a_2} \right)$$

Dividing both sides by RT:

$$\frac{nFV}{RT} = \ln \left(\frac{a_1}{a_2} \right)$$

Applying the rule that the difference of logs is equal to the log of the quotient:

$$\frac{nFV}{RT} = \ln a_1 - \ln a_2$$

Adding $\ln a_2$ to both sides:

$$\frac{nFV}{RT} + \ln a_2 = \ln a_1$$

Making both sides of the equation a power of e:

$$e^{\frac{nFV}{RT} + \ln a_2} = e^{\ln a_1}$$

Canceling the natural log and exponential functions on the right-hand side:

$$e^{\frac{nFV}{RT} + \ln a_2} = a_1$$

In most cases, the ionic activity of a_2 will be relatively constant, and so $\ln a_2$ will be relatively constant as well. With this in mind, we may simplify the equation further, using k as our constant value:

Substituting k for $\ln a_2$:

$$e^{\frac{nFV}{RT} + k} = a_1$$

Applying the rule that the sum of exponents is the product of powers:

$$e^k e^{\frac{nFV}{RT}} = a_1$$

If k is constant, then e^k will be constant as well (calling the new constant C):

$$Ce^{\frac{nFV}{RT}} = a_1$$

Analytical instruments based on potentiometry must evaluate this inverse function to "undo" the Nernst equation to arrive at an inferred measurement of ion activity in the sample given the small voltage produced by the sensing membrane. These instruments typically have temperature sensors as well built in to the sensing membrane assembly, since it is apparent that temperature (T) also plays a role in the generation of this voltage. Once again, this mathematical function is typically evaluated in a microprocessor.

References

Lipták, Béla G., *Instrument Engineers' Handbook – Process Measurement and Analysis Volume I*, Fourth Edition, CRC Press, New York, NY, 2003.

Stewart, James, *Calculus: Concepts and Contexts*, 2nd Edition, Brooks/Cole, Pacific Grove, CA, 2001.

Chapter 25

Final control elements

25.1 Control valves

One of the most common final control elements in industrial control systems is the *control valve*. A "control valve" works to restrict the flow of fluid through a pipe at the command of an automated signal, such as the signal from a loop controller or logic device (such as a PLC). Some control valve designs are intended for discrete (on/off) control of fluid flow, while others are designed to *throttle* fluid flow somewhere between fully open and fully closed (shut), inclusive. The electrical equivalent of an on/off valve is a switch, while the electrical equivalent of a throttling valve is a variable resistor.

Control valves are comprised of two major parts: the *valve body*, which contains all the mechanical components necessary to influence fluid flow; and the *valve actuator*, which provides the mechanical power necessary to move the components within the valve body. Often times, the major difference between an on/off control valve and a throttling control valve is the type of actuator applied to the valve[1]: on/off actuators need only position a valve mechanism two one of two extreme positions (fully open or fully closed). Throttling actuators must be able to accurately position a valve mechanism anywhere between those extremes.

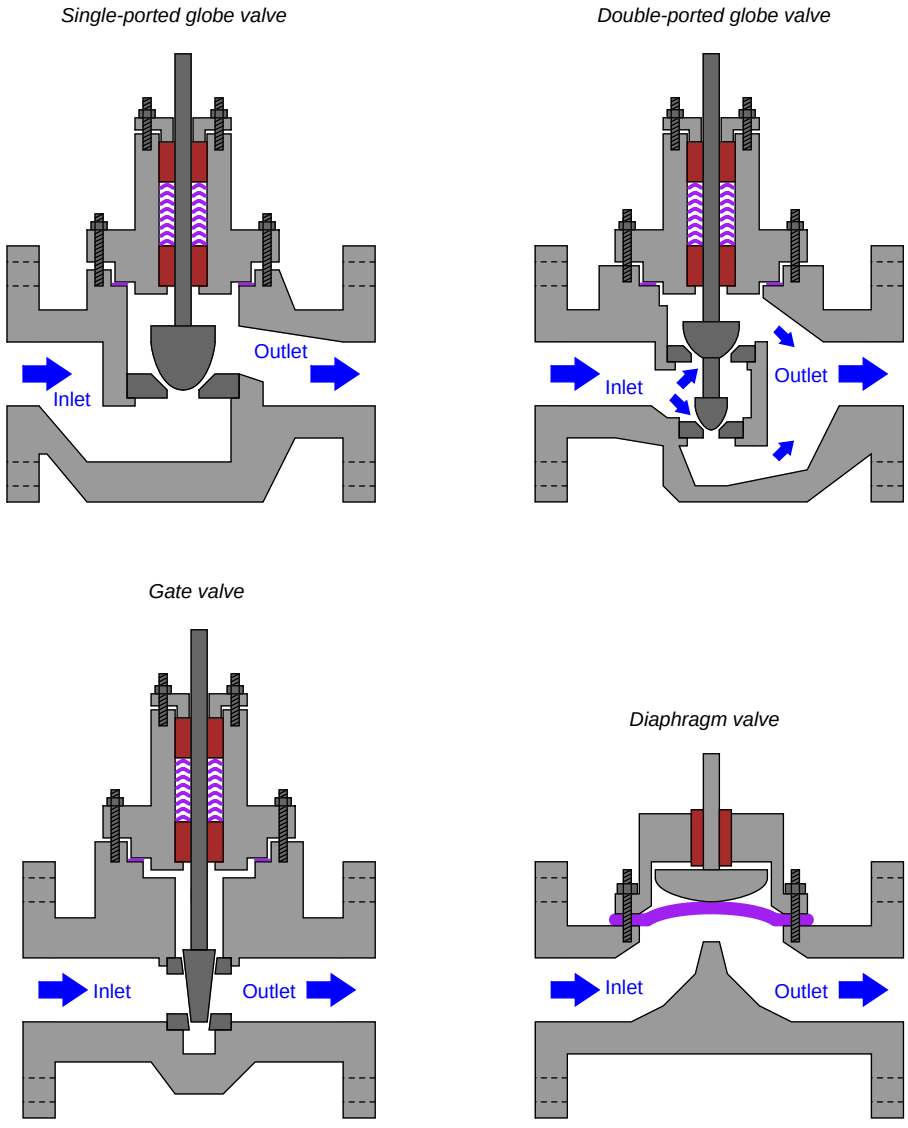
Within a control valve body, the specific components performing the work of throttling (or completely shutting off) of fluid flow are collectively referred to as the valve *trim*. For each major type of control valve, there are usually many variations of trim design. The choice of valve type, and of specific trim for any type of valve, is a decision dictated by the type of fluid being controlled, the nature of the control action (on/off versus throttling), the process conditions (expected flow rate, temperature, pressures, etc.), and economics.

An appendix of this book (appendix B, beginning on page 1743) photographically documents the complete disassembly of a typical control valve. The valve happens to be a Fisher E-body globe valve with a pneumatic diaphragm actuator.

[1] To be honest, there are some valve body designs that work far better in on/off service (e.g. ball valves and plug valves) while other designs do a better job at throttling (e.g. double-ported globe valves). Many valve designs, however, may be pressed into either type of service merely by attaching the appropriate actuator.

25.1.1 Sliding-stem valves

A *sliding-stem* valve body is one that actuates with a linear motion. Some examples of sliding-stem valve body designs are shown here:

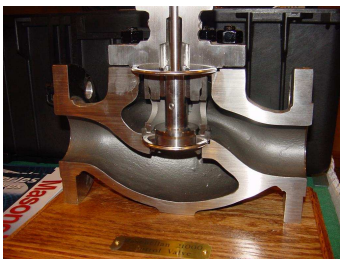


Most sliding-stem control valves are *direct acting*, which means the valve opens up wider as the stem is drawn out of the body. Conversely, a direct-acting valve shuts off (closes) when the stem is pushed into the body. Of course, a *reverse-acting* valve body would behave just the opposite: opening up as the stem is pushed in and closing off as the stem is drawn out.

Globe valves

Globe valves restrict the flow of fluid by altering the distance between a movable plug and a stationary seat (in some cases, a pair of plugs and matching seats). Fluid flows through a hole in the center of the seat, and is more or less restricted by how close the plug is to that hole. The globe valve design is one of the most popular sliding-stem valve designs used in throttling service. A photograph of a small (2 inch) globe valve body appears here:

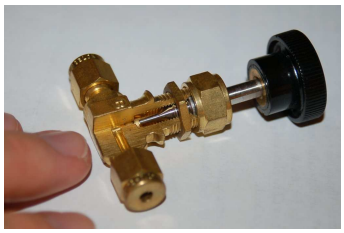
A set of three photographs showing a cut-away Masoneilan model 21000 globe valve body illustrates just how the moving plug and stationary seat work together to throttle flow in a direct-acting globe valve. The left-hand photo shows the valve body in the fully closed position, while the middle photo shows the valve half-open, and the right-hand photo shows the valve fully open:



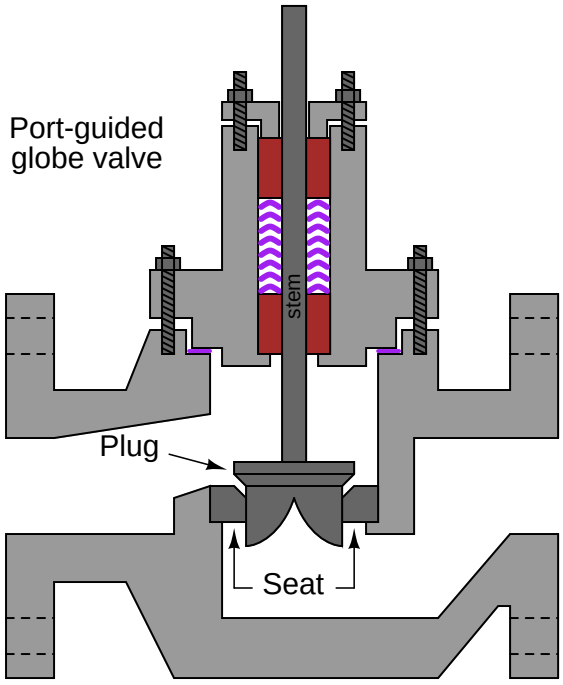
As you can see from these photographs, the valve plug is guided by the stem so it always lines up with the centerline of the seat. For this reason, this particular style of globe valve is called a *stem-guided* globe valve.

A variation on the stem-guided globe valve design is the *needle valve*, where the plug is extremely small in diameter and usually fits well into the seat hole rather than merely sitting on top of it. Needle valves are very common as manually-actuated valves used to control low flow rates of air or oil. A set of three photographs shows a needle valve in the fully-closed, mid-open, and fully-open positions (left-to-right):

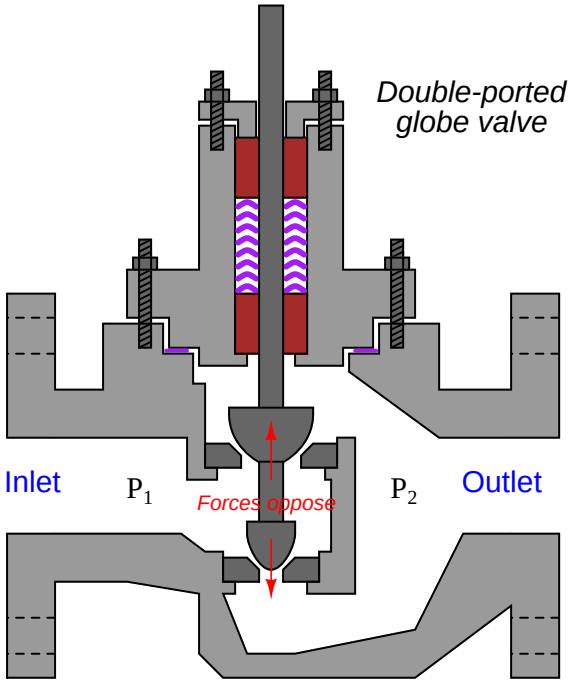




Yet another variation on the globe valve is the *port-guided* valve, where the plug has an unusual shape that projects into the seat. Thus, the seat ring acts as a guide for the plug to keep the centerlines of the plug and seat always aligned, minimizing guiding stresses that would otherwise be placed on the stem. This means that the stem may be made smaller in diameter than if the valve trim were stem-guided, minimizing sliding friction and improving control behavior.

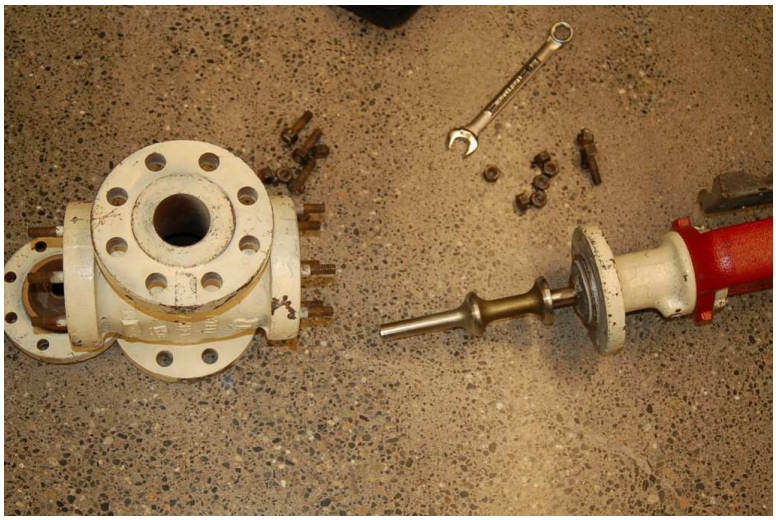


Some globe valves use a pair of plugs (on the same stem) and a matching pair of seats to throttle fluid flow. These are called *double-ported* globe valves. The purpose of a double-ported globe valve is to minimize the force applied to the stem by process fluid pressure across the plugs:



Differential pressure of the process fluid ($P_1 - P_2$) across a valve plug will generate a force parallel to the stem as described by the formula $F = PA$, with A being the plug's effective area presented for the pressure to act upon. In a single-ported globe valve, there will only be one force generated by the process pressure. In a double-ported globe valve, there will be *two opposed* force vectors, one generated at the upper plug and another generated at the lower plug. If the plug areas are approximately equal, then the forces will likewise be approximately equal and therefore nearly cancel. This makes for a control valve that is easier to actuate (i.e. the stem position is less affected by process fluid pressures).

The following photograph shows a disassembled Fisher "A-body" double-ported globe valve, with the double plug plainly visible on the right:



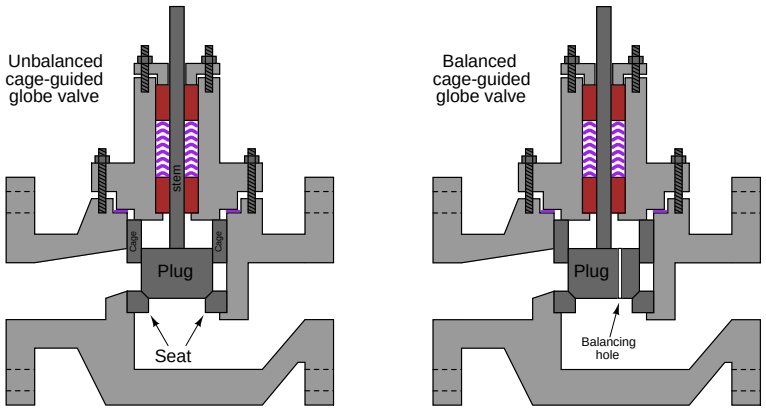
While double-ported globe valves certainly enjoy the advantage of easier actuation compared to their single-ported cousins, they also suffer from a distinct disadvantage: the near impossibility of tight shut-off. With *two* plugs needing to come to simultaneous rest on *two* seats to achieve a fluid-tight seal, there is precious little room for error or dimensional instability. Even if a double-ported valve is prepared in a shop for the best shut-off possible[2], it may not completely shut off when installed due to dimensional changes caused by process fluid heating or cooling the valve stem and body. This is especially problematic when the stem is made of a different material than the body. Globe valve stems are commonly manufactured from stainless steel bar stock, while globe valve bodies are commonly cast of iron. Cold-formed stainless steel has a different coefficient of thermal expansion than hot-cast iron, which means the plugs will no longer simultaneously seat once the valve warms or cools much from the temperature it was at when it seated tightly.

[2]The standard preparatory technique is called *lapping*. To "lap" a valve plug and seat assembly, an abrasive paste known as *lapping compound* is applied to the valve plug(s) and seat(s) at the areas of mutual contact when the valve is disassembled. The valve mechanism is reassembled, and the stem is then rotated in a cyclic motion such that the plug(s) grind into the seat(s), creating a matched fit. The precision of this fit may be checked by disassembling the valve, cleaning off all remaining lapping compound, applying a metal-staining compound such as *Prussian blue*, then reassembling. The stem is rotated once more such that the plug(s) will rub against the seat(s), wearing through the applied stain. Upon disassembly, the worn stain may be inspected to reveal the extend of metal-to-metal contact between the plug(s) and the seat(s). If the contact area is deemed insufficient, the lapping process may be repeated.

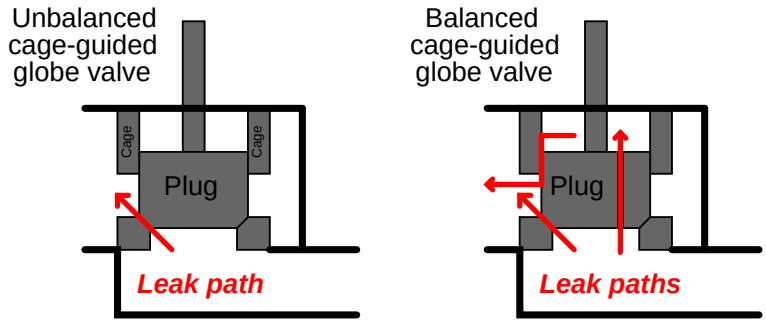
A more modern version of the globe valve design uses a piston-shaped plug inside a surrounding *cage* with ports cast or machined into it. These *cage-guided* globe valves throttle flow by uncovering more or less of the port area in the surrounding cage as the plug moves up and down. The cage also serves to guide the plug so the stem need not be subjected to lateral forces as in a stem-guided valve design. A photograph of a cut-away control valve shows the appearance of the cage (in this case, with the plug in the fully closed position). Note the "T"-shaped ports in the cage, through which fluid flows as the plug moves up and out of the way:

An advantage of the cage-guided design is that the valve's flowing characteristics may be easily altered just by replacing the cage with another having different size or shape of holes. Many different cage styles are available for certain plug (piston) sizes, which means the plug need not be replaced while changing the cage. This is decidedly more convenient than the plug change necessary for changing characteristics of stem-guided or port-guided globe valve designs.

Cage-guided globe valves are available with both *balanced* and *unbalanced* plugs. A balanced plug has one or more ports drilled from top to bottom, allowing fluid pressure to equalize on both sides of the plug. This helps minimize the forces acting on the plug which must be overcome by the actuator:



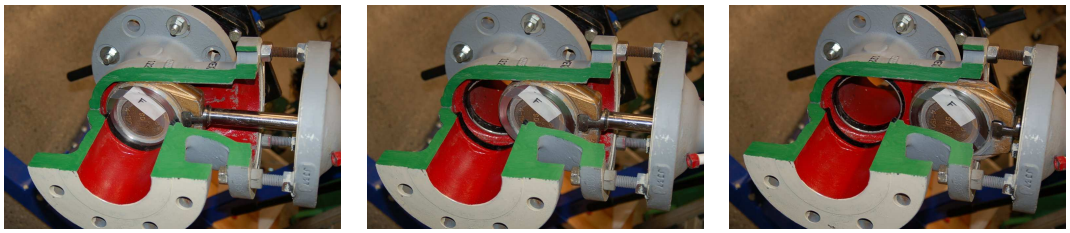
Unbalanced plugs generate a force equal to the product of the differential pressure across the plug and the plug's area ($F = PA$), which may be quite substantial in some applications. Balanced plugs do not generate this same force because they equalize the pressure on both sides of the plug, however, they exhibit the disadvantage of one more leak path when the valve is in the fully closed position (through the balancing ports, past the piston ring, and out the cage ports):



Gate valves

Gate valves work by inserting a dam ("gate") into the path of the flow to restrict it, in a manner similar to the action of a sliding door. Gate valves are more often used for on/off control than for throttling.

The following set of photographs shows a hand-operated gate valve (cut away and painted for use as an instructional tool) in three different positions, from full closed to full open (left to right):

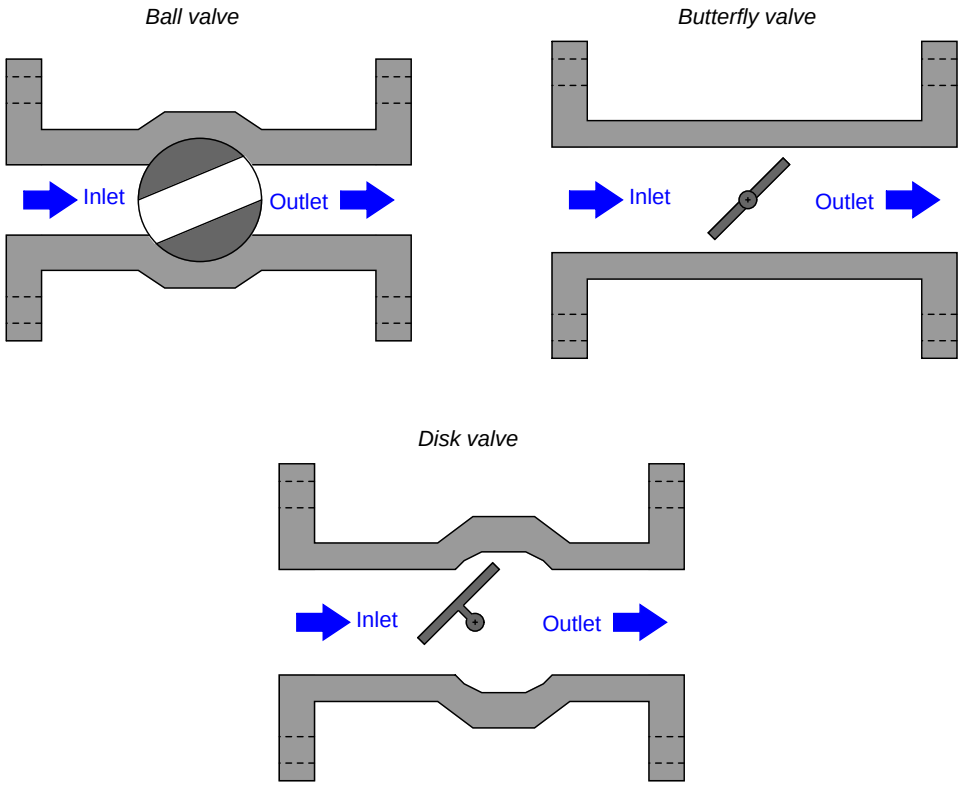


Diaphragm valves

Diaphragm valves use a flexible sheet pressed close to the edge of a solid dam to narrow the flow path for fluid. These valves are well suited for flows containing solid particulate matter such as slurries, although precise throttling may be difficult to achieve due to the elasticity of the diaphragm. The next photograph shows a diaphragm valve actuated by an electric motor, used to control the flow of treated sewage:

25.1.2 Rotary-stem valves

A different strategy for controlling the flow of fluid is to insert a rotary element into the flow path. Instead of sliding a stem into and out of the valve body to actuate a throttling mechanism, rotary valves rely on the rotation of a shaft to actuate the trim. An important advantage of rotary control valves over sliding-stem designs such as the globe valve and diaphragm valve is a virtually obstructionless path for fluid when the valve is wide-open[3].



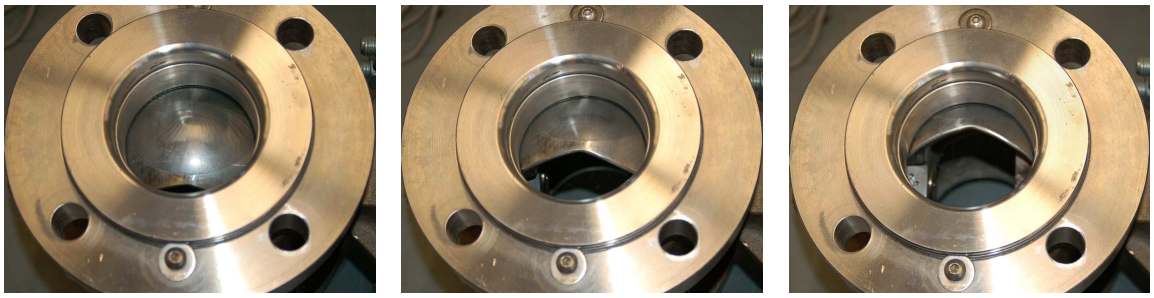
[3]Of course, gate valves also offer obstructionless flow when wide-open, but their poor throttling characteristics give most rotary valve designs the overall advantage.

Ball valves

In the ball valve design, a spherical ball with a passageway cut through the center rotates to allow fluid more or less access to the passageway. When the passageway is parallel to the direction of fluid motion, the valve is wide open; when the passageway is aligned perpendicular to the direction of fluid motion, the valve is fully shut (closed).

The following set of photographs shows a hand-operated ball valve in three different positions, from nearly full closed to nearly full open (left to right):

Simple ball valves with full-sized bores in the rotating ball are generally better suited for on/off service than for throttling (partially-open) service. A better design of ball valve for throttling service is the *characterized* or *segmented* ball valve, shown in various stages of opening in the following set of photographs:



The V-shaped notch cut into the opening lip of the ball provides a narrower area for fluid flow at low opening angles, providing more precise flow control than a plain-bore ball valve.

Butterfly valves

Butterfly valves are quite simple to understand: the "butterfly" element is a disk that rotates perpendicular to the path of fluid flow. When parallel to the axis of flow, the disk presents minimal obstruction; when perpendicular to the axis, the disk completely blocks any flow. Fluid-tight shut-off is difficult to obtain in the classic butterfly design unless the seating area is lined with a soft (elastic) material.

Disk valves

Disk valves (often referred to as *eccentric disk valves*, or as *high-performance butterfly valves*) are a variation on the butterfly design intended to improve seat shut-off. The disk's center is offset from the shaft centerline, causing it to approach the seat with a "cam" action that results in high seating pressure. Thus, tight shut-off of flow is possible even when using metal seats and disks.

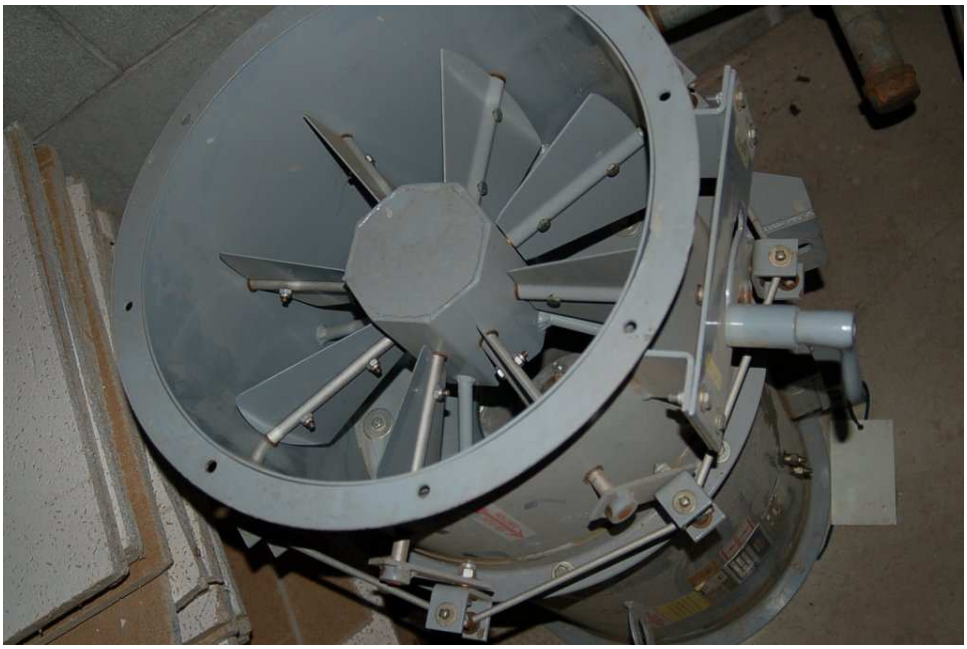
The following photograph shows the body of a Fisher E-plug control valve, with the disk in a partially-open position:

25.1.3 Dampers and louvres

A *damper* (otherwise known as a *louvre*) is a multi-element flow control device generally used to throttle large flows of air at low pressure. Dampers find common application in furnace and boiler draft control, and in HVAC (Heating, Ventilation, and Air Conditioning) systems.

Common damper designs include parallel and radial. Parallel-vane dampers resemble a Venetian blind, with multiple rectangular vanes synchronously rotated to throttle flow through a rectangular opening. A photograph of a parallel-vane damper is shown here, part of an induced-draft (suction) air fan system on a separator at a cement plant. The vanes are not visible in this photograph because they reside inside the metal air duct, but the actuator mechanism and linkages connecting seven vane shafts together are:

Radial-vane dampers use multiple vanes arranged like petals of a flower to throttle flow through a circular opening. A photograph of a radial-vane damper is shown here (note the levers and linkages on the periphery of the tube, synchronizing the motions of the eight vanes so they rotate at the same angle):

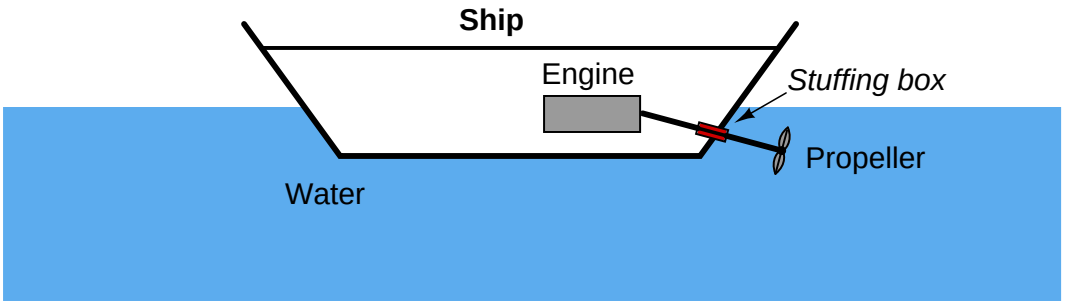


Dampers find use in many non-industrial applications as well. Take for instance these greenhouse vents, actuated by pneumatic (air-powered) piston actuators:

25.1.4 Valve packing

Regardless of valve type, all stem-actuated control valves require some form of seal allowing motion of the stem from some external device (an *actuator*) while sealing process fluid so no leaks occur between the moving stem and the body of the valve. The general term for this sealing mechanism is *packing*.

This mechanical feature is not unlike the *stuffing box* used to seal seawater from entering a boat or ship at the point where the propeller shaft penetrates the hull:



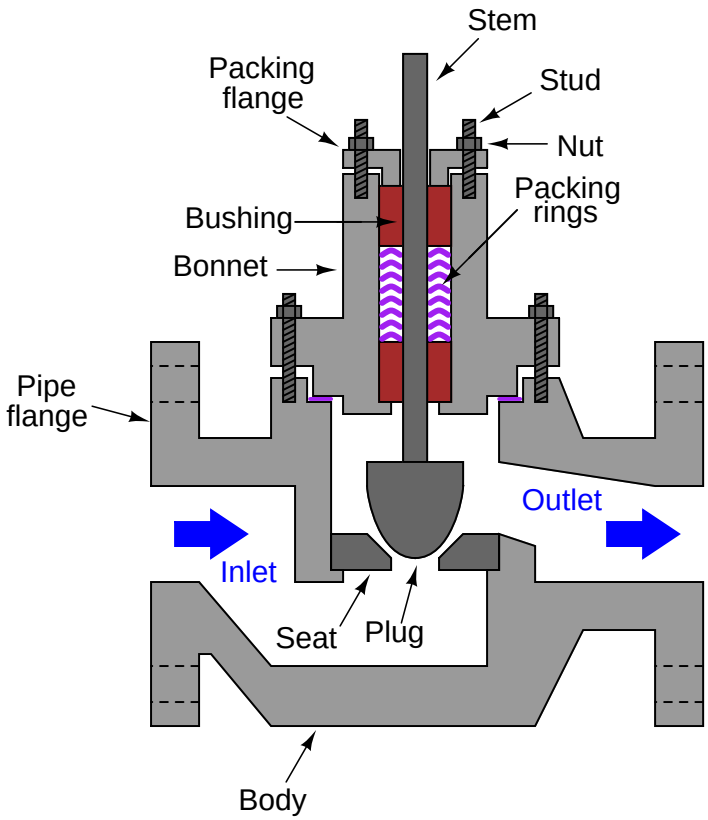
The fundamental problem is the same for the ship as it is for the control valve: how to allow a moving shaft to pass through what needs to be an impenetrable barrier to some fluid (in the case of the ship's hull, the fluid is seawater). The solution is to wrap the shaft in a flexible material that maintains a close fit to the shaft without binding its motion. A traditional packing material for ship propeller shafts is *flax*. Some form of lubrication is usually provided so this packing material does not impose excessive friction on the shaft's motion[4].

Modern marine stuffing boxes use advanced materials such as Teflon (PTFE) or graphite instead of flax, which wear longer and leak less water. In the world of control valves, the traditional packing material used to be asbestos (shaped into rings or ropes, much like flax used to be shaped for use in stuffing boxes), but is now commonly Teflon or graphite as well.

In the case of a ship's stuffing box, a little bit of water leakage is not a problem since all ships are equipped with bilge pumps to pump out collected water over time. However, leakage is simply unacceptable in many industrial control valve applications where we must minimize *fugitive emissions*. A "fugitive emission" is any unwanted escape of process substance into the surrounding environment, usually from leaks around pump and valve shafts. Special "environmental" packing sets are available for control valve applications where this is a concern.

[4] Some packing materials, most notably Teflon and graphite, tend to be *self-lubricating*.

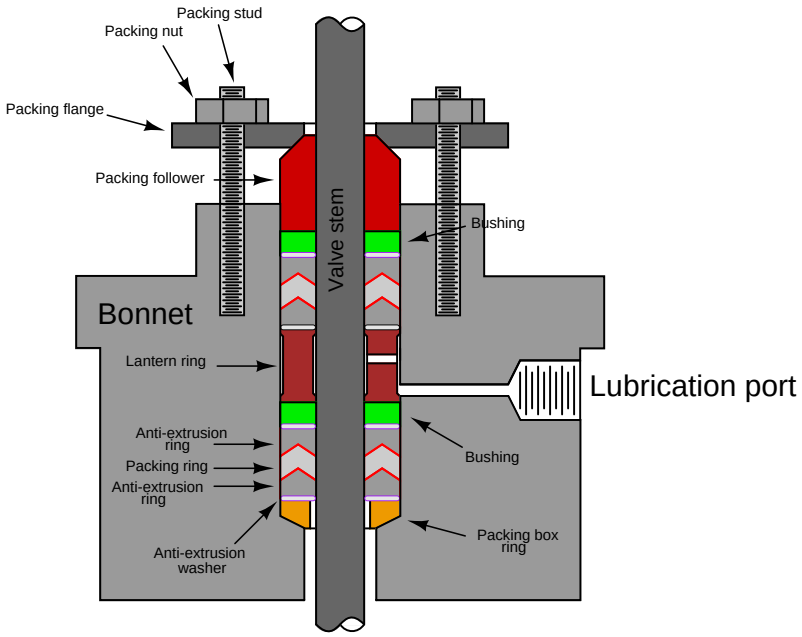
Packing in a sliding-stem valve fits in a section of the valve body called the *bonnet*, shown in this simplified diagram of a single-ported, stem-guided globe valve:



Here, the packing material takes the form of several concentric rings, stacked on the valve stem like washers on a bolt. These packing rings are forced down from above by the *packing flange* to apply a compressive force around the circumference of the valve stem.

Two nuts threaded onto studs maintain proper force on the packing rings. Care must be taken not to over-tighten these nuts and over-compress the packing material, or else the packing will create excessive friction on the valve stem. Not only will this friction impede precise valve stem motion, but it will also create undue wear on the stem and packing, increasing the likelihood of future packing leakage.

A closer look at the bonnet shows a multitude of components working together to form a low-friction, pressure-tight seal for the moving valve stem:



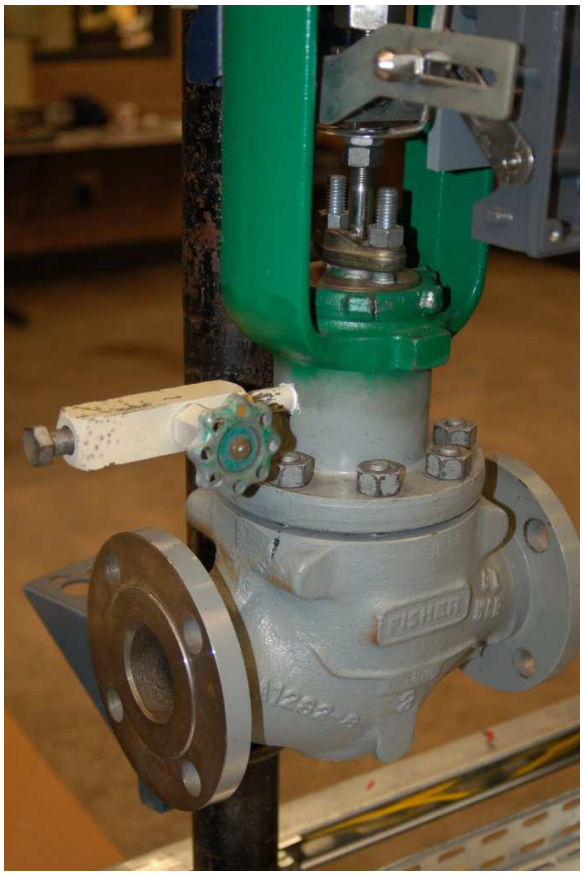
In this diagram, we see two sets of packing rings separated by a metal piece called a *lantern ring*. The lantern ring acts as a spacer allowing lubricant introduced through the lubrication port to enter into both packing sets from the middle of the bonnet.

Photographs taken of an actual valve packing assembly removed from the bonnet (left), and reassembled on the valve stem (right) reveal the structure of the packing and associated components.

There is no lantern ring in this packing assembly, but there is a coil spring[5]. This makes it a *live-loaded* packing as opposed to a *jam* packing. Jam packings (without springs) must be periodically adjusted to compensate for compressive fatigue and/or wear of the packing rings.

[5] An alternative to coil-shaped springs for live-loading of valve packing are *Belleville washers*. These washers are made of spring steel and have a slightly concave shape, giving them resistance to compression along the shaft axis. Belleville washers are always stacked in opposed pairs.

In packing applications requiring external lubrication, a *stem packing lubricator* may be connected to the lubrication port on the bonnet. This device uses a long, threaded bolt as a piston to push a quantity grease into the packing assembly:



To operate a lubricator, the hand valve on the lubricator is first secured in the closed (shut) position, then the bolt is fully unscrewed until it falls out of the lubricator body. An appropriate lubricating grease is squeezed into the bolt hole in the lubricator body, and the bolt threaded back into place until hand-tight. Using a wrench or socket to tighten the bolt a bit more (generating pressure in the grease) and opening the hand valve allows grease to enter the packing chamber. The bolt is then screwed in fully, pushing the entire quantity of grease into the packing. As a final step, the hand valve is fully shut so there is no way for process liquid to leak out past the bolt threads.

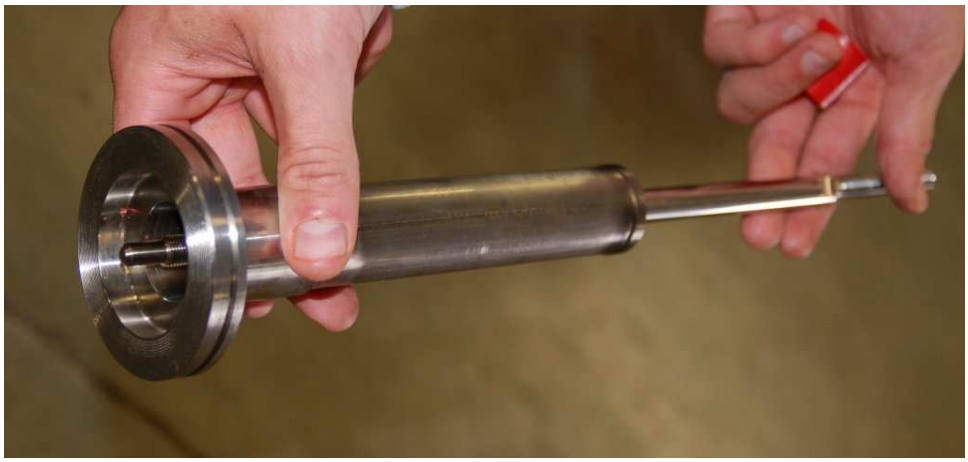
The two most common packing materials in use today are Teflon (PTFE) and graphite. Teflon is the better of the two with regard to friction and stem wear[6]. Teflon is also quite resistant to attack from a wide variety of chemical substances. Unfortunately, it has a limited temperature range and cannot withstand intense nuclear radiation (making it unsuitable for use near reactors in

[6]Based on friction values shown on page 131 of Fisher's *Control Valve Handbook* (Third Edition), Teflon packing friction is typically 5 to 10 times less than graphite packing for the same stem size!

nuclear power plants). Graphite is another self-lubricating packing material, and it has a far greater temperature range than Teflon[7] as well as the ability to withstand harsh nuclear radiation, but creates much more stem friction than Teflon. Graphite packing also has the unfortunate property of permitting *galvanic corrosion* between the stem and bonnet metals due to its electrical conductivity. Sacrificial zinc washers are sometimes added to graphic packing assemblies to help mitigate this corrosion, but this only postpones rather than prevents corrosive damage to the stem.

Hybrid packing materials, such as carbon-reinforced Teflon, also exist in an attempt to combine the best characteristics of both technologies.

A completely different approach to packing is a device called a *bellows seal*: an accordion-like metal tube fastened to the valve stem and to the bonnet, forming a leak-proof seal with negligible friction. The accordion ribs give the bellows seal an ability to stretch and compress with a sliding stem's linear motion. Since the bellows is an uninterrupted metal tube, there is no place at all for leaks to develop:



The accordion-shaped bellows is contained and protected inside the thick metal tube visible in this photograph. One end of the bellows is welded to the valve stem, and the other end is welded to the protective tube. With the tube firmly clamped in the bonnet of the valve, a leak-free seal exists.

While a properly operating bellows seal is indeed leak-free, there is always the potential of a rupture in the bellows, which could develop into a leak of catastrophic proportions. For this reason, bellows seals are almost always followed by standard packing around the stem. In the unlikely event of a bellows rupture, this standard packing will provide a serviceable seal until the bellows can be replaced.

[7] Graphite packing is usable in services ranging from cryogenic temperatures to 1200 degrees Fahrenheit, as opposed to Teflon which is typically rated between -40^o F and 450^o F.

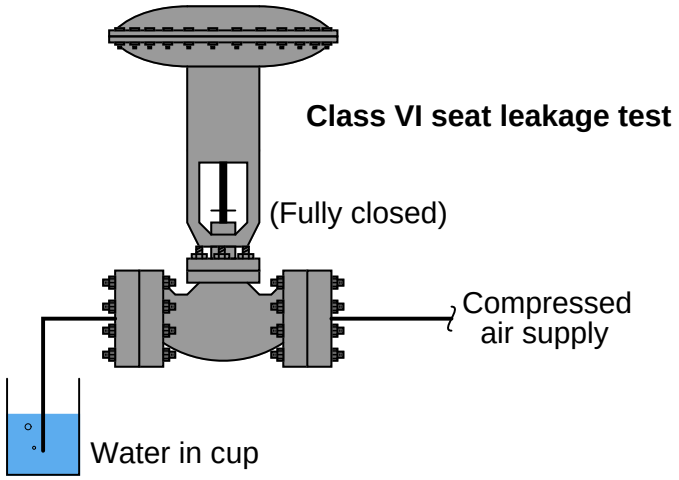
25.1.5 Valve seat leakage

In many process applications, it is important that the control valve be able to completely stop fluid flow when placed in the "closed" position. Although this may seem to be a fundamental requirement of any valve, it is not necessarily so. Many control valves spend most of their operating lives in a partially-open state, rarely opening or closing fully. Additionally, some control valve designs are notorious for the inability to completely shut off (e.g. double-ported globe valves). Given the common installation of manual "block" valves upstream and downstream of a control valve, there is usually a way to secure zero flow through a pipe even if a control valve is incapable of tight shut-off. For some control valve applications, however, tight shut-off is a mandatory requirement.

For this reason we have several classifications for control valves, rating them in their ability to fully shut off. Seat leakage tolerances are given roman numeral designations, as shown in this table[8]:

Class	Maximum allowable leakage rate	Test pressure drop
I	(no specification given)	(no specification given)
II	0.5% of rated flow capacity, air or water	45-60 PSI or max. operating
III	0.1% of rated flow capacity, air or water	45-60 PSI or max. operating
IV	0.01% of rated flow capacity, air or water	45-60 PSI or max. operating
V	0.0005 ml/min water per inch orifice size per PSI	Max. operating
VI	Bubble test, air or nitrogen	50 PSI or max. operating

The "bubble test" used for Class VI seat leakage is based on the leakage rate of air or nitrogen gas past the closed valve seat as measured by counting the rate of gas bubbles escaping a bubble tube submerged under water. For a 6 inch valve, this maximum bubble rate is 27 bubbles per minute (or about 1 bubble every two seconds):



It is from this leakage test procedure that the term *bubble-tight shut-off* originates. Class VI shut-off is often achievable only through the use of "soft" seat materials such as Teflon rather than

[8]Data in this table taken from Fisher's *Control Valve Handbook*.

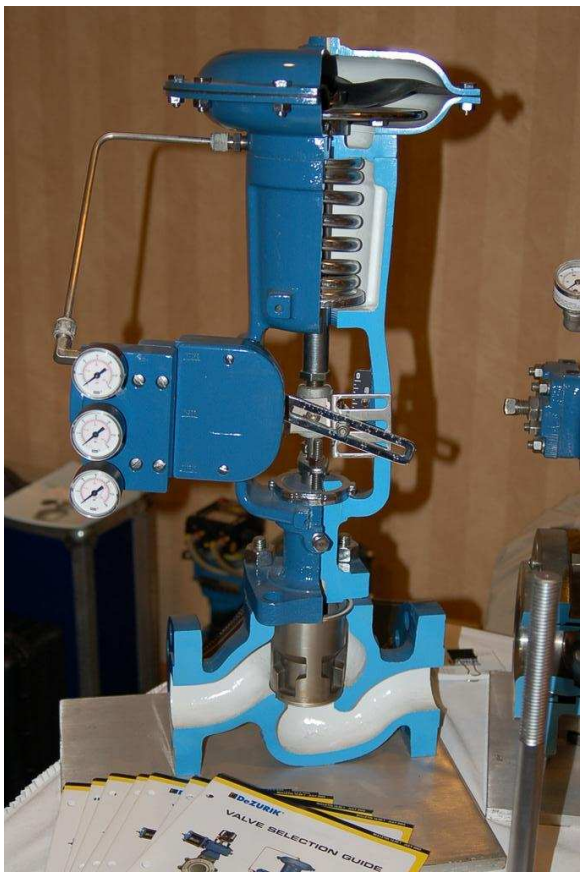
hard metal-to-metal contact between the valve plug and seat. Of course, this method of achieving bubble-tight shut-off comes at the price of limited operating temperature range and the inability to withstand nuclear radiation exposure.

25.1.6 Control valve actuators

The purpose of a control valve *actuator* is to provide the motive force to operate a valve mechanism. Both sliding-stem and rotary control valves enjoy the same selection of actuators: *pneumatic*, *hydraulic*, *electric motor*, and *hand* (manual).

Pneumatic actuators

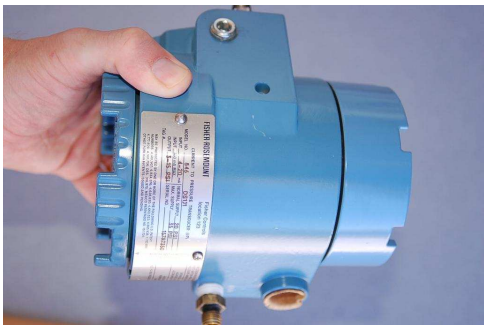
Pneumatic actuators use air pressure pushing against either a flexible diaphragm or a piston to move a valve mechanism. The following photograph shows a cut-away control valve, with a pneumatic diaphragm actuator mounted above the valve body. You can see the large coil spring providing default positioning of the valve (air pressure acting against the diaphragm moves the valve against the spring) and the rubber diaphragm at the very top. Air pressure applied to the bottom side of the diaphragm lifts the sliding stem of the valve in the upward direction, against the spring's force which tries to push the stem down:



The air pressure required to motivate a pneumatic actuator may come directly from the output of a pneumatic process controller, or from a *signal transducer* (or *converter*) translating an electrical signal into an air pressure signal. Such transducers are commonly known as *I/P* or "I to P"

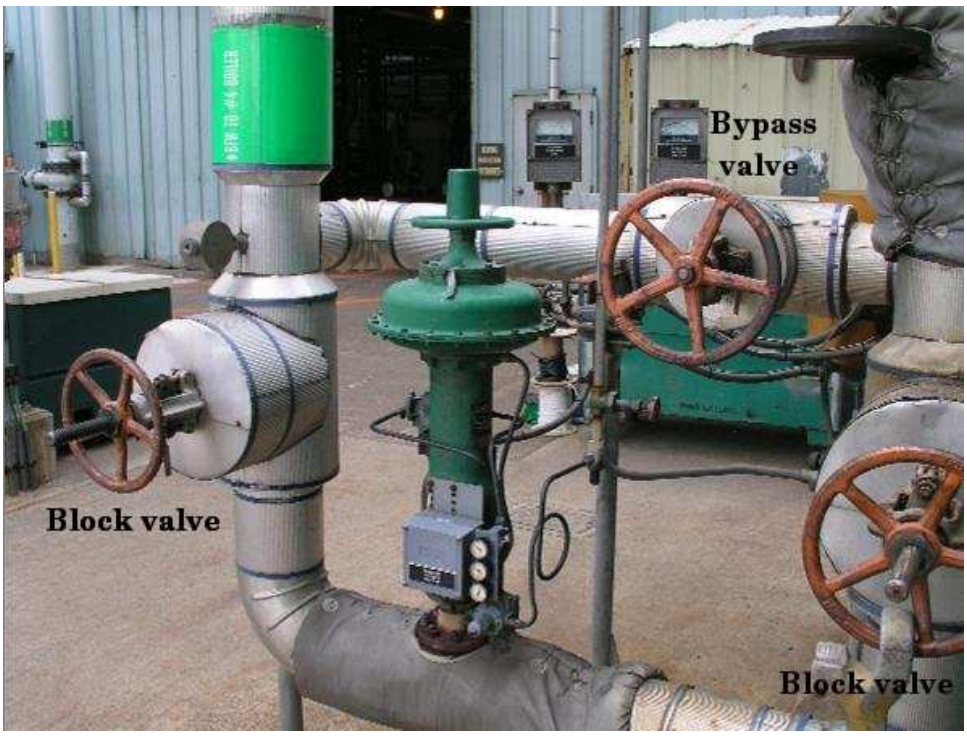
converters, since they typically translate an electric current signal (I) of 4 to 20 mA DC into an air pressure signal (P) of 3 to 15 PSI.

The following photographs show I/P transducers of different make and model. A Fisher model 846 appears in the upper-left photograph, while an older Fisher model 546 appears in the upper-right (with cover removed). A Foxboro model E69F I/P appears in the lower-left photograph, while a Moore Industries model IPT appears in the lower-right:



Despite their differing designs and appearances, they all function the same: accepting an analog DC current signal input and a clean supply air pressure of about 20 PSI, outputting a variable air pressure signal proportional to the electric current input. An interesting feature to compare between these four I/P transducers is their relative ruggedness. Every transducer shown except the Moore Industries model (lower-right) is built to withstand direct exposure to a process atmosphere, hence the heavy cast-metal housings and electrical conduit fittings. The Moore Industries unit is intended for a sheltered location, and may be plugged in to a "manifold" with several other I/P transducers to form a compact bank of transducers capable of driving air pressure signals to several valve actuators.

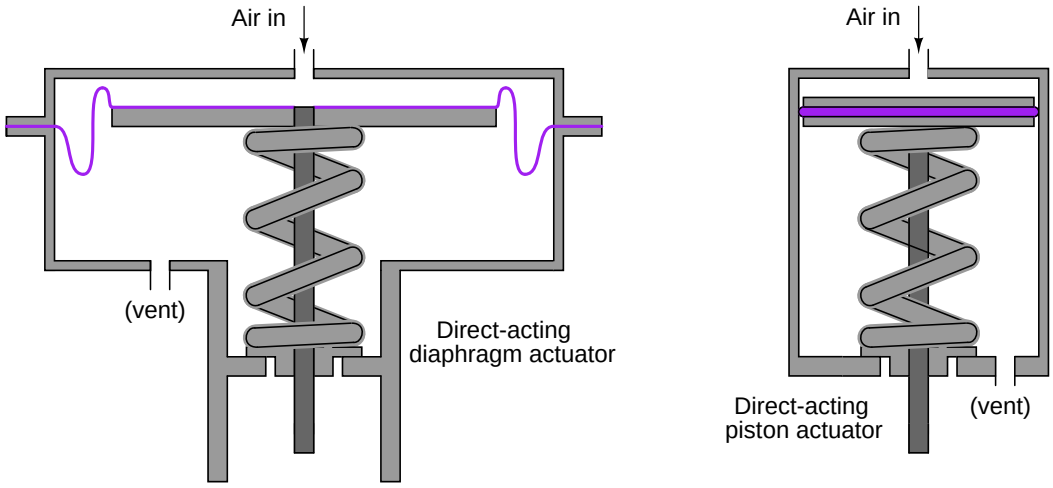
Some pneumatic valve actuators are equipped with *hand jacks* which are used to manually position the valve in the event of air pressure failure. The next photograph shows a sliding-stem control valve with pneumatic diaphragm actuator and a "handwheel" on the top:



Note the three manual valves located around the control valve: two to *block* flow through the control valve and one to *bypass* flow around the control valve in the event of control valve failure or maintenance. These manual valves happen to be of the *gate* design, with *rising-stem* actuators to clearly show their status (stem protruding = open valve ; stem hidden = closed valve). Such block-and-bypass manual valve arrangements are quite common in the process industries where control valves fulfill critical roles and some form of manual control is needed as an emergency alternative.

Note also the air pressure tubing between the valve actuator and the air supply pipe, bent into a loop. This is called a *vibration loop*, and it exists to minimize strain on the metal tubing from vibration that may occur.

Pneumatic actuators may take the form of pistons rather than diaphragms. Illustrations of each type are shown here for comparison:



Piston actuators generally have longer stroke lengths than diaphragm actuators, and are able to operate on much greater air pressures[9]. Since actuator force is a function of fluid pressure and actuator area ($F = PA$), this means piston actuators are able to generate more force than diaphragm actuators. The combination of greater force and greater displacement yields more work potential for piston actuators than diaphragm actuators of equivalent size, since mechanical work is the product of force and displacement ($W = Fx$).

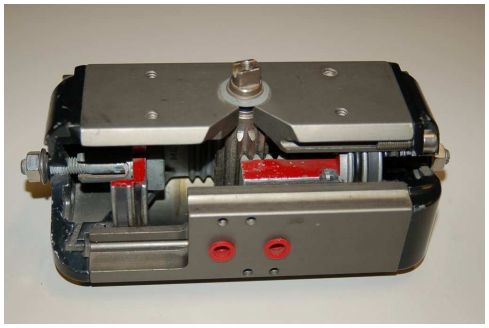
Perhaps the greatest disadvantage of piston actuators as applied to control valves is friction between the piston's pressure-sealing ring and the cylinder wall. This is not a problem for on/off control valves, but it may be a significant problem for throttling valves where precise positioning is desired. Diaphragm actuators do not exhibit the same degree of friction as piston actuators because the elastic diaphragm rolls and flexes rather than rubs against a stationary surface as is the case with piston sealing rings.

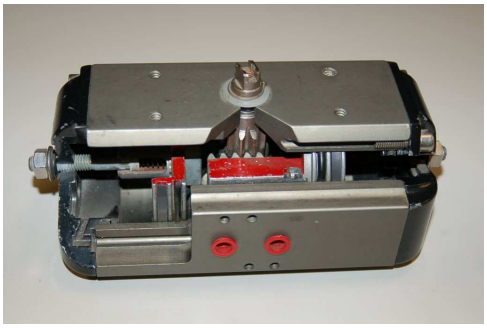
[9]The greater pressure rating of a piston actuator comes from the fact that the only "soft" component (the sealing ring) has far less surface area exposed to the high pressure than a rolling diaphragm. This results in significantly less stress on the elastic ring than there would be on an elastic diaphragm exposed to the same pressure. There really is no limit to the stroke length of a piston actuator as there is with the stroke length of a diaphragm actuator. It is possible to build a piston actuator miles long, but such a feat would be impossible for a diaphragm actuator, where the diaphragm must stretch (or roll) the entire stroke length.

A double-piston pneumatic actuator appears in the next photograph, providing the mechanical force needed to turn an on/off butterfly valve:

In this particular actuator design, a pair of pneumatically-actuated pistons move a rack-and-pinion mechanism to convert linear piston motion into rotary shaft motion to move the butterfly trim. Note the rotary indicator (yellow in color) at the end of the rotary valve stem, showing what position the butterfly valve is in. Note also the travel switch box (black in color) housing multiple limit switches providing remote indication of valve position to the control room.

Photographs of a cut-away rack-and-pinion piston actuator (same design, just smaller) show how the pistons' linear motion is converted into rotary motion to actuate a rotary valve:





Another pneumatic piston actuator design uses a simple crank lever instead of a rack-and-pinion gear set to convert linear piston motion into rotary motion. This next photograph shows such a piston actuator connected to a ball valve:

Hydraulic actuators

Hydraulic actuators use liquid pressure rather than gas pressure to move the valve mechanism. Nearly all hydraulic actuator designs use a piston rather than a diaphragm to convert fluid pressure into mechanical force. The high pressure rating of piston actuators lends itself well to typical hydraulic system pressures, and the lubricating nature of hydraulic oil helps to overcome the characteristic friction of piston-type actuators. Given the high pressure ratings of most hydraulic pistons, it is possible to generate tremendous actuating forces with a hydraulic actuator, even if the piston area is modest. For example, an hydraulic pressure of 2,000 PSI applied to one side of a 3 inch diameter piston will generate a linear thrust of over 14,000 pounds (7 tons)!

In addition to the ability of hydraulic actuators to easily generate extremely large forces, they also exhibit very *stable* positioning owing to the non-compressibility of hydraulic oil. Unlike pneumatic actuators, where the actuating fluid (air) is "elastic," the oil inside a hydraulic actuator cylinder does not yield appreciably under stress. If the passage of oil to and from a hydraulic cylinder is blocked by small valves, the actuator will become firmly "locked" into place. This may be a decided advantage for certain valve-positioning applications, where the actuator must resist forces generated

on the valve trim by process fluid pressures.

Some hydraulic actuators contain their own electrically-controlled pumps to provide the fluid power, so the valve is actually controlled by an electric signal. Other hydraulic actuators rely on a separate fluid power system (pump, reservoir, cooler, hand or solenoid valves, etc.) to provide hydraulic pressure on which to operate. Hydraulic pressure supply systems, however, tend to be more limited in physical span than pneumatic distribution systems due to the need for thick-walled tubing (to contain the high oil pressure), the need to purge the system of all gas bubbles, and the problem of maintaining a leak-free distribution network. It is usually not practical to build a hydraulic oil supply and distribution system large enough to cover the entirety of an industrial facility. Another disadvantage of hydraulic systems compared to pneumatic is lack of intrinsic power storage. Compressed air systems, by virtue of air's compressibility (elasticity), naturally store energy in any pressurized volumes, and so provide a certain degree of "reserve" power in the event that the main compressor shut down. Hydraulic systems do not naturally exhibit this desirable trait.

A hydraulic piston actuator attached to a large shut-off valve (used for on/off control rather than throttling) appears in the next photograph. Two hydraulic cylinders may be seen above the round valve body, mounted horizontally. Like the pneumatic piston valve shown earlier, this valve actuator uses a rack-and-pinion mechanism to convert the hydraulic pistons' linear motion into rotary motion to turn the valve trim:

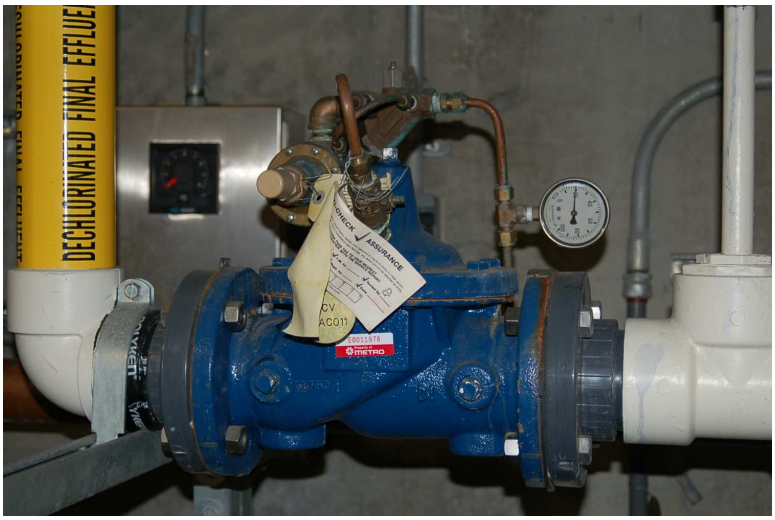
A feature not evident in this photograph is a hydraulic hand pump that may be used to manually actuate the valve in the event of hydraulic system failure.

Self-operated valves

Although not a type of actuator itself, a form of actuation worthy of mention is where the process fluid pressure itself actuates a valve mechanism. This self-operating principle may be used in throttling applications or on/off applications, in gas or liquid services alike. The process fluid may be directly tubed to the actuating element (diaphragm or piston), or passed through a small mechanism called a *pilot* to modulate that pressure before reaching the valve actuator. This latter design allows the main valve's motion to be controlled by an adjustable device (the pilot).

A very common application for pilot-operated control valves is gas pressure regulation, especially for fuel gas such as propane or natural gas used to fuel large industrial burners. This next photograph shows a Fisher gas pressure regulator used for regulating the pressure of natural gas fueling an industrial burner:

This next pilot-operated valve is used in a liquid (wastewater) service rather than gas:



A consumer-grade application for pilot-operated valves is irrigation system control, where the solenoid valves used to switch water flow on and off to sprinkler heads use pilot mechanisms rather than operate the valve mechanism directly with magnetic force. A small solenoid valve opens and closes to send water pressure to an actuating diaphragm, which then operates the larger valve mechanism to start and stop the flow of water to the sprinkler. The use of a pilot minimizes the amount of electrical power needed to actuate the solenoid.

A special case of self-operated valve is the *Pressure Relief Valve* (PRV) or *Pressure Safety Valve* (PSV). These valves are normally shut, opening only when sufficient fluid pressure develops across them to relieve that process fluid pressure and thereby protect the pipes and vessels upstream. Like the other self-operated valves, these safety valves may directly actuate using process fluid pressure or they may be triggered by a pilot mechanism sending process fluid pressure to the actuator only above certain pressures. Relief valves using pilots have the advantage of being widely adjustable, whereas non-pilot safety valves usually have limited adjustment ranges.

Relief valves typically have two pressure ratings: the pressure value required to initially open ("lift") the valve, and the pressure value required to re-seat the valve (the *blowdown* pressure). A relief valve's lift pressure will always exceed its blowdown pressure, giving the valve a hysteretic behavior.

This photograph shows a pressure relief valve on an industrial hot water system, designed to release pressure to atmosphere if necessary to prevent damage to process pipes and vessels in the system:

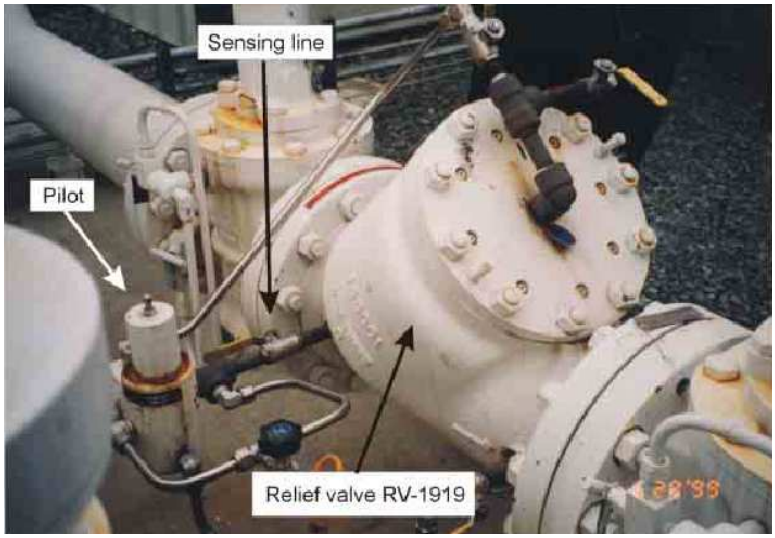
The vertical pipe is the atmospheric vent line, while the bottom flange of this PRV connects to the pressurized hot water line. A large spring inside the relief valve establishes the release pressure.

Another style of pressure relief valve appears in this next photograph. Manufactured by the Groth corporation, this is a combination pressure/vacuum relief valve assembly for an underground tank, designed to vent excess pressure to atmosphere *or* introduce air to the tank in the event of excess vacuum forming inside:

The purpose for a pressure/vacuum relief valve such as this is to relieve stresses applied to the walls of the tank due to difference of pressure inside and out. Large storage tanks are typically thin-wall for reasons of economics, and cannot withstand significant pressures or vacuums. An improperly vented storage tank may burst with only slight pressure inside, or collapse inwardly with only a slight vacuum inside. Combination pressure/vacuum relief valves such as this Groth unit reduce the chances of either failure from happening.

Of course, an alternative solution to this problem is to continuously vent the tank with an open vent pipe at the top. If the tank is always vented to atmosphere, it cannot build up either a pressure or a vacuum inside. However, continuous venting means vapors could escape from the tank if the liquid stored inside is volatile. Escaping vapors may constitute product loss and/or negative environmental impact, in which case it is prudent to vent the tank with an automatic valve such as this only when needed to prevent pressure-induced stress on the tank walls.

As previously mentioned, some pressure relief valves are pilot-controlled, their "lift" and "blowdown" pressures established by spring adjustments in the pilot mechanism rather than by adjustments made to the main valve mechanism. A photograph[10] of a pilot-operated pressure relief valve used on a liquid petroleum pipeline appears here:



[10]This photograph courtesy of the National Transportation Safety Board's report of the 1999 pipeline rupture in Bellingham, Washington. Improper setting of this relief valve pilot played a role in the pipeline rupture.

Electric actuators

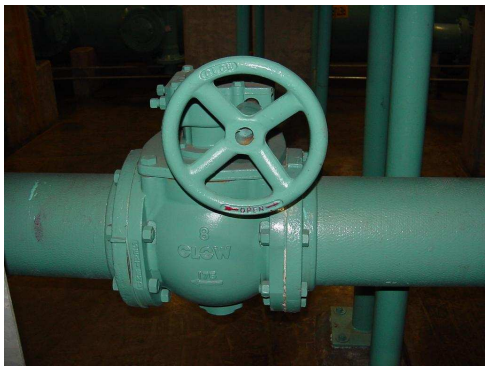
Electric motors have long been used to actuate large valves, either in on/off mode or in throttling services. Advances in motor design and motor control circuitry has brought motor-operated valve (MOV) technology to the point where it regularly competes with legacy actuator technologies such as pneumatic.

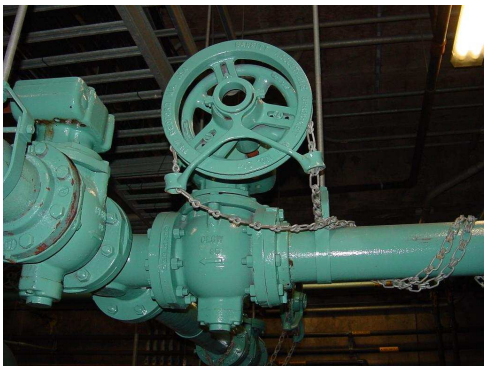
An electric actuator appears in the next photograph, providing rotary actuation to a ball valve. This particular electric actuator comes with a hand crank for manual operation, in the event that the electric motor (or the power provided to it) fails:

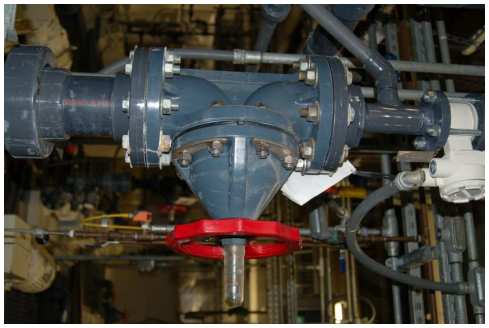
The next photograph shows a different brand of valve actuator (Rotork) turning a large butterfly valve. Although there nothing visible in this photograph betrays the nature of this actuator's signaling, it happens to be *digital* rather than analog, receiving position commands through a *Profibus* digital network rather than an analog 4-20 mA current signal:

Hand (manual) actuators

Valves may also be actuated by hand power alone. The following valves are all "manual" valves, requiring the intervention of a human operator to actuate:





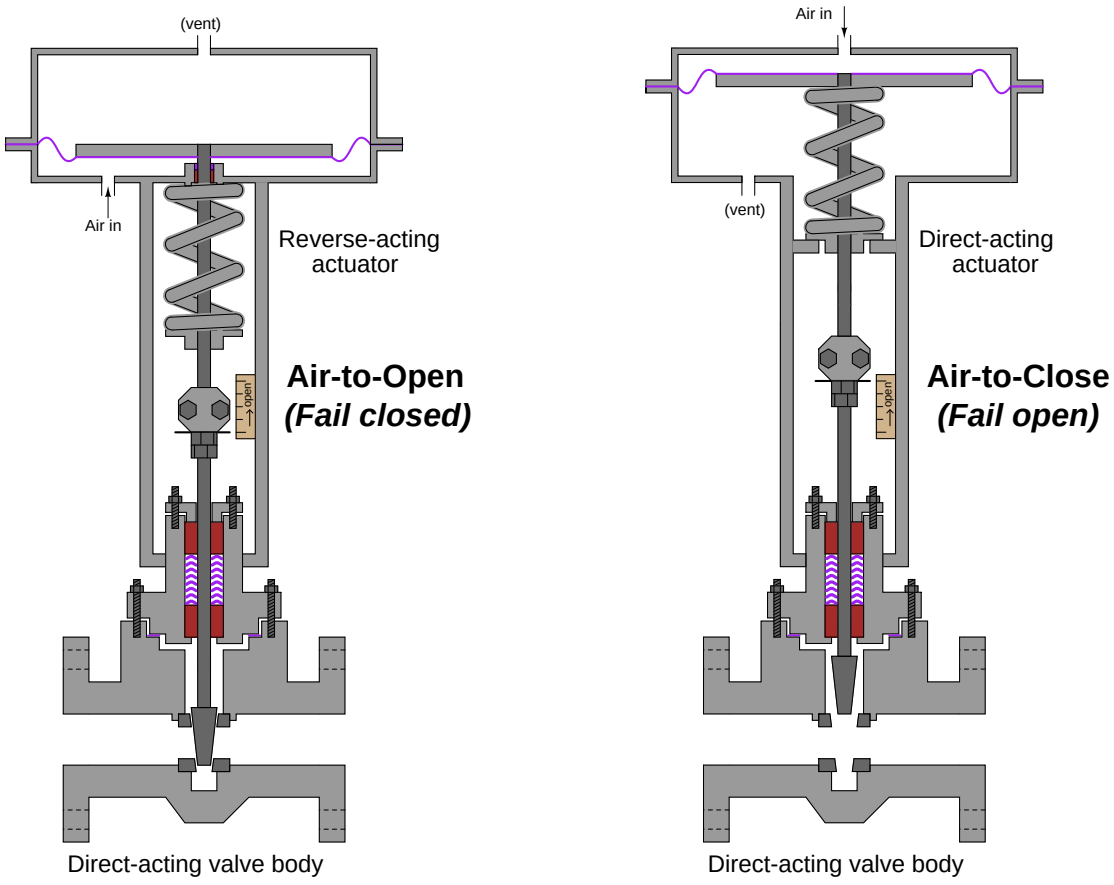


Note the threaded stem of the left-hand valve in the last photograph pair. This stem rises and falls with the handle's turning, providing visual indication of the valve's status. Such an actuator is called a *rising-stem* design.

25.1.7 Valve failure mode

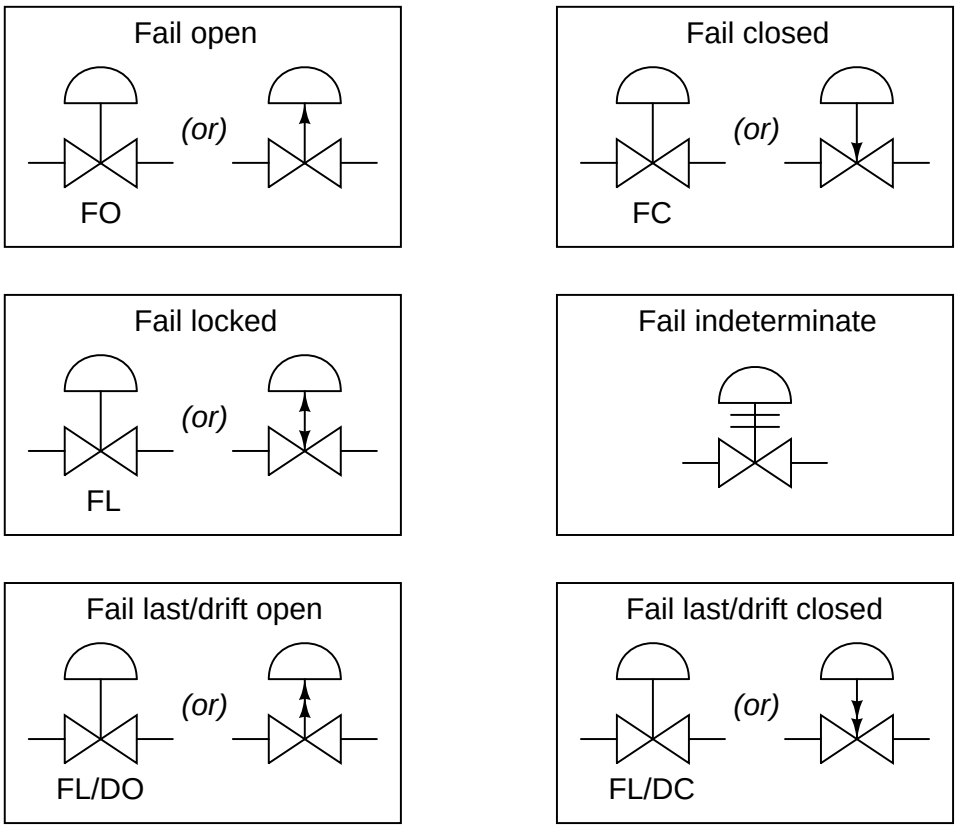
An important design parameter of a control valve is the position it will "fail" to if it loses motive power. For electrically actuated valves, this is typically the last position the valve was in before loss of electric power. For pneumatic and hydraulic actuated valves, the option exists of having a large spring provide a known "fail-safe" position (either open or closed) in the event of fluid pressure (pneumatic air pressure or hydraulic oil pressure) loss.

The fail-safe mode of a pneumatic/spring valve is a function of both the actuator's action and the valve body's action. For sliding-stem valves, a *direct-acting* actuator pushes down on the stem with increasing pressure while a *reverse-acting* actuator pulls up on the stem with increasing pressure. Sliding-stem valve bodies are classified as *direct-acting* if they open up when the stem is lifted, and classified as *reverse-acting* if they shut off (close) when the stem is lifted. Thus, a sliding-stem, pneumatically actuated control valve may be made *air-to-open* or *air-to-close* simply by matching the appropriate actuator and body types. The most common combinations mix a direct-acting valve body with either a reverse- or direct-acting valve actuator, as shown in this illustration:



Valve fail mode may be shown in instrument diagrams by either an arrow pointing in the direction

of failure (assuming a direct-acting valve body where stem motion toward the body closes and stem motion away from the body opens the valve trim) and/or the abbreviations "FC" (fail closed) and "FO" (fail open). Other failure modes are possible, as indicated by this set of valve symbols:



In order for a pneumatic or hydraulic valve to fail in the *locked* state, an external device must trap fluid pressure in the actuator's diaphragm or piston chamber in the event of supply pressure loss.

Valves that fail in place and drift in a particular direction are usually actuated by double-acting pneumatic piston actuators. These actuators do not use a spring to provide a definite fail mode, but rather use air pressure both to open and to close the valve. In the event of an air pressure loss, the actuator will neither be able to open nor close the valve, and so it will tend to remain in position. If the valve is of the globe design with unbalanced trim, forces exerted on the valve plug will move it in one direction (causing drift).

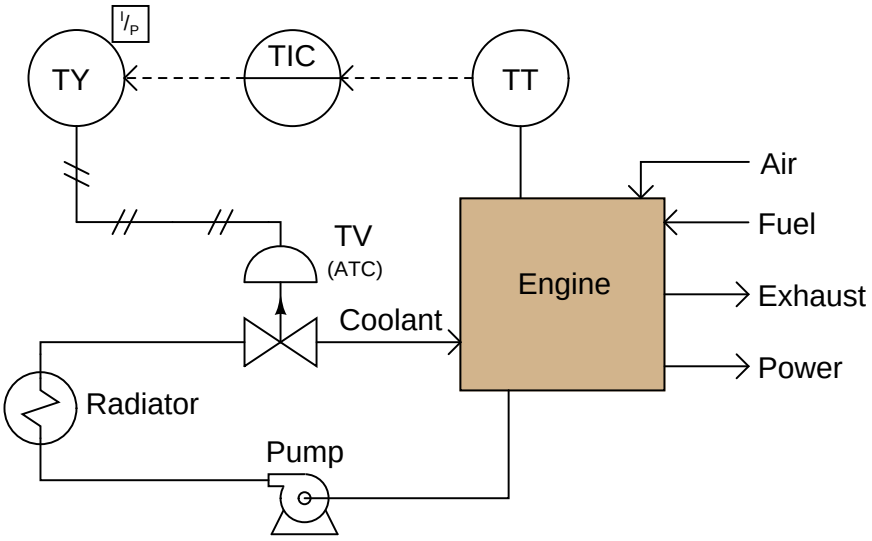
It is important to note how the failure mode of a valve is often linked to its control action (air-to-open, air-to-close)[11]. That is, an air-to-open pneumatic control valve will fail closed on loss of

[11] Exceptions exist for valves designed to fail in place, where a valve may be engineered to "lock" in position through the action of an external device whether the valve itself is air-to-open or air-to-close.

air pressure, and visa-versa. This is an important fact because good safety engineering demands that the risk factors of the process determine proper valve failure mode rather than control system convention or habit. People may find it easier to understand the operation of an air-to-open control valve than an air-to-close valve (more signal = more process fluid flow), but this should not be a guiding principle in valve selection. Air-to-open control valves fail closed by their very nature (unless equipped with automatic "locking" devices in which case they will fail in place), which means they are appropriate for a particular process control application *only* if that process is safer with a failed-closed valve than with a failed-open valve. If the process is safer with a fail-open valve, then the pneumatically-actuated control valve specified for that application needs to be air-to-close.

In fact, this basic principle forms the basis – or at least it *should* form the basis – of decisions made for all instrument actions in critical control loops: first determine the safest mode of valve failure, then select and/or configure instrument actions in such a way that the most probable modes of signal path failure will result in the control valve consistently moving to that (safest) position.

For example, consider this automated cooling system for a large power-generating engine:



Clearly, it is more hazardous to the engine for the valve to fail closed than it would be for the valve to fail open. If the valve fails closed, the engine will surely overheat from lack of cooling. If it fails open, the engine will merely run cooler than designed, the only negative consequence being decreased efficiency. With this in mind, the only sensible choice for a control valve is one that fails open (air-to-close).

However, our choices in instrument action do not end with the control valve. How should the temperature transmitter, the controller, and the I/P transducer be configured to act? In each case, the answer should be to act in such a way that the valve will default to its fail-safe position (wide open) in the event of the most likely input signal fault.

Stepping "backward" through the control system from the valve to the temperature sensor, the next instrument we encounter is the I/P transducer. Its job, of course, is to convert a 4-20 mA current signal into a corresponding pneumatic pressure that the valve actuator can use. Since we know that the valve's failure mode is based on a loss of actuating air pressure, we want the I/P to be

configured in such a way that it outputs minimum pressure in the event of an electrical fault in its 4-20 mA input signal wiring. The most common fault for a current loop is an *open*, where current goes to 0 milliamps. Thus, the configuration of the I/P transducer should be *direct*, such that a 4 to 20 mA input signal produces a 3 to 15 PSI output pressure, respectively (i.e. minimum input current yields minimum output pressure).

The next instrument in the loop is the controller. Here, we want the most likely input signal failure to result in a minimum output signal, so the valve will (once again) default to its "fail safe" position. Consequently, we should configure the controller for *direct* action just like we did with the I/P transducer (i.e. a decreasing PV signal from a broken wire or loose connection in the input circuit results in a decreasing output signal).

Finally, we come to the last instrument in the control loop: the temperature transmitter (TT). As with most instruments, we have the option of configuring it for direct or reverse action. Should we choose direct (i.e. hotter engine = more mA output) or reverse (hotter engine = less mA output)? Here, our choice needs to be made in such a way that the overall effect of the control system is *negative feedback*. In other words, we need to configure the transmitter such that a hotter engine results in increased coolant flow (a wider-open control valve). Since we know the rest of the system has been designed so a minimum signal anywhere tends to drive the valve to its fail-safe mode (wide open), we must choose a *reverse-acting* transmitter, so a hotter engine results in a decreased milliamp signal from the transmitter. If the transmitter has a sensor "burnout" mode switch, we should flip this switch into the low-scale burnout position, so a burned-out sensor will result in 4 mA output (low end of the 4-20 mA scale), thus driving the valve into its safest (wide-open) position.

Such a configuration – with its air-to-close control valve and a reverse-acting transmitter – may seem strange and counter-intuitive, but it is the safest design for this engine cooling system. We arrived at this "odd" configuration of instruments by first choosing the safest control valve failure mode, then choosing instrument actions in such a way that the most likely signal-path failures anywhere in the system would result in the same, consistent valve response. Of course it should go without saying that accurate documentation in the form of a loop diagram with instrument actions clearly shown is an absolutely essential piece of the whole system. If the safety of a control system depends on using any "non-standard" instrument configurations, those configurations had better be documented so those maintaining the system in the future will know what to expect!

25.1.8 Actuator bench-set

Valve actuators provide force to move control valve trim. For precise positioning of a control valve, there must be a calibrated relationship between applied force and valve position. Most pneumatic actuators exploit *Hooke's Law* to translate applied air pressure to valve stem position.

$$F = kx$$

Where,

F = Force applied to spring in newtons (metric) or pounds (British)

k = Constant of elasticity, or "spring constant" in newtons per meter (metric) or pounds per foot (British)

x = Displacement of spring in meters (metric) or feet (British)

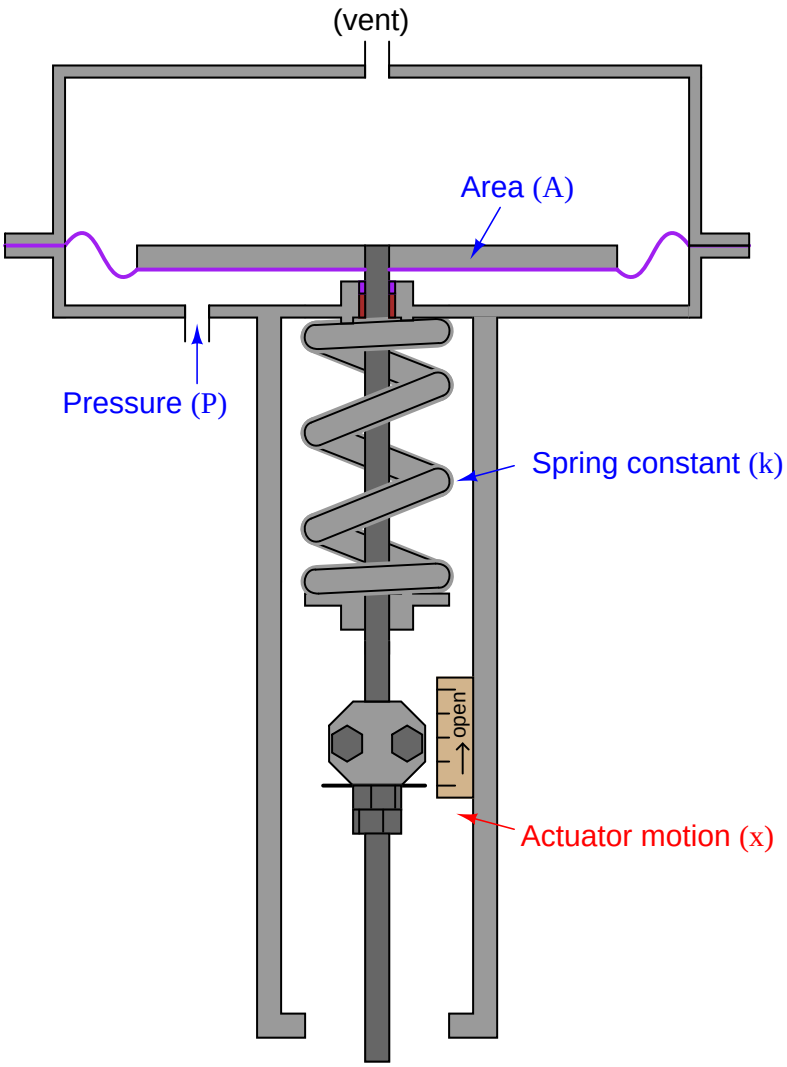
Hooke's Law is a linear function, which means that spring motion will be linearly related to applied force from the actuator element (piston or diaphragm). Since the working area of a piston or diaphragm is constant, the relationship between actuating fluid pressure and force will be a simple proportion ($F = PA$). By algebraic substitution, we may alter Hooke's Law to include pressure and area:

$$F = kx$$

$$PA = kx$$

Solving for spring compression as a function of pressure, area, and spring constant:

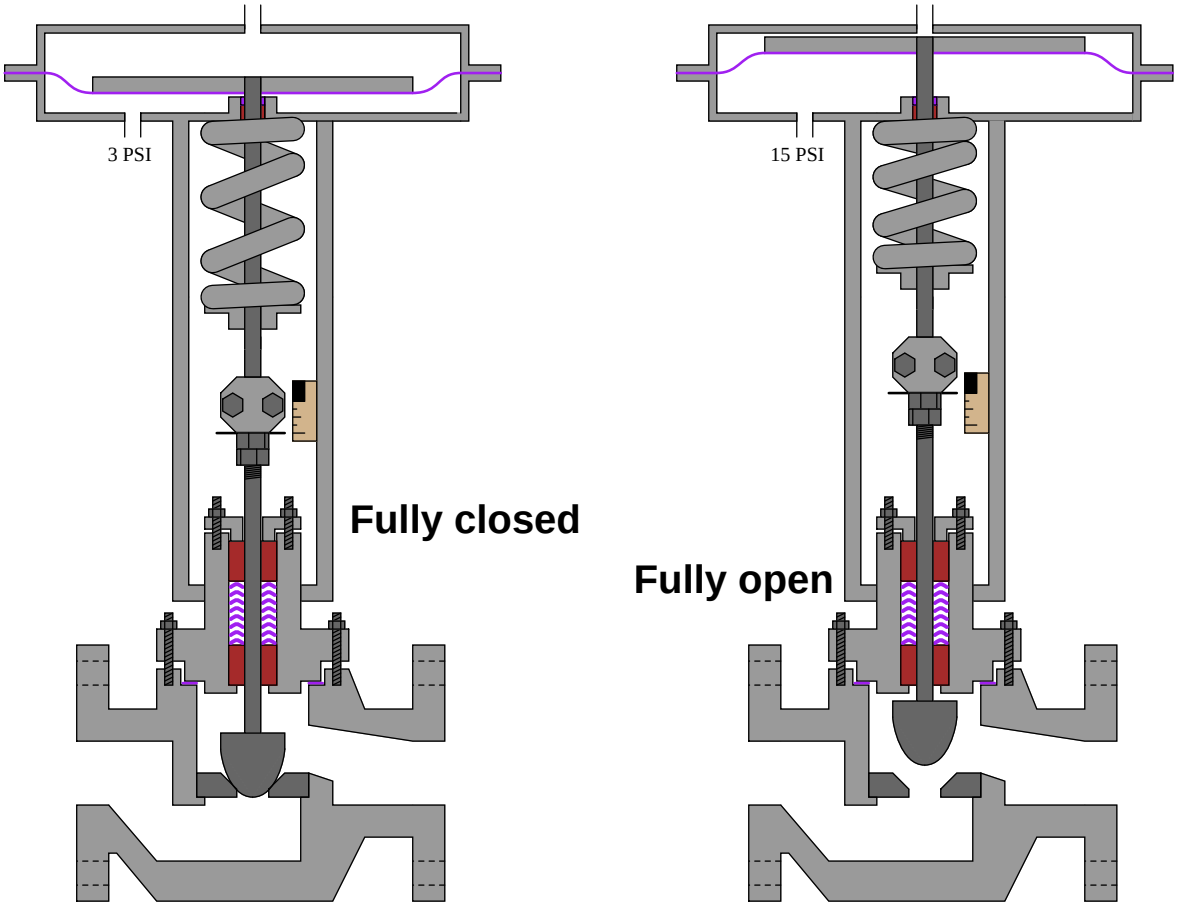
$$x = \frac{PA}{k}$$



When a control valve is assembled from an actuator and a valve body, the two mechanisms must be coupled together in such a way that the valve moves between its fully closed and fully open positions with an expected range of air pressures. A common standard for pneumatic control valve actuators is 3 to 15 PSI[12].

[12] 3 PSI could mean fully closed and 15 PSI fully open, or visa-versa, depending on what form of actuator is coupled to what form of valve body. A direct-acting actuator coupled to a direct-acting valve body will be open at low pressure and closed at high pressure (increasing pressure pushing the valve stem toward the body, closing off the valve trim),

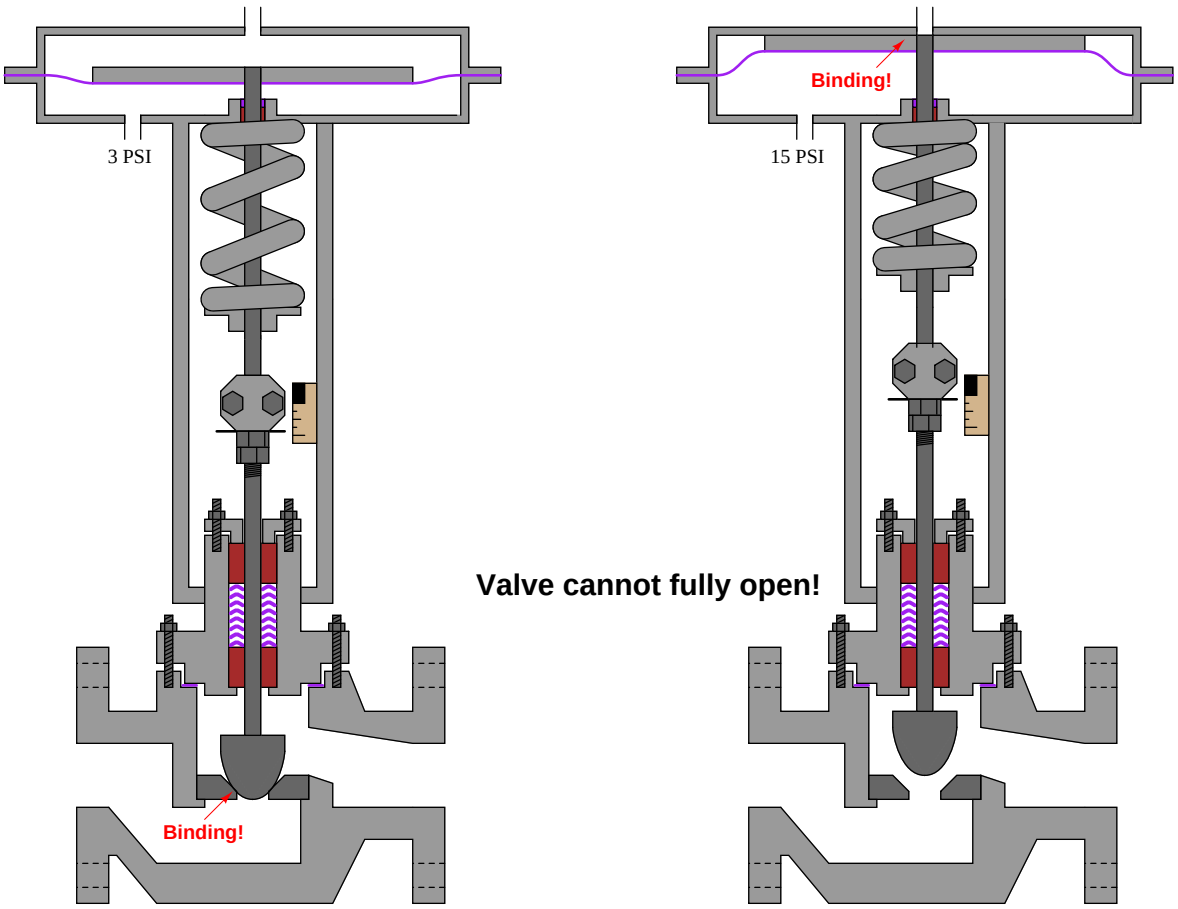
There are really only two mechanical adjustments that need to be made when coupling a pneumatic diaphragm actuator to a sliding-stem valve: the *stem connector* and the *spring adjuster*. The stem connector mechanically joins the sliding stems of both actuator and valve body so they move together as one stem. This connector must be adjusted so neither the actuator nor the valve trim prevents full travel of the valve trim:



Note how the plug is fully against the seat when the valve is closed, and how the travel indicator indicates fully open at the point where the actuator diaphragm nears its fully upward travel limit. This is how things should be when the stem connector is properly adjusted.

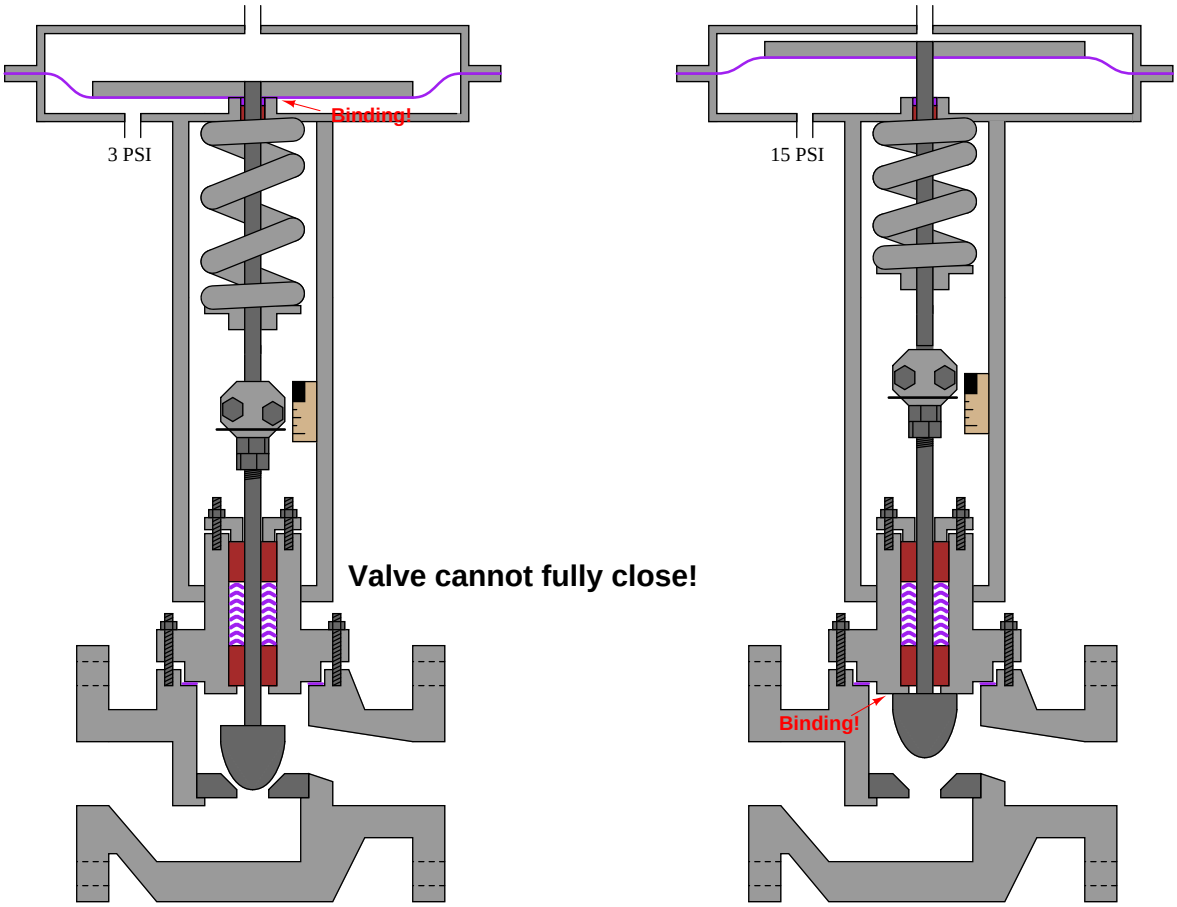
resulting in *air-to-close* action. Reversing either actuator or valve type (e.g. reverse actuator with direct valve or direct actuator with reverse valve) will result in *air-to-open* action.

If the stem connector is set with the actuator and valve stems spaced too far apart (i.e. the total stem length is too long), the actuator diaphragm will bind travel at the upper end and the valve plug will bind travel at the lower end. The result is a valve that cannot ever fully open:



A control valve improperly adjusted in this manner will never achieve full-flow capacity, which may have an adverse impact on control system performance.

If the stem connector is set with the actuator and valve stems too closely coupled (i.e. the total stem length is too short), the actuator diaphragm will bind travel at the lower end and the valve plug will bind travel at the upper end. The result is a valve that cannot ever fully close:



This is a very dangerous condition: a control valve that lacks the ability to fully shut off. The process in which this valve is installed may be placed in jeopardy if the valve lacks the ability to stop the flow of fluid through it!

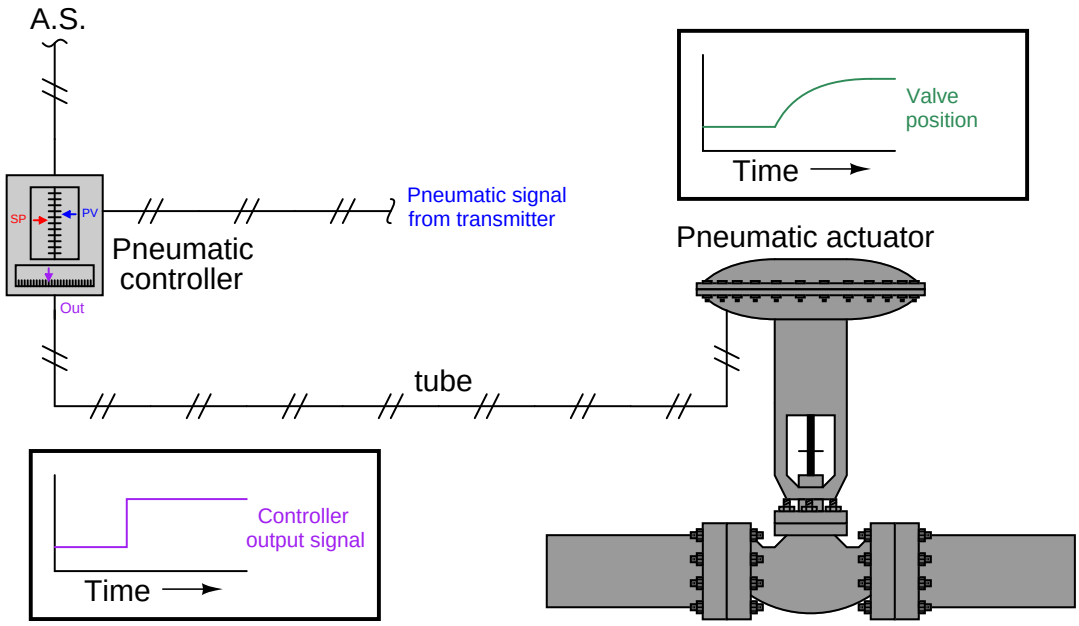
Once the stem length has been properly set by adjusting the stem connector, the spring adjuster must be set for the proper *bench set* pressure. This is the pneumatic signal pressure required to lift the plug off the seat. For an air-to-open control valve with a 3 to 15 PSI signal range, the "bench set" pressure would be 3 PSI.

Bench set is a very important parameter for a control valve because it establishes the seating pressure of the plug when the valve is fully closed. Proper seating pressure is critical for tight shut-off, which carries safety implications in some process services. Consult the manufacturer's instructions when adjusting the bench set pressure for any sliding-stem control valve. These instructions will

typically guide you through both the stem connector and the spring adjuster procedures, to ensure both parameters are correctly set.

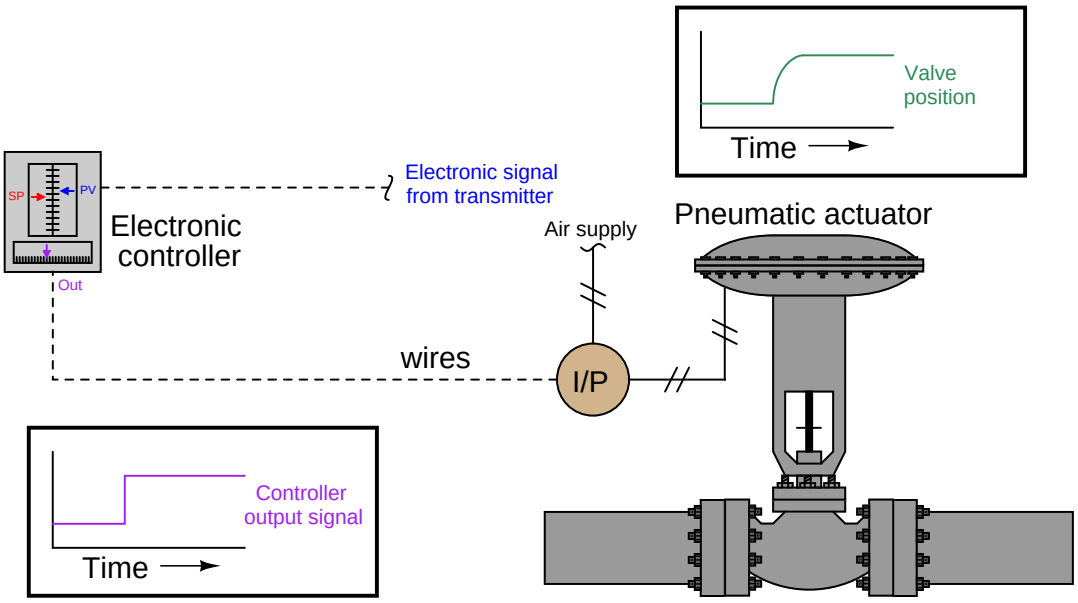
25.1.9 Pneumatic actuator response

A limitation inherent to pneumatic valve actuators is the amount of air flow required to or from the actuator to cause rapid valve motion. This is an especially acute problem in all-pneumatic control systems, where the distance separating a control valve from the controller may be substantial:



The combined effect of air-flow friction in the tube, flow limitations inherent to the controller mechanism, and volume inside the valve actuator conspire to create a sluggish valve response to sudden changes in controller output signal, not unlike the response of an RC (resistor-capacitor) time-delay circuit where a step-change in voltage input results in an inverse exponential output signal.

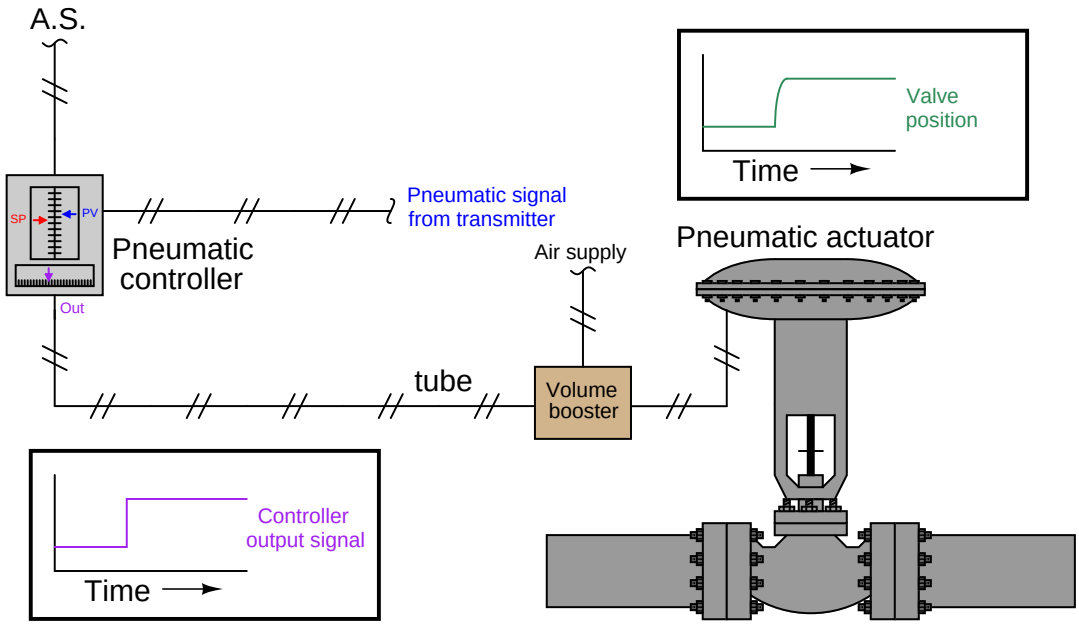
If the pneumatic valve actuator is driven by an I/P transducer instead of directly by a pneumatic controller, the problem is lessened by the ability to locate the I/P close to the actuator, thus greatly minimizing tube friction and thus minimizing the "time constant" (τ) of the control valve's response:



Still, if the pneumatic actuator is particularly large in volume, an I/P transducer may experience trouble supplying the necessary air flow rate to rapidly actuate the control valve. Certainly the problem of time delay is reduced, but not eliminated, by the close-coupled location of the I/P transducer to the actuator.

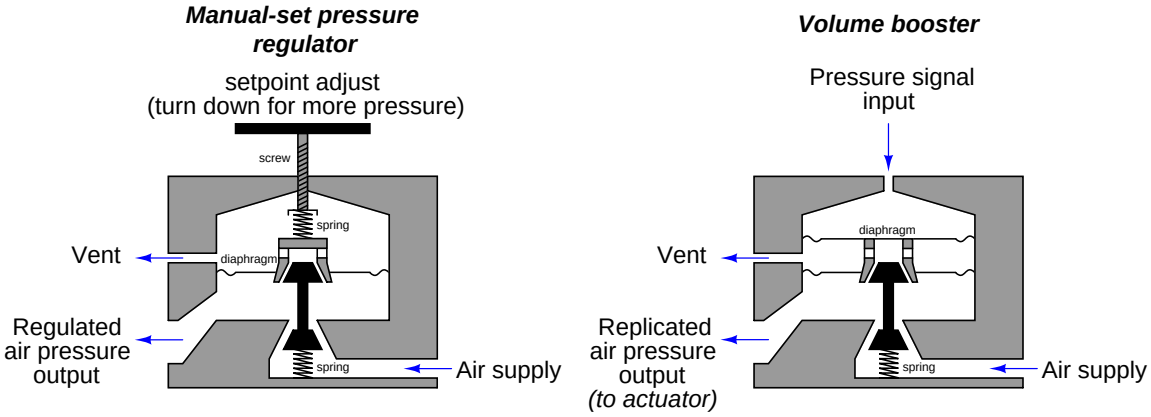
One way to improve valve response in either type of system (full-pneumatic or I/P-driven) is to use a device known as a *volume booster* to source and vent compressed air for the valve actuator. A "volume booster" is a pneumatic device designed to reproduce a pneumatic pressure signal (1:1 ratio), but with far greater output flow capacity. In electrical terms, a volume booster is analogous to a *voltage follower*: a circuit designed to boost current to a load, without boosting or diminishing voltage. A 3 to 15 PSI pneumatic pressure signal applied to the input of a volume booster will result in an identical output signal (3 to 15 PSI), but with greatly enhanced flow capacity.

A pneumatic control system equipped with a volume booster would look something like this:



Of course, enhanced air flow to and from the actuator does not completely eliminate time delays in valve response. So long as the flow rate into or out of an actuator is finite, some time will be required to change pressure inside the actuator and thus change valve position. However, if the actuator volume cannot be reduced for practical reasons of actuating force (larger diaphragm or piston area needed for more force, also resulting in more volume for any given stroke length), then the only variable capable of reducing time lag is increased air flow rate, and a volume booster directly addresses that deficiency.

Internally, a volume booster's construction is not unlike a manually-adjusted pressure regulator[13]:



In either mechanism, an internal diaphragm senses output pressure and acts against a restraining force (either a spring preloaded by a hand adjustment screw or an external pressure signal acting on another diaphragm) to position an air flow throttling/venting mechanism. If the output pressure is less than desired, the diaphragm moves down to open the air sourcing plug and supply additional air volume to the output. If the output pressure is greater than desired, the diaphragm moves up to shut off the sourcing plug and open up the venting port to relieve air pressure to atmosphere.

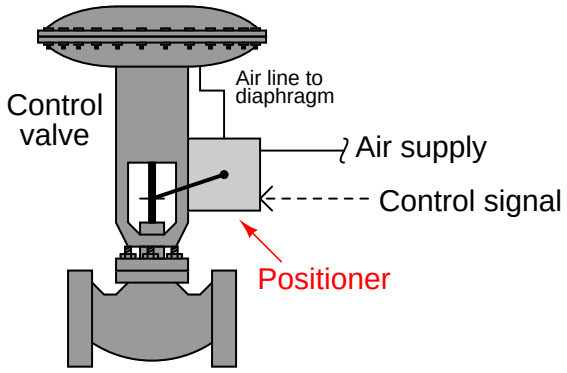
[13]The volume booster design shown here is loosely based on the Fisher model 2625 volume boosting relay.

25.1.10 Valve positioners

Springs work quite nicely to convert mechanical force into mechanical motion (Hooke's Law – $F = kx$) for valve actuators if and only if the sole forces involved are the diaphragm or piston force against the spring's resistance force. If any other force acts upon the system, the relationship between actuating fluid pressure and valve stem travel will not necessarily be proportional.

Unfortunately, there typically are other forces acting on a valve stem besides the actuating fluid pressure's force and the spring's reaction force. Friction from the stem packing is one force, and reaction force at the valve plug caused by differential pressure across the plug's area is another[14]. These forces conspire to re-position the valve stem so stem travel does not precisely correlate to actuating fluid pressure.

A common solution to this dilemma is to add a *positioner* to the control valve assembly. A positioner is a motion-control device designed to actively compare stem position against the control signal, adjusting pressure to the actuator diaphragm or piston until the correct stem position is reached:



Positioners essentially act as control systems within themselves[15]: the valve's stem position is the process variable (PV), the command signal to the positioner is the setpoint (SP), and the positioner's signal to the valve actuator is the manipulated variable (MV) or output. Thus, when a process controller sends a command signal to a valve equipped with a positioner, the positioner receives that command signal and does its best to ensure the valve stem position follows along.

[14] One way to minimize dynamic forces on a globe valve plug is to use a *double-ported* plug design, or to use a *balanced* plug on a cage-guided globe valve. A disadvantage to both these valve plug designs, though, is greater difficulty achieving tight shut-off.

[15] The technical term for this type of control system is *cascade*, where one controller's output becomes the setpoint for a different controller. In the case of a valve positioner, the positioner receives a valve stem position setpoint from the main process controller.

The following photograph shows a Fisher model 3582 pneumatic positioner mounted to a control valve. The positioner is the grey-colored box with three pressure gauges on the right-hand side:

A more modern positioner appears in the next photograph, the Fisher DVC6000 (again, the grey-colored box with pressure gauges on the right-hand side):

Positioners such as the DVC6000 are considered "smart" devices, containing digital electronic microprocessors to monitor and control valve stem position in accordance with the control signal, and also store data relevant to diagnostics[16].

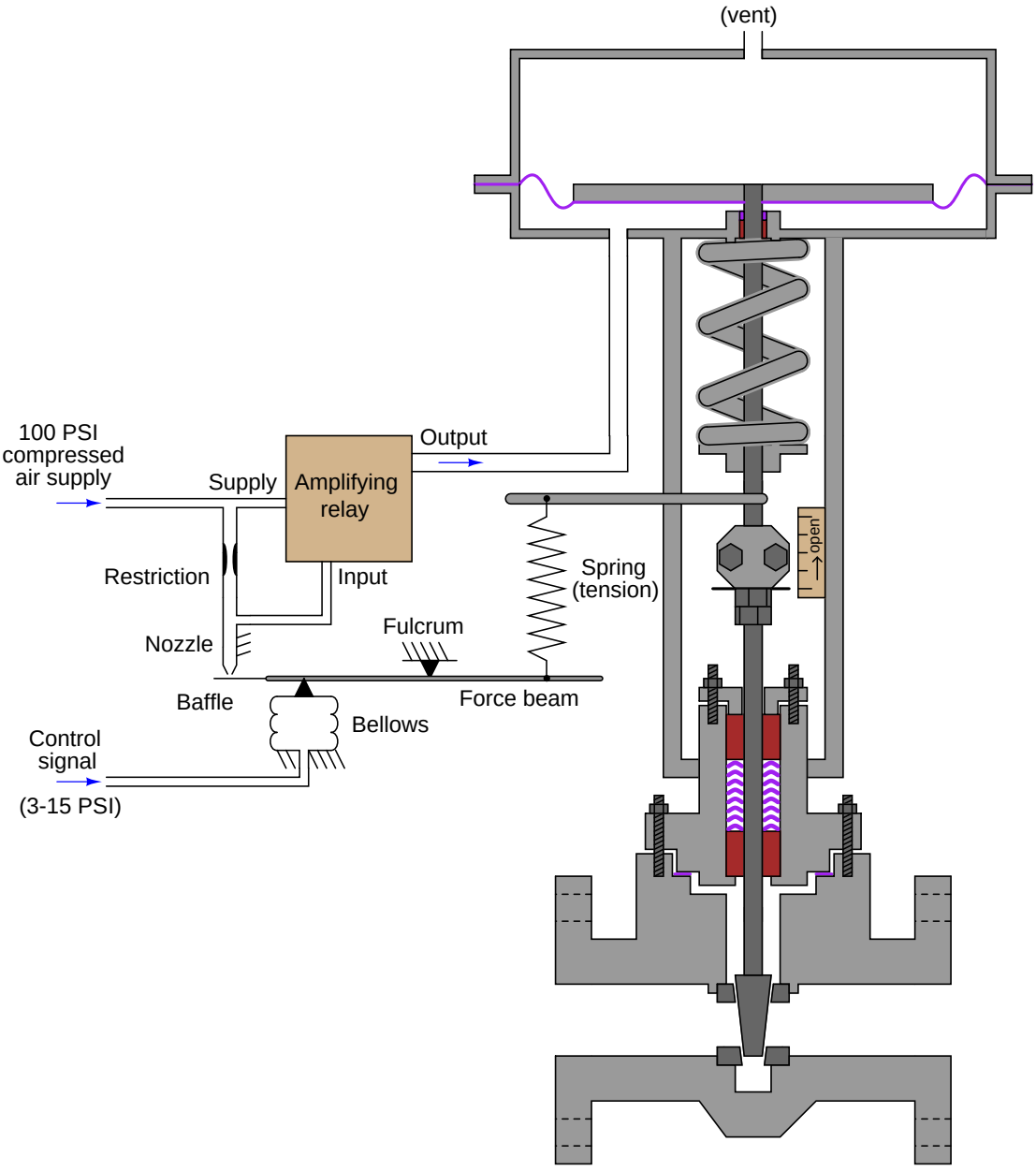
Control valve positioners are typically constructed in such a way to source and vent high air flow rates, such that the positioner also fulfills the functionality of a volume booster. Thus, a positioner not only ensures more precise valve stem positioning, but also faster stem velocity (and shorter time delays) than if the valve actuator were directly "powered" by an I/P transducer.

While beneficial on spring-equipped valve actuators, positioners are absolutely essential for positioners lacking springs such as double-acting pneumatic piston actuators. Some form of positioning mechanism is also required on electric motor actuators intended for throttling service, because an electric motor is not "aware" of its own shaft position in order that it may precisely

[16] Examples of diagnostic data recorded by smart positioners includes error (command signal − actual valve position), pressure versus motion relationships (used to measure valve packing friction), supply air pressure, and ambient temperature. Smart positioners even have the ability to totalize valve stem travel over long periods of time, enabling predictive maintenance alerts for wearing components such as packing and piston sealing rings.

move a control valve. Thus, a positioner circuit using a potentiometer or LVDT/RVDT sensor to detect valve stem position and a set of transistor outputs to drive the motor is necessary to make an electric actuator responsive to an analog control signal.

A simple force-balance pneumatic valve positioner design appears in the following illustration:

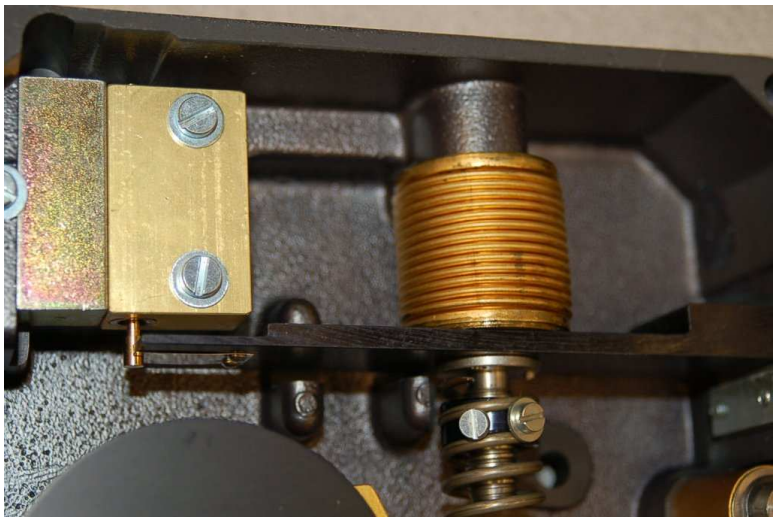


The control signal for this valve is a 3 to 15 PSI pneumatic signal, coming from either an I/P transducer or a pneumatic controller (neither one shown in the illustration). This control signal pressure applies an upward force on the force beam, such that the baffle tries to approach the nozzle. Increasing backpressure in the nozzle causes the pneumatic amplifying relay to output a greater air pressure to the valve actuator, which in turn lifts the valve stem up (opening up the valve). As the valve stem lifts up, the spring connecting the force beam to the valve stem becomes further stretched, applying additional force to the right-hand side of the force beam. When this additional force balances the bellows' force, the system stabilizes at a new equilibrium.

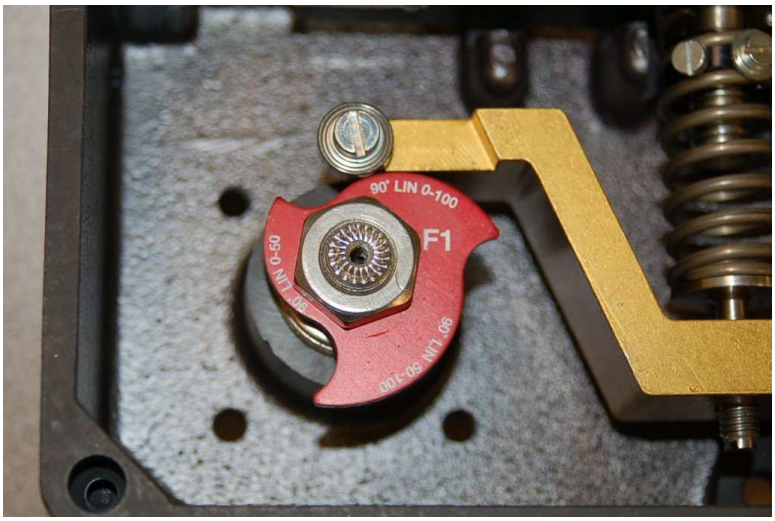
Like all force-balance systems, the force beam motion is greatly constrained by the balancing forces, such that its motion is negligible for all practical purposes. In the end, equilibrium is achieved by one force balancing another, like two teams of people pulling oppositely on a length of rope: so long as the two teams' forces remain equal in magnitude and opposite in direction, the rope will not deviate from its original position.

The following photograph shows a PMV model 1500 force-balance positioner used to position a rotary valve actuator, with the cover on (left) and removed (right):

The 3-15 PSI pneumatic control signal enters into the bellows, which pushes downward on the horizontal force beam. A pneumatic pilot valve assembly at the left-hand side of the force beam detects any motion, increasing air pressure to the valve actuating diaphragm if any downward motion is detected and releasing air pressure from the actuator if any upward motion is detected:

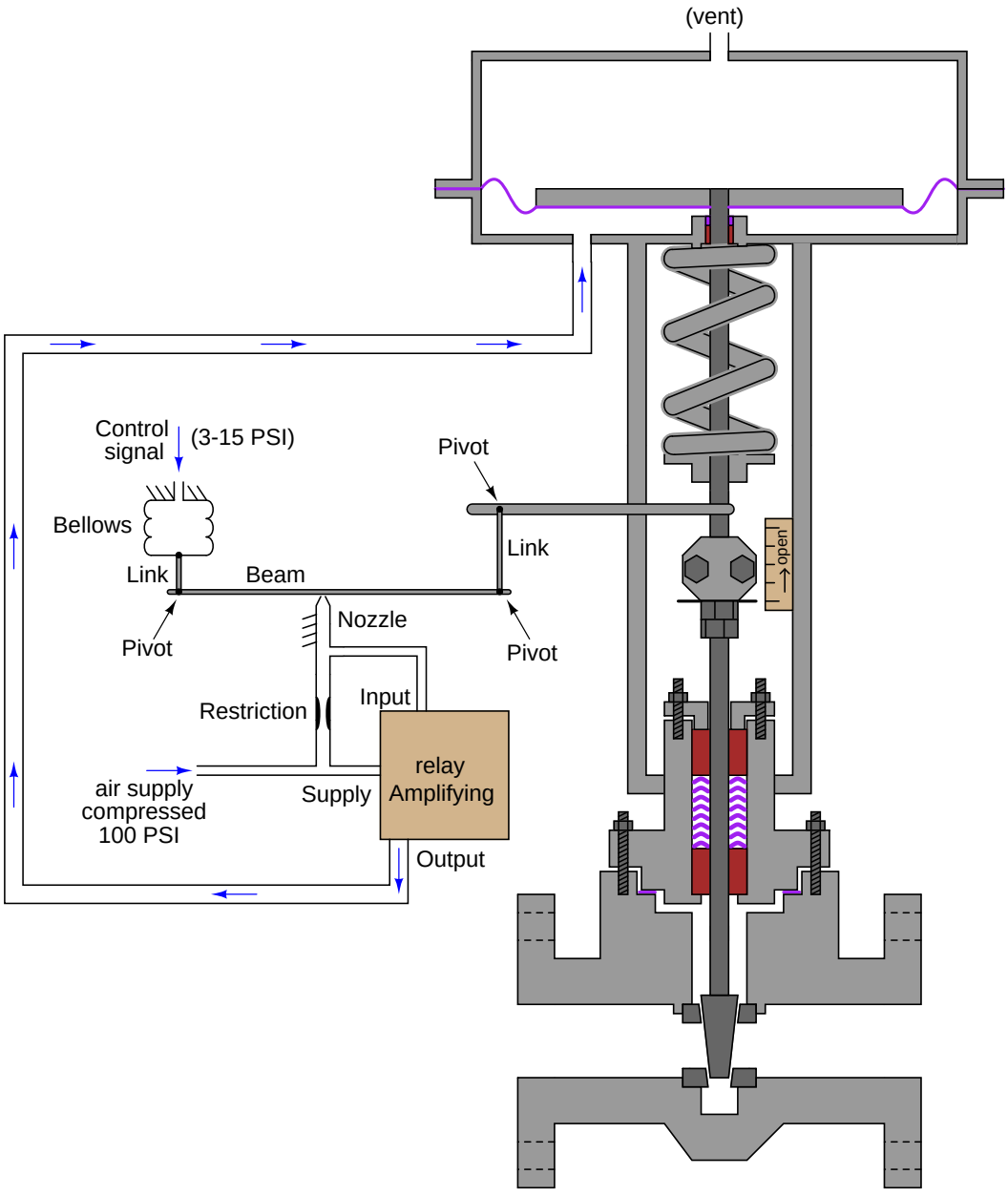


As compressed air is admitted to the valve actuator by this pilot valve assembly, the rotary valve will begin to rotate open. The shaft's rotary motion is converted into a linear motion inside the positioner by means of a *cam*: a disk with an irregular radius designed to produce linear displacement from angular displacement:



A roller-tipped *follower* at the end of another beam rides along the cam's circumference. Beam motion caused by the cam is translated into linear force by the compression of a coil spring directly against the force of the pneumatic bellows on the force beam. When the cam moves far enough to compress the spring enough to balance the additional force generated by the bellows, the force beam will come to an equilibrium position and the valve will stop moving.

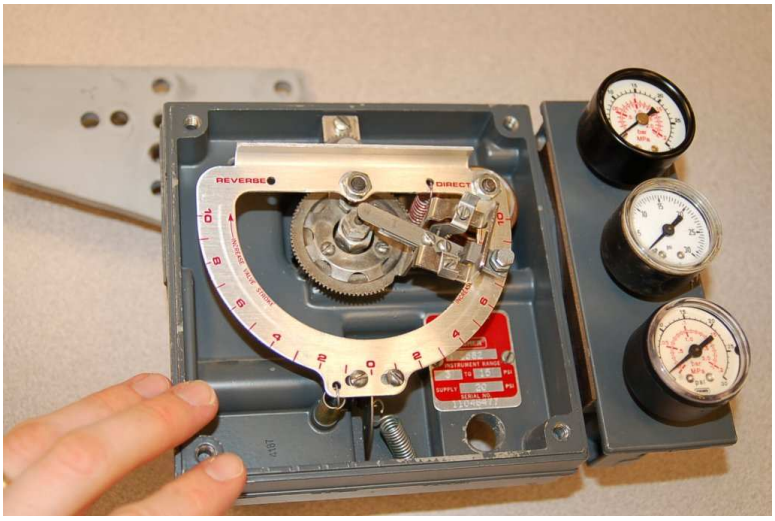
Motion-balance pneumatic valve positioner designs also exist, whereby the motion of the valve stem counteracts motion (not force) from another element. The following illustration shows how a simple motion-balance positioner would work:



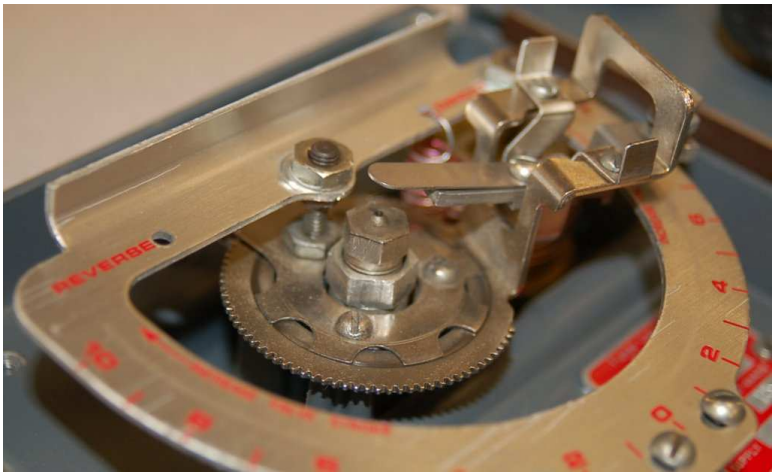
In this mechanism, an increasing signal pressure causes the beam to advance toward the nozzle,

generating increased nozzle backpressure which then causes the pneumatic amplifying relay to send more air pressure to the valve actuator. As the valve stem lifts up, the upward motion imparted to the right-hand end of the beam counters the beam's previous advance toward the nozzle. When equilibrium is reached, the beam will be in an angled position with the bellows' motion balanced by valve stem motion.

The following photograph shows a close view of a Fisher model 3582 pneumatic motion-balance positioner's mechanism:



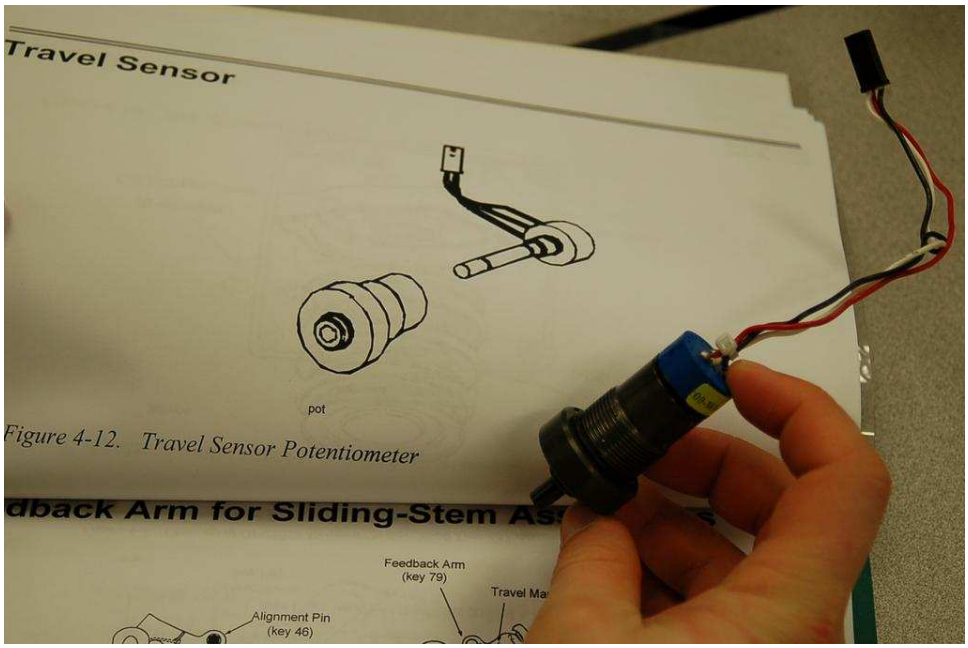
At the heart of this mechanism is a D-shaped metal ring translating bellows motion and valve stem motion into flapper (baffle) motion. As the bellows (located underneath the upper-right corner of the D-ring) expands with increasing pneumatic signal pressure, it rocks the beam along its vertical axis. With the positioner set for direct-acting operation, this rocking motion drives the flapper closer to the nozzle, increasing backpressure and sending more compressed air to the valve actuator:



As the valve stem moves, a feedback lever rotates a cam underneath the bottom-most portion of the D-ring. A roller follower riding on that cam translates the valve stem's motion to another rocking motion on the beam, this time along the horizontal axis. Depending on how the cam has been fixed to the feedback shaft, this motion may rock the flapper away from the nozzle or further toward the nozzle. This selection of cam orientation must match the action of the actuator: either direct (air to extend the stem) or reverse (air to retract the stem).

The D-ring mechanism is rather ingenious, as it allows convenient adjustment of span by angling the flapper (baffle) assembly at different points along the ring's circumference. If the flapper assembly is set close to horizontal, it will be maximally sensitive to bellows motion and minimally sensitive to valve stem motion, forcing the valve to move further to balance small motions of the bellows (long stroke length). Conversely, if the flapper assembly is set close to vertical, it will be maximally sensitive to valve stem motion and minimally sensitive to bellows motion, resulting in little valve stroke (i.e. the bellows needs to expand greatly in order to balance a small amount of stem motion).

Electronic valve positioners, such as the Fisher model DVC6000, use an electronic sensor to detect valve stem position, compare that sensed position against the control signal by subtraction (error = position − signal), then send an appropriate pneumatic pressure to the valve actuator to minimize that error. A photograph of the potentiometer from a Fisher DVC6000 positioner appears here:



The DVC6000 positioner also contains air pressure sensors to monitor actuator air pressure as the valve moves. Being able to measure both stem position and actuator air pressure in real time allows the positioner to correlate one variable to the other in the form of a graph. Such a graph contains much useful diagnostic information for troubleshooting valve problems such as excessive packing friction, bent valve stems, and valve trim damage.

"Smart" positioners used on electric actuators have the capability to provide similar diagnostic data, correlating stem position with actuator torque (measured either indirectly by motor current or directly by a torque sensor in the gear train). Such data is quite valuable in predictive maintenance programs, used to identify when valve packing friction becomes excessive, or if valve trim components become damaged and no longer seat together properly. These diagnostic tools apply even to open/close motor-operated valves not used for throttling service, and are especially useful on gate, plug, and ball-type shut-off valves where seat engagement is substantial for tight shut-off.

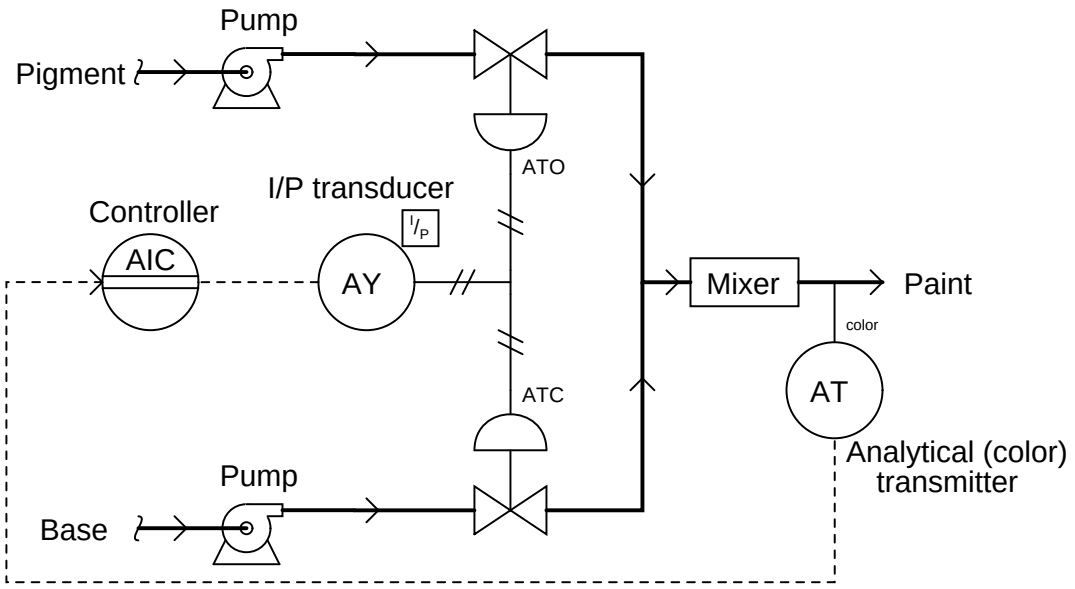
25.1.11 Split-ranging

There are many process control applications in industry where it is desirable to have multiple control valves respond to the output of a common controller. Control valves configured to follow the command of the same controller are said to be *split-ranged*, or *sequenced*.

Split-ranged control valves may take different forms of sequencing. A few different modes of control valve sequencing are commonly seen in industry: *complementary*, *exclusive*, and *progressive*[17].

Complementary valve sequencing

The first is a mode where two valves serve to proportion a mixture of two fluid streams, such as this example where base and pigment liquids are mixed together to form colored paint:

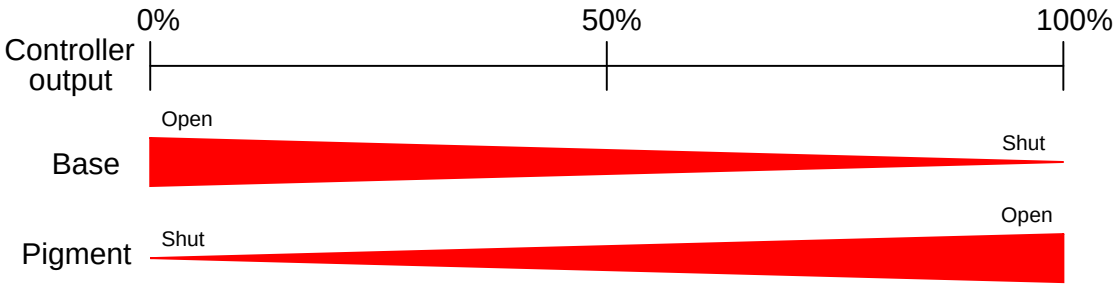


[17] I have searched in vain for standardized names to categorize different forms of control valve sequencing. The names "complementary," "exclusive," and "progressive" are my own invention. If I have missed someone else's categorization of split-ranging in my research, I sincerely apologize.

Both base and pigment valves operate from the same 3 to 15 PSI pneumatic signal output by the I/P transducer (AY), but one of the valves is Air-To-Open while the other is Air-To-Close. The following table shows the relationship between valve opening for each control valve and the controller's output:

Controller output (%)	I/P output (PSI)	Pigment valve (stem position)	Base valve (stem position)
0 %	3 PSI	fully shut	fully open
25 %	6 PSI	25% open	75% open
50 %	9 PSI	half-open	half-open
75 %	12 PSI	75% open	25% open
100 %	15 PSI	fully open	fully shut

An alternative expression for this split-range valve behavior is a graph showing each valve opening as a colored stripe of varying width (wider representing further open). For this particular mode of split-ranging, the graph would look like this:



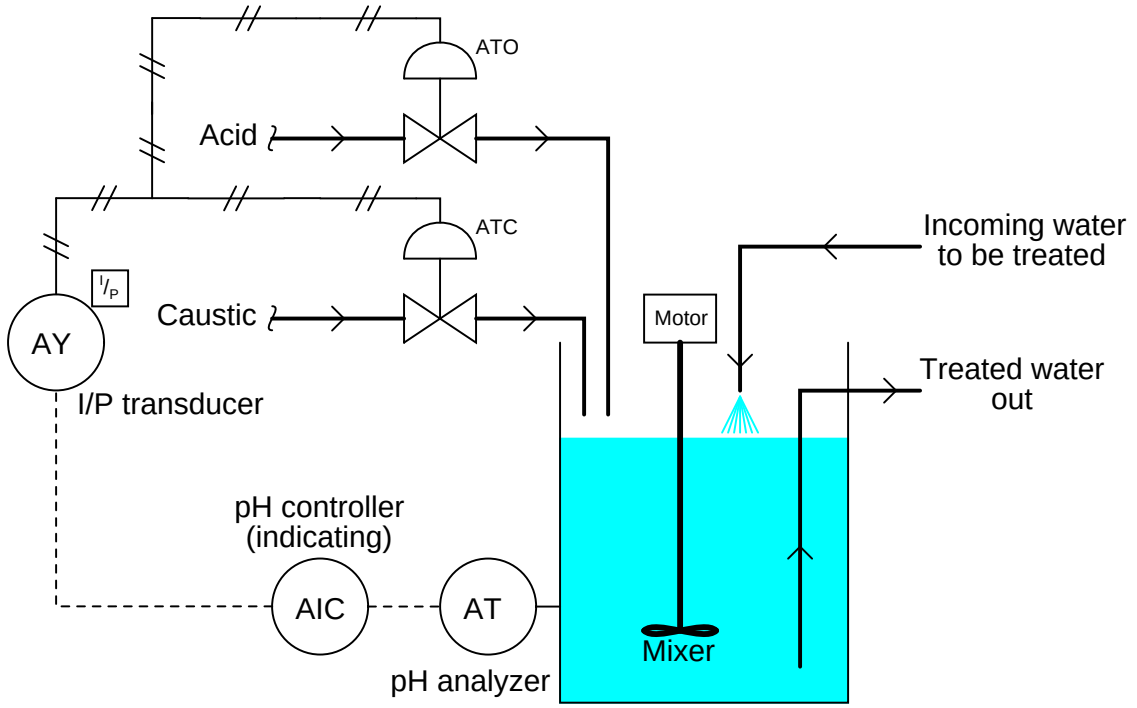
With this form of split-ranging, there is never a condition in the controller's output range where both valves are fully open or fully shut. Rather, each valve *complements* the other's position[18].

[18] In mathematics, a "complement" is a value whose sum with another quantity always results in a fixed total. Complementary angles, for instance, always add to 90^o (a right angle).

Exclusive valve sequencing

Other applications for split-ranged control valves call for a form of valve sequencing where both valves are fully closed at a 50% controller output signal, with one valve opening fully as the controller output drives toward 100% and the other valve opening fully as the controller output goes to 0%. The nature of this valve sequencing is to have an "either-or" throttled path for process fluid. That is, *either* process fluid flows through one valve *or* through the other, but never through both at the same time.

A practical example of this form of split-ranging is in reagent feed to a pH neutralization process, where the pH value of process liquid is brought closer to neutral by the addition of either acid or caustic:



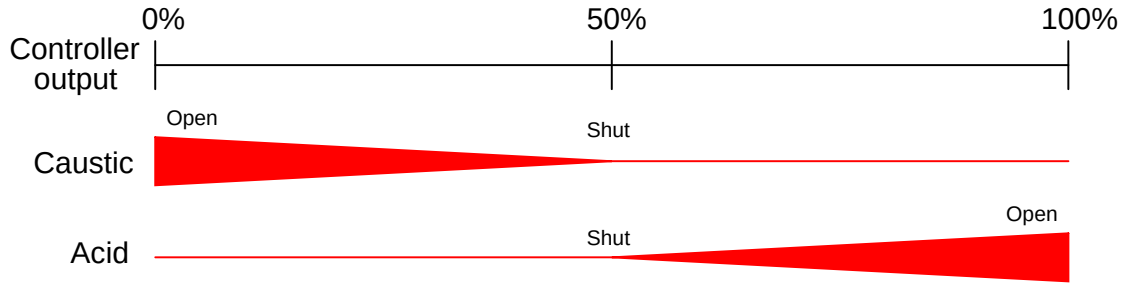
Here, a pH analyzer monitors the pH value of the mixture and a single pH controller commands two reagent valves to open when needed. If the process pH begins to increase, the controller output signal increases as well (direct action) to open up the acid valve. The addition of acid to the mixture will have the effect of lowering the mixture's pH value. Conversely, if the process pH begins to decrease, the controller output signal will decrease as well, closing the acid valve and opening the caustic valve. The addition of caustic to the mixture will have the effect of raising the mixture's pH value.

Both reagent control valves operate from the same 3 to 15 PSI pneumatic signal output by the I/P transducer (AY), but the two valves' calibrated ranges are not the same. The Air-To-Open acid valve has an operating range of 9 to 15 PSI, while the Air-To-Close caustic valve has an operating

range of 9 to 3 PSI. The following table shows the relationship between valve opening for each control valve and the controller's output:

Controller output (%)	I/P output (PSI)	Acid valve (stem position)	Caustic valve (stem position)
0 %	3 PSI	fully shut	fully open
25 %	6 PSI	fully shut	half-open
50 %	9 PSI	fully shut	fully shut
75 %	12 PSI	half-open	fully shut
100 %	15 PSI	fully open	fully shut

Again, we may express the two valves' exclusive relationship in the form of a graph, with colored stripes representing valve opening:



Exclusive-sequenced control valves are used in applications where it would be undesirable to have both valves open simultaneously. In the example given of a pH neutralization process, the goal here is for the controller to be able to call forth either acid reagent or caustic reagent to "push" the pH value either direction as needed. However, simultaneously adding both acid and caustic to the process would be wasteful, as one reagent would simply neutralize the other with no benefit to the process liquid itself.

Progressive valve sequencing

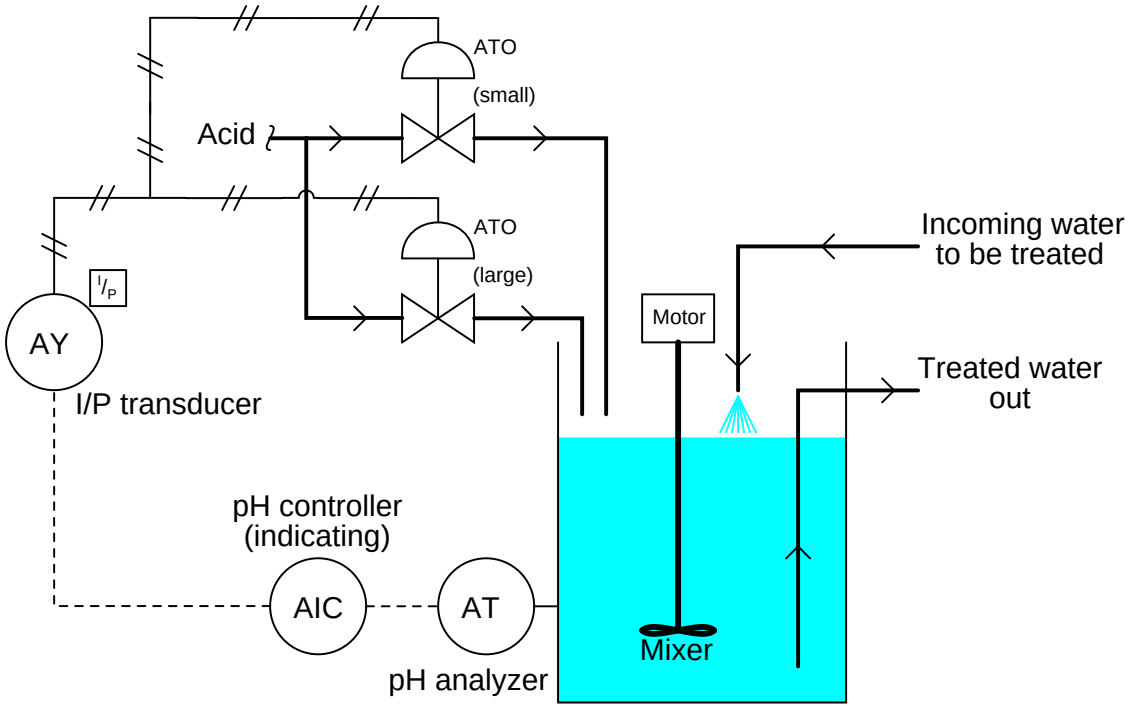
A third form of control valve sequencing is used to expand the operating range of flow control for some fluid beyond that which a single control valve could muster. Once again pH control provides a suitable example to illustrate an application of this form of sequencing.

pH is an especially challenging application of process control because the dynamic range of the process is enormous. Each unit of pH value change represents a *ten-fold* change in hydrogen ion concentration within the process liquid. This means the difference in ion concentration between a process liquid having a value of 10 pH and a process liquid having a value of 7 pH is a factor of *one thousand!* Consequently, the flow rate of reagent necessary to neutralize a process liquid stream may vary widely. It is quite possible that a control valve sized to handle minimum flow will simply be too small to meet the demands of high flow when needed. Yet, a control valve sized large enough to meet the maximum flow rate may be too large to precisely "turn down" when just a trickle of reagent is needed.

This general control problem was encountered by automotive engineers in the days when *carburetors* were used to mix gasoline with air prior to combustion in an engine. A carburetor is a mechanical air flow control device using a "butterfly" valve element to throttle air flow into the engine, and a venturi element producing vacuum to aspirate fuel droplets into the air stream to create an air-fuel mixture. A carburetor with a butterfly valve and flow tube sized to idle well and respond to the needs of in-town driving would not flow enough air to provide high-end performance. Conversely, a large carburetor suitable for performance driving would be almost uncontrollable for low-speed and idling operation. Their solution to this problem was the *progressive carburetor*, having two butterfly valves to throttle the flow of air into the engine. One butterfly valve handled low amounts of air flow only, while a larger butterfly valve opened up only when the accelerator pedal was nearly at its maximum position. The combination of two differently-sized butterfly valves – progressively opened – gave drivers the best of both worlds. Now, an automobile engine could perform well both at low power levels and at high power levels.

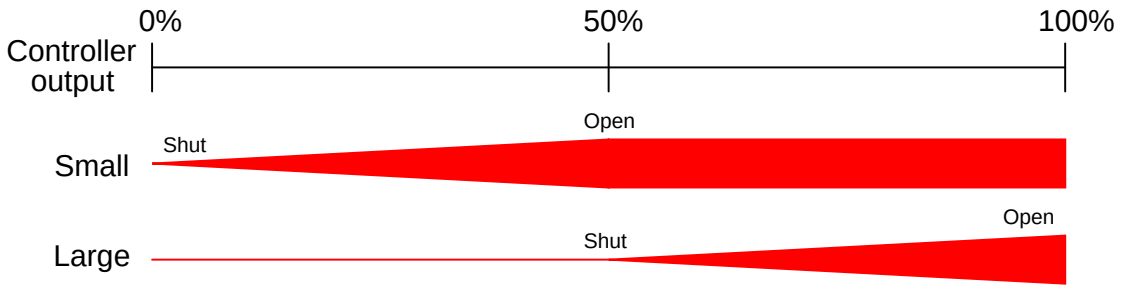
On a fundamental level, the problem faced in pH control as well as by early automotive engineers is the same thing: insufficient *rangeability*. Some processes demand a greater range of control than any single valve can deliver, and it is within these processes that a pair of progressively-sequenced control valves is a valid solution.

Applying this solution to a pH control process where the incoming liquid always has a high pH value, and must be neutralized with acid:



Proper sequencing of the small and large acid control valves is shown in the table and the graph:

Controller output (%)	I/P output (PSI)	Small acid valve (stem position)	Large acid valve (stem position)
0 %	3 PSI	fully shut	fully shut
25 %	6 PSI	half-open	fully shut
50 %	9 PSI	fully open	fully shut
75 %	12 PSI	fully open	half-open
100 %	15 PSI	fully open	fully open

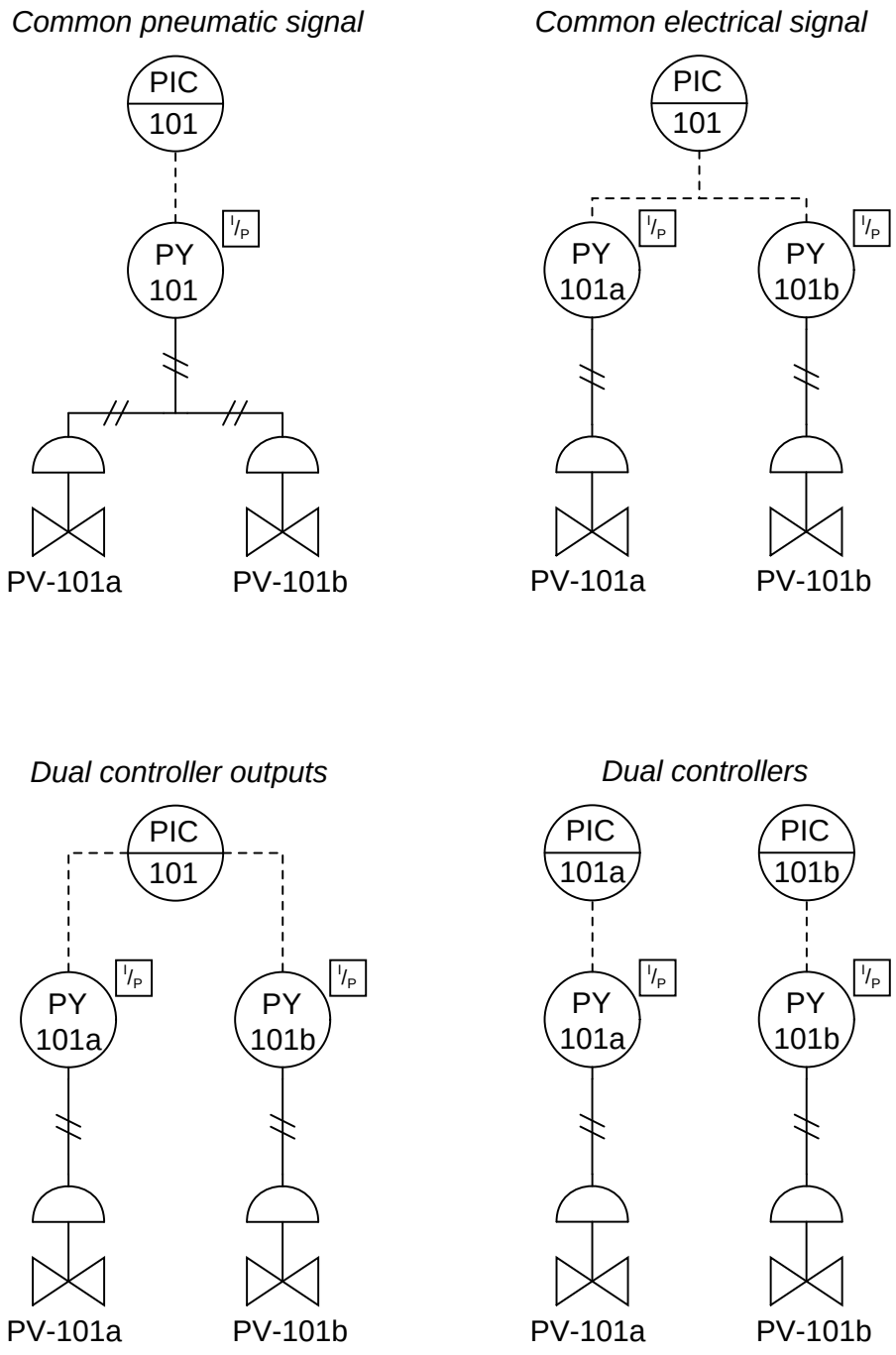


With the two acid control valves sequenced progressively, the control system will have sufficient rangeability to handle widely varying process conditions.

Valve sequencing methods

In all previous control valve sequencing examples shown, both control valves received the same pneumatic signal from a common I/P (current-to-pressure) converter. Sequencing between the two valves was a matter of proper bench-set pressure ranges.

Several alternative methods exist for control valve sequencing, as shown in the following illustration:



The common pneumatic signal approach (one controller, one I/P transducer) is simple but suffers from the disadvantage of slow response, since one I/P transducer must drive two pneumatic actuators. Response time may be improved by adding a pneumatic *volume booster* between the I/P and the valve actuators, or by adding a positioner to at least one of the valves. Either of these solutions works by the same principle: reducing the air volume demand on the one I/P transducer.

Wiring two I/P transducers in series so they share the exact same current is another way to split-range two control valves. This approach does not suffer from slow response, since each valve has its own dedicated I/P transducer to supply it with actuating air. We now have a choice where we implement the split ranges: we can do it in each of the I/P transducers (with non-standard I/P calibrations) *or* in the valve bench-set ranges as before. Since it is generally easier to re-range an I/P than it is to rebuild a control valve with a different spring (to give it a different actuating pressure range), this approach has the advantage of convenient configuration. A disadvantage of this approach is the extra demand placed on the controller's output signal circuitry: one must be careful to ensure the two series-connected I/P converters do not drop too much voltage at full current, or else the controller may have difficulty driving both in series. Another (potential) disadvantage of series-connected valve devices in one current loop is the inability to install "smart" instruments communicating with the HART protocol, since multiple devices on the same loop will experience address conflicts[19].

A very common way to implement split-ranging is to use a controller with multiple 4-20 mA outputs. This is very easy to do if the controller is part of a large system (e.g. a DCS or a PLC with multiple output channels). Now, each valve has its own dedicated wire pair for control. A further advantage of dual controller outputs is the ability to perform the split-range sequencing within the controller itself, which is often easier than re-ranging an I/P or calibrating a valve positioner.

Dual controllers are an option only for specialized applications requiring different degrees of responsiveness for each valve, usually for exclusive or progressive split-ranging applications only. Care must be taken to ensure the controllers' output signals do not wander outside of their intended ranges, or that the controllers do not begin to "fight" each other in trying to control the same process variable[20].

An important consideration – and one that is easily overlooked – in split-range valve systems is *fail-safe mode*. As discussed in a previous section of this chapter (on page 1305), the basis of fail-safe control system design is that the control valve(s) must be chosen to fail in the mode that is safest for the process in the event of actuating power loss or control signal loss. The actions of all other instruments in the loop should then be selected to complement the valves' natural operating mode.

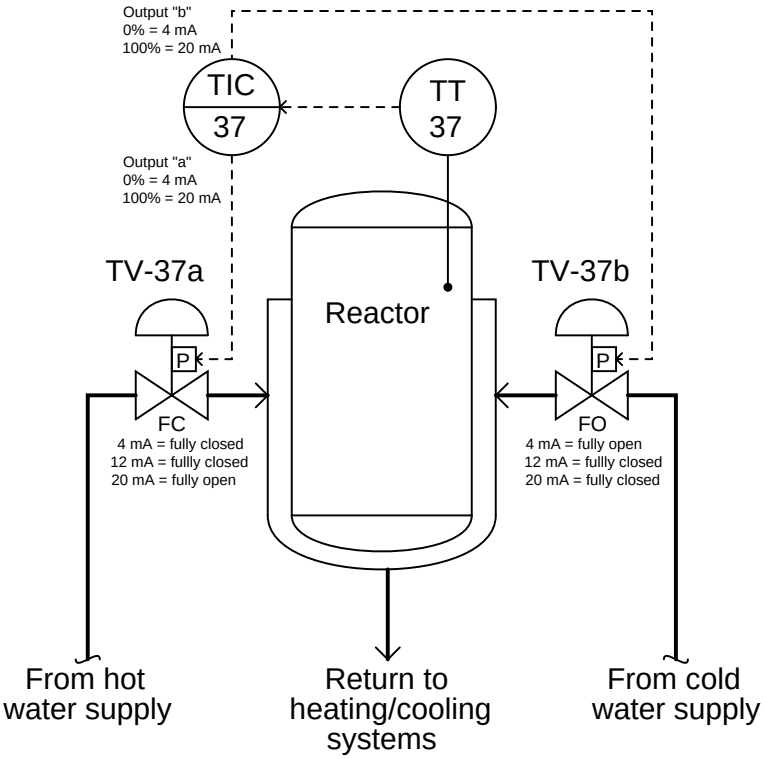
In control systems where valves are split-ranged in either complementary or exclusive fashion, one control valve will be fully closed and the other will be fully open at each extreme end of the signal range (e.g. at 4 mA and at 20 mA). If the control valves are driven by the same controller signal, the failure modes of the two valves must likewise be opposite each other: one will fail open

[19] Although the HART standard does support "multidrop" mode where multiple devices exist on the same current loop, this mode is digital-only with no analog signal support. Not only do many host systems not support HART multidrop mode, but the relatively slow data communication rate of HART makes this choice unwise for most process control applications. If analog control of multiple HART valve positioner devices from the same 4-20 mA signal is desired, the address conflict problem may be resolved through the use of one or more *isolator* devices, allowing all devices to share the same analog current signal but isolating each other from HART signals.

[20] Both controllers should be equipped with provisions for reset windup control (or have no integral action at all), such that the output signal values are predictable enough that they behave as a synchronized pair rather than as two separate controllers.

while the other fails closed if the signal goes dead or if air pressure is lost. However, if it is deemed safer for the process to have the two valves fail in the same state – for example, to both fail closed in the event of air pressure or signal loss – it is still possible to sequence them for complementary or exclusive control action by driving the two valves with different output signals. In other words, split-ranging two control valves so they normally behave in opposite fashion does *not* necessarily mean the two valves must fail in opposite states.

As an example, consider the following temperature control system supplying either hot water or chilled water to a "jacket" surrounding a chemical reactor vessel. The purpose of this system is to control temperature within the reactor to a constant setpoint value, regardless of the chemical reaction's thermal properties. If the reaction inside the vessel is *exothermic* (releasing heat), the control system will respond by sending chilled water to the jacket to remove that heat. If the reaction inside the vessel is *endothermic* (absorbing heat), the control system will respond with hot water to the jacket to add heat. Chemical piping in and out of the reactor vessel has been omitted from this P&ID for simplicity, so we can focus just on the reactor's temperature control system:

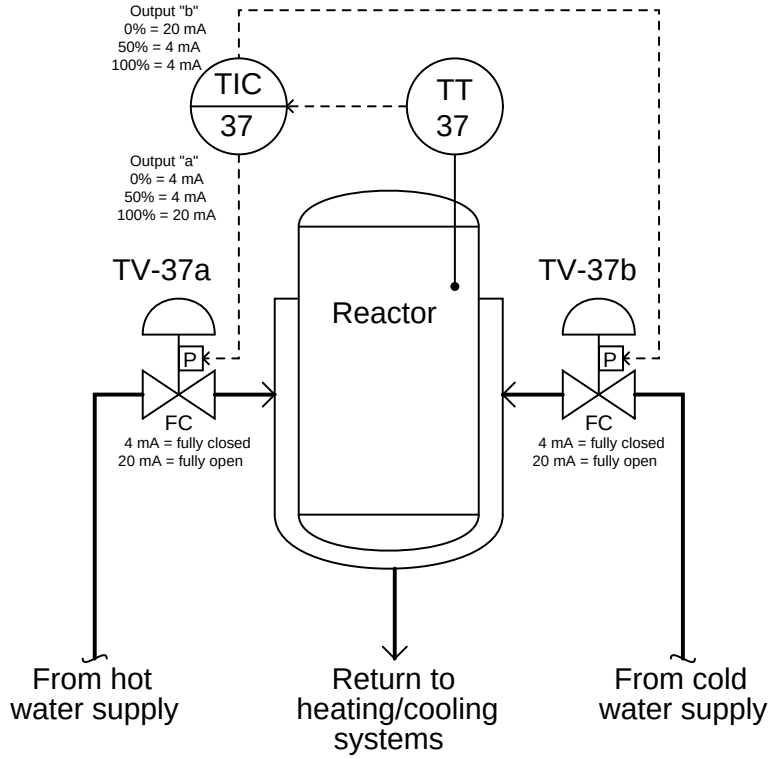


Here, the controller has been configured for dual-output operation, where the output value drives two identical 4-20 mA signals to the control valve positioners, which directly input the current signals from the controller without the need for I/P transducers in between. The hot water valve (TV-37a) is fail-closed (FC) while the cold water valve (TV-37b) is fail-open (FO). Half-range positioner calibrations provide the exclusive sequencing necessary to ensure the two valves are never open simultaneously – TV-37b operates on the lower half of the 4-20 mA signal range (4-12 mA), while

TV-37a operates on the upper half (12-20 mA).

Consider the effects from the controller (TIC-37) losing power. Both 4-20 mA signals will go dead, driving both valves to their fail-safe modes: hot water valve TV-37a will fully close, while cold water valve TV-37b will fully open. Now consider the effects of air pressure loss to both valves. With no air pressure to operate, the actuators will spring-return to their fail-safe modes: once again hot water valve TV-37a will fully close, while cold water valve TV-37b will fully open. In both failure events, the two control valves assume consistent states, ensuring the reactor will cool down rather than heat up.

Now imagine someone reconfigures the system, using identical control valves (signal-to-open, fail-closed) for both hot and cold water supply, and a different program in the controller to exclusively sequence two different 4-20 mA current signals:



Consider the effects from the controller (TIC-37) losing power. Both 4-20 mA signals will go dead, driving both valves to their fail-safe modes: fully closed. Now consider the effects of air pressure loss to both valves. With no air pressure to operate, the actuators will spring-return to their fail-safe modes: once again both control valves fully close. In both failure events, the two control valves assume consistent states where the reactor is neither heated nor cooled, but rather left to assume its own temperature. The failure modes of both valves are still consistent regardless of the nature of the fault, but note how this scheme allows both valves to fail in the same mode if that is what we deem safest for the process.

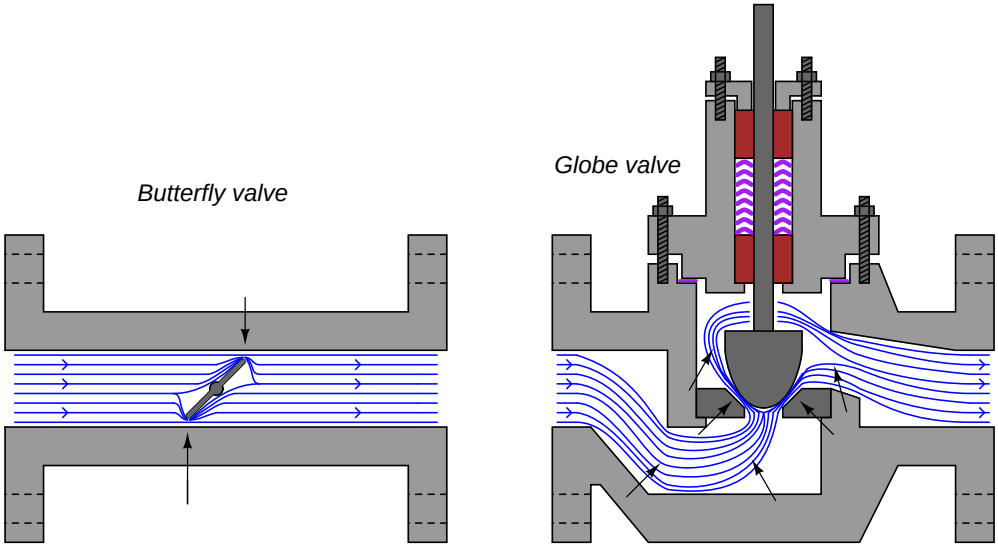
As with all fail-safe system designs, we begin by choosing the proper fail-safe mode for each

control valve *as determined by the nature of the process*, not by what we would consider the simplest or easiest-to-understand instrument configurations. Only after we have chosen each valve's failure mode do we design the rest of the system to behave they way we wish. This includes split-range sequencing: where and how we sequence the valve operation is a decision to be made only after the valves' natural fail-safe states are chosen based on the needs of process safety.

25.1.12 Control valve sizing

When control valves operate between fully open and fully shut, they serve much the same purpose in process systems as resistors do in electric circuits: to dissipate energy. Like resistors, the form that this dissipated energy takes is mostly heat, although some of the dissipated energy manifests in the form of vibration and noise[21].

In most control valves, the dominant mechanism of energy dissipation comes as a result of turbulence introduced to the fluid as it travels through constrictive portions of the valve trim. The following illustration shows these constrictive points within two different control valve types (shown by arrows):

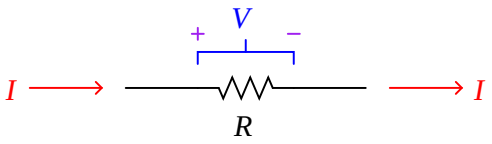


The act of choosing an appropriate control valve for the expected energy dissipation is called *valve sizing*.

[21] Valve noise may be *severe* in some cases, especially in certain gas flow applications. An important performance metric for control valves is noise production expressed in decibels (dB).

Physics of energy dissipation in a turbulent fluid stream

As one might expect, control valves are rated in their ability to throttle fluid flow, much as resistors are rated in their ability to throttle the flow of electrons in a circuit. For resistors, the unit of measurement for electron flow restriction is the *ohm*: 1 ohm of resistance results in a voltage drop of 1 volt across that resistance given a current through the resistance equal to 1 ampere:



Arrows point in direction of conventional flow

The mathematical relationship between current, voltage, and resistance for any resistor is *Ohm's Law*:

$$R = \frac{V}{I}$$

Where,

R = Electrical resistance in ohms
V = Electrical voltage drop in volts
I = Electrical current in amperes

Ohm's Law is a simple, linear relationship, expressing the "friction" encountered by electric charge carriers as they slowly drift through a solid object.

When a fluid moves turbulently through any restriction, energy is inevitably dissipated in that turbulence. The amount of energy dissipated is proportional to the kinetic energy of the turbulent motion, which is proportional to the square of velocity according to the classic kinetic energy equation for moving objects:

$$E_k = \frac{1}{2}mv^2$$

If we were to re-write this equation to express the amount of kinetic energy represented by a volume of moving fluid with velocity v, it would look like this:

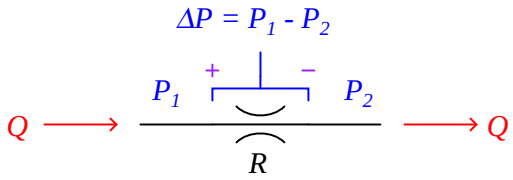
$$\text{Kinetic energy per unit volume} = \frac{1}{2}\rho v^2$$

We know that the amount of energy dissipated by turbulence in such a fluid stream will be some proportion (k) of the total kinetic energy, so:

$$\text{Energy dissipated per unit volume} = \frac{1}{2}k\rho v^2$$

Any energy lost in turbulence eventually manifests as a loss in fluid pressure. Thus, a control valve throttling a fluid flowstream will have a greater upstream pressure than downstream pressure

(assuming all other factors such as pipe size and height above ground level being the same downstream as upstream):



This pressure drop ($P_1 - P_2$, or ΔP) is equivalent to the voltage drop seen across any current-carrying resistor, and may be substituted for dissipated energy per unit volume in the previous equation[22]. We may also substitute $\frac{Q}{A}$ for velocity v because we know volumetric flow rate (Q) is the product of fluid velocity and pipe cross-section area ($Q = Av$) for incompressible fluids such as liquids:

$$P_1 - P_2 = \frac{1}{2} k \rho \left(\frac{Q}{A} \right)^2$$

Next, we will solve for a quotient with pressure drop ($P_1 - P_2$) in the numerator and flow rate Q in the denominator so the equation bears a resemblance to Ohm's Law ($R = \frac{V}{I}$):

$$P_1 - P_2 = \frac{1}{2} k \rho \frac{Q^2}{A^2}$$

$$\frac{P_1 - P_2}{Q^2} = \frac{k \rho}{2A^2}$$

$$\frac{\sqrt{P_1 - P_2}}{Q} = \sqrt{\frac{k \rho}{2A^2}}$$

[22] In case you were wondering, it is appropriate to express energy loss per unit volume in the same units of measurement as pressure. For a more detailed discussion of dimensional analysis, see section 2.9.12 starting on page 120 where Bernoulli's equation is examined and you will see how the units of $\frac{1}{2}\rho v^2$ and P are actually the same.

Either side of the last equation represents a sort of "Ohm's Law" for turbulent liquid restrictions: the left-hand side expressing fluid "resistance" in the state variables of pressure drop and volumetric flow, and the right-hand term expressing fluid "resistance" as a function of fluid density and restriction geometry. We can see how pressure drop ($P_1 - P_2$) and volumetric flow rate (Q) are not linearly related as voltage and current are for resistors, but that nevertheless we still have a quantity that acts like a "resistance" term:

$$R = \frac{\sqrt{P_1 - P_2}}{Q} \qquad\qquad R = \sqrt{\frac{k\rho}{2A^2}}$$

Where,

R = Fluid "resistance"
P_1 = Upstream fluid pressure
P_2 = Downstream fluid pressure
Q = Volumetric fluid flow rate
k = Turbulent energy dissipation factor
ρ = Mass density of fluid
A = Cross-sectional area of restriction

The fluid "resistance" of a restriction depends on several variables: the proportion of kinetic energy lost due to turbulence (k), the density of the fluid (ρ), and the cross-sectional area of the restriction (A). In a control valve throttling a liquid flow stream, only the first and last variables are subject to change with stem position, fluid density remaining relatively constant.

In a wide-open control valve, especially valves offering a nearly unrestricted path for moving fluid (e.g. ball valves, eccentric disk valves), the value of A will be at a maximum value essentially equal to the pipe's area, and k will be nearly zero[23]. In a fully shut control valve, A is zero, creating a condition of infinite "resistance" to fluid flow.

It is customary in control valve engineering to express the "restrictiveness" of any valve in terms of how much flow it will pass given a certain pressure drop and fluid specific gravity (G_f). This measure of valve performance is called *flow capacity* or *flow coefficient*, symbolized as C_v. A greater flow capacity value represents a less restrictive (less "resistive") valve, able to pass greater rates of flow for the same pressure drop. This is analogous to expressing an electrical resistor's rating in terms of conductance (G) rather than resistance (R): how many amperes of current it will pass with 1 volt of potential drop ($I = GV$ instead of $I = \frac{V}{R}$).

If we return to one of our earlier equations expressing pressure drop in terms of flow rate, restriction area, dissipation factor, and density, we will be able to manipulate it into a form expressing flow rate (Q) in terms of pressure drop and density, collecting k and A into a third term which will become flow capacity (C_v):

$$P_1 - P_2 = \frac{1}{2} k \rho \frac{Q^2}{A^2}$$

[23] In a case of minimal throttling, almost none of the fluid's kinetic energy is lost to turbulence, but rather passes right through the valve unrestricted.

First, we must substitute specific gravity (G_f) for mass density (ρ) using the following definition of specific gravity:

$$G_f = \frac{\rho}{\rho_{water}}$$

$$\rho_{water} G_f = \rho$$

Substituting and continuing with the algebraic manipulation:

$$P_1 - P_2 = \frac{1}{2} k \rho_{water} G_f \frac{Q^2}{A^2}$$

$$\frac{P_1 - P_2}{G_f} = \frac{1}{2} k \rho_{water} \frac{Q^2}{A^2}$$

$$\left(\frac{2A^2}{k\rho_{water}}\right)\left(\frac{P_1 - P_2}{G_f}\right) = Q^2$$

$$Q = \sqrt{\frac{2A^2}{k\rho_{water}}} \sqrt{\frac{P_1 - P_2}{G_f}}$$

The first square-rooted term in the equation, $\sqrt{\frac{2A^2}{k\rho_{water}}}$, is the valve capacity or C_v factor. Substituting C_v for this term results in the simplest form of valve sizing equation (for incompressible fluids):

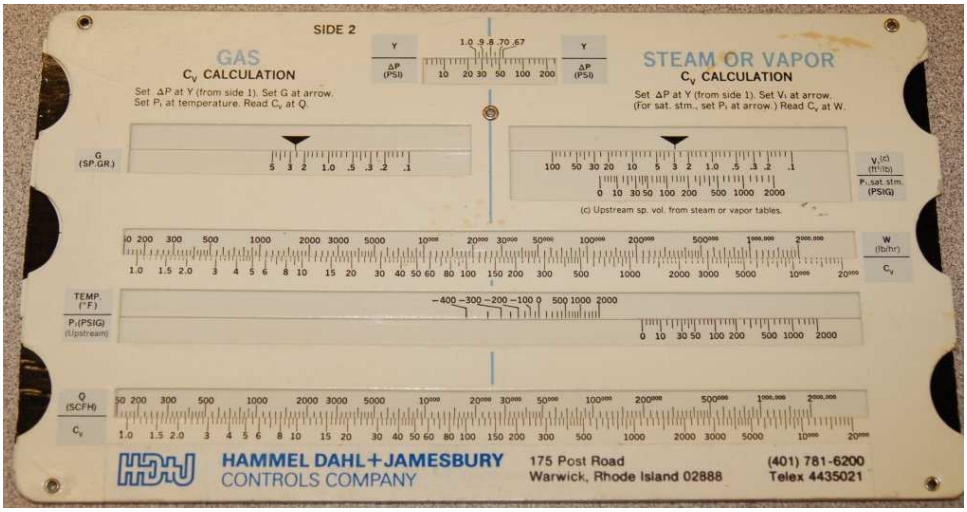
$$Q = C_v \sqrt{\frac{P_1 - P_2}{G_f}}$$

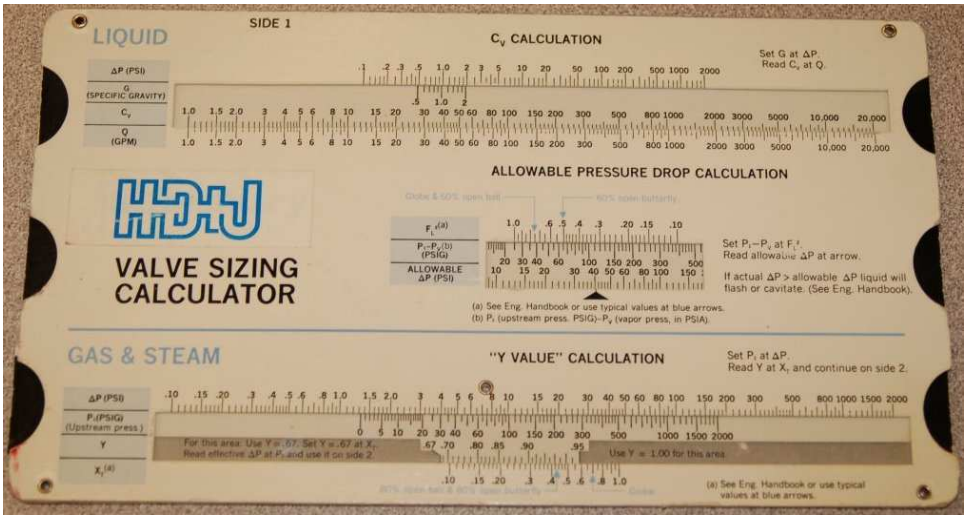
In the United States of America, C_v is defined as the number of gallons per minute of water that will flow through a valve with 1 PSI of pressure drop[24]. A similar valve capacity expression used elsewhere in the world rates valves in terms of how many cubic meters per hour of water will flow through a valve with a pressure drop of 1 bar. This latter flow capacity is symbolized as K_v.

For the best results predicting required C_v values for control valves in any service, it is recommended that you use valve sizing software provided by control valve manufacturers. Modern valve sizing software is easy to use, especially when referenced to specific models of control valve sold by that manufacturer, and is able to account for a diverse multitude of factors affecting proper sizing.

[24]The specification of certain British units of measurement for flow and pressure drop means that there is more to C_v than just $\sqrt{\frac{2A^2}{k\rho_{water}}}$. C_v also incorporates a factor necessary to account for the arbitrary choice of units.

Control valve sizing is complex enough that some valve manufacturers used to give away "slide rule" calculator devices so customers could choose the C_v values they needed with relative ease. Photographs of a two-sided valve sizing slide rule are shown here for historical reference:



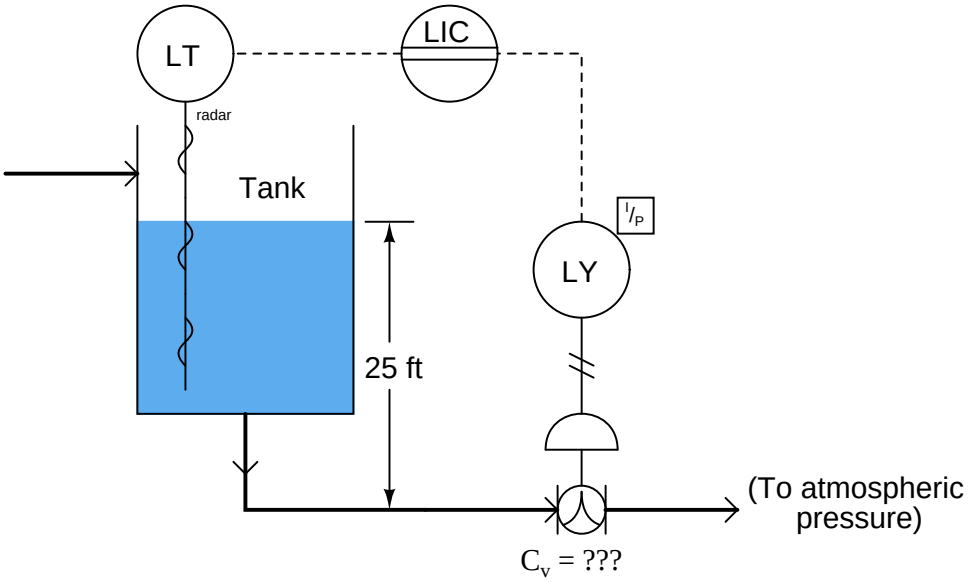


Importance of proper valve sizing

The flow coefficient of a control valve is a numerical value usually expressing maximum flow capacity. For example, a control valve with a C_v rating of 45 should flow 45 gallons per minute of water through it with a 1 PSI pressure drop *when wide open*. The flow coefficient value for this same control valve will be less than 45 when the valve position is anything less than fully open. When the control valve is in the fully shut position, its C_v value will be zero. Thus, it should be understood that C_v is truly a variable – not a constant – for any control valve, even though control valves are often specified simply by their maximum flow capacity.

It should be obvious that any control valve must be sized large enough (i.e. possess sufficient maximum C_v capacity) to flow the greatest expected flow rate in any given process installation. A valve that is too small for an application will not be able to pass enough process fluid through it when needed.

Given this fact, it may seem safe to choose a valve sized much larger than what is needed, just to avoid the possibility of not having enough flow capacity. For instance, consider this control valve sizing problem, where a characterized ball valve controls the flow rate of water out of a surge tank to maintain a constant water level 25 feet higher than the height of the valve:



According to the process engineers, the maximum expected flow rate for this valve is 470 GPM. What should the maximum C_v rating be for this valve? To begin, we must know the expected pressure drop across the valve. The 25 foot water column height upstream provides us with the means to calculate P_1:

$$P = \gamma h$$

$$P_1 = (62.4 \text{ lb/ft}^3)(25 \text{ feet})$$

$$P_1 = 1560 \text{ PSF} = 10.8 \text{ PSI}$$

There is no need to calculate P_2, since the P&ID shows us that the downstream side of the valve is vented to atmosphere, and is thus 0 PSI gauge pressure. This gives us a pressure drop of 10.8 PSI across the control valve, with an expected maximum flow rate of 470 GPM. Manipulating our flow capacity equation to solve for C_v:

$$Q = C_v \sqrt{\frac{P_1 - P_2}{G_f}}$$

$$C_v = \frac{Q}{\sqrt{\frac{P_1 - P_2}{G_f}}}$$

$$C_v = \frac{470 \text{ GPM}}{\sqrt{\frac{10.8 \text{ PSI}}{1}}}$$

$$C_v = 143$$

This tells us we need a control valve with a C_v value of *at least* 143 to meet the specified (maximum) flow rate. A valve with insufficient C_v would not be able to flow the required 470 gallons per minute of water with only 10.8 PSI of pressure drop.

Does this mean we may safely over-size the valve? Would there be any problem with installing a control valve with a C_v value of 300? The general answer to these questions is that over-sized valves may be problematic. Not only is there the possibility of allowing too much flow under wide-open conditions (consider whatever process vessels and equipment lie downstream of the oversized valve), but also that the process will be difficult to control under low-flow conditions.

In order to understand how an over-sized control valve leads to unstable control, an exaggerated example is helpful to consider: imagine installing a fire hydrant valve on your kitchen sink faucet[25]. Certainly, a wide-open hydrant valve would allow sufficient water flow into your kitchen sink. However, most of this valve's usable range of throttling will be limited to the first *percent* of stem travel. After the valve is opened just a few percent from fully shut, restrictions in the piping of your house's water system will have limited the flow rate to its maximum, thus rendering the rest of the

[25] For those readers who may be unfamiliar with American terminology, a *fire hydrant* is a large hand valve installed at intervals along public roadways, allowing connection of fire hoses to an underground water supply pipe in the event of an emergency fire. These valves are quite large, and would be comically oversized if installed inside a person's house, for any purpose.

valve's stem travel capacity utterly useless. It would be challenging indeed to try filling a drinking cup with water from this hydrant valve: just a little bit too much stem motion and the cup would be subjected to a full-flow stream of water!

Control valve over-sizing is a common problem in industry, often created by future planning for expanded process flow. "If we buy a large valve now," so the reasoning goes, "we won't have to replace a smaller valve with a large valve when the time comes to increase our production rate." In the interim period when that larger valve must serve to control a meager flow rate, however, problems caused by poor control quality may end up costing the enterprise more than the cost of an additional valve.

Gas valve sizing

Sizing a control valve for gas or vapor service is more complicated than for liquid service, due to the compressibility of gases and vapors. As a gas or vapor compresses with changes in pressure, its density changes correspondingly. In previous mathematical analyses of fluid flow restriction, one of our assumptions was that fluid density (ρ) remained constant. This assumption may hold true for some flowing gas conditions as well, provided minimal pressure changes within the path of flow. However, for most control valve applications where the very purpose of the valve is to introduce substantial pressure changes in a fluid stream, the assumption of constant fluid density is unrealistic.

Shown here is one of the simpler gas valve sizing equations you will encounter:

$$Q = 963\, C_v \sqrt{\frac{\Delta P(P_1 + P_2)}{G_g T}}$$

Where,

Q = Gas flow rate, in units of Standard Cubic Feet per Hour (SCFH)
C_v = Valve capacity coefficient
ΔP = Pressure dropped across valve, pounds per square inch differential (PSID)
P_1 = Upstream valve pressure, pounds per square inch absolute (PSIA)
P_2 = Downstream valve pressure, pounds per square inch absolute (PSIA)
G_g = Specific gravity of gas (Air at standard temperature and pressure = 1.0)
T = Absolute temperature of gas in degrees Rankine (^{o}R)

This equation holds true only for "subcritical" flow, where the moving gas stream velocity never approaches the speed of sound[26]. Other equations exist for calculating flow rates of gas through control valves when sonic flows are achieved. Note the inclusion of absolute pressures in this equation, and not just differential pressure (ΔP, or $P_1 - P_2$). This is intended to correct for effects related to compression of the gas under pressure.

Valve sizing is complicated enough, both for liquid and gas service, that the use of valve sizing computer software is strongly recommended as opposed to hand-calculations. The number of important parameters, nonlinear factors, and alternative equations relevant to control valve sizing are numerous enough to bewilder most technicians (and more than a few engineers). Valve sizing

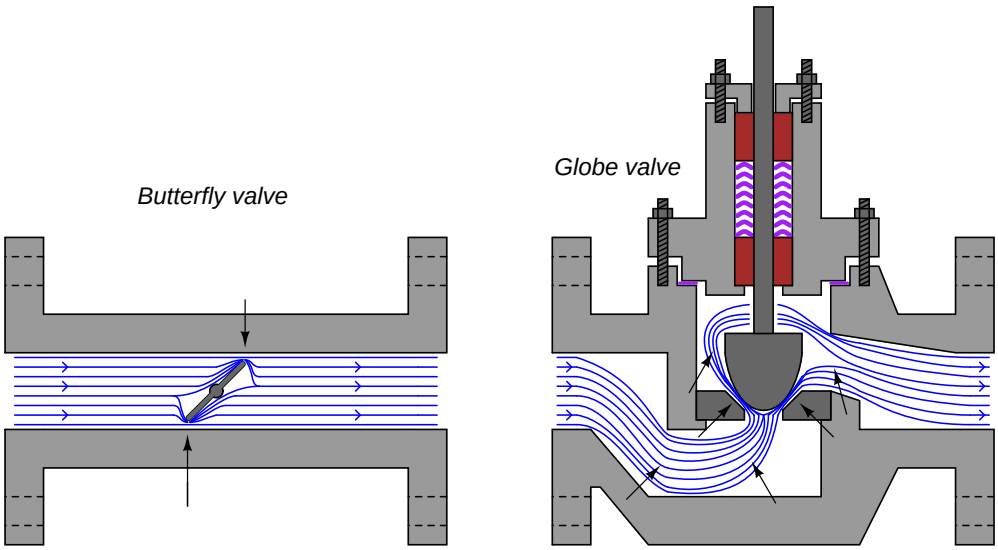
[26] The *ISA Handbook of Control Valves* cites this equation as being valid for conditions where the valve's downstream pressure (P_2) is equal to or greater than one-half the upstream pressure (P_1), with both pressures expressed in absolute units. In other words, $P_2 \geq 0.5P_1$ or $P_1 \leq 2P_2$. An upstream:downstream pressure ratio in excess of 2:1 usually means flow through a valve will become *choked*.

software will also predict noise levels generated by the valve, and in many cases specify actual valve trim styles offered by the manufacturer for mitigating problems such as noise.

Relative flow capacity

The flow capacity of a valve (C_v) is a quantitative rating of its ability to pass a fluid flow for a set of given pressure and density conditions. C_v may be predicted, or empirically measured, for any type of control valve given the proper information.

Not all control valve types exhibit the same C_v coefficients, however, for the same pipe size. A 4 inch butterfly valve, for example, has a much greater full-open C_v rating than a 4 inch globe valve, due to the much more direct path it offers to a moving fluid. A simple comparison of these two valve types clearly shows why this is true (note the "constriction" points labeled with arrows):



A globe valve is simply more efficient at generating fluid turbulence – and therefore dissipating fluid kinetic energy – than a butterfly valve of the same pipe size, because the globe valve design forces the fluid to change direction more often and in different ways.

One way to help quantify a particular valve design's ability to throttle fluid flow is to express this ability as a ratio of flow coefficient (C_v) versus cross-sectional pipe area. The basic principle here is that we should expect the C_v of any particular valve design to be proportional to pipe area (e.g. a ball valve with twice the pipe area should have twice the flow capacity, all other factors being equal), and therefore a ratio of these two quantities should be fairly constant for any valve design. Since we know the area of a pipe is proportional to the square of either radius or diameter ($A = \frac{\pi d^2}{4}$ or $A = \pi r^2$), we may simplify this ratio by omitting all constants such as π and simply relating C_v factor to the square of pipe diameter (d^2). This ratio is called the *relative flow capacity*, or C_d:

$$C_d = \frac{C_v}{d^2}$$

Several valve capacity factors (C_d) for different control valve types are shown here[27], assuming full-area trim and a full-open position:

Valve design type	C_d
Single-port globe valve, ported plug	9.5
Single-port globe valve, contoured plug	11
Single-port globe valve, characterized cage	15
Double-port globe valve, ported plug	12.5
Double-port globe valve, contoured plug	13
Rotary ball valve, segmented	25
Rotary ball valve, standard port (diameter $\approx 0.8d$)	30
Rotary butterfly valve, 60^o, no offset seat	17.5
Rotary butterfly valve, 90^o, offset seat	29
Rotary butterfly valve, 90^o, no offset seat	40

As you can see from a comparison of C_d values, a no-offset butterfly valve has nearly *4 times* the flow capacity of a single-ported contoured-plug globe valve of the same pipe size ($C_d = 40$ versus $C_d = 11$). This makes butterfly valves advantageous in applications where large flow capacities must be achieved at minimal cost, such as in air handling (HVAC) systems for commercial buildings and combustion air controls for large industrial burners.

[27]Source for C_d factors: Chapter 4.17: Valve Sizing of Béla Lipták's *Instrument Engineer's Handbook, Process Control (Volume II), Third Edition*, page 590.

25.1.13 Control valve characterization

When control valves are tested in a laboratory setting, they are connected to a piping system that is able to provide a nearly constant pressure difference between upstream and downstream ($P_1 - P_2$). With a fluid of constant density and a constant pressure drop across the valve, flow rate becomes a direct function of flow coefficient (C_v). This is clear from an examination of the basic valve capacity equation:

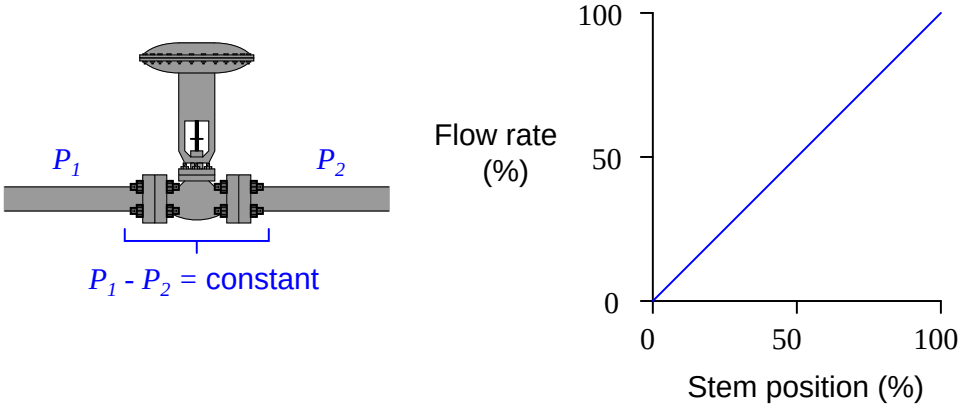
$$Q = C_v \sqrt{\frac{P_1 - P_2}{G_f}}$$

(If pressures and specific gravity are constant . . .)

$$Q = kC_v$$

As discussed in an earlier section of this chapter (see page 1340), the amount of "resistance" offered by a restriction of any kind to a turbulent fluid depends on the cross-sectional area of that restriction and also the proportion of fluid kinetic energy dissipated in turbulence. If a control valve is designed such that the combined effect of these two parameters vary linearly with stem motion, the C_v of the valve will likewise be proportional to stem position. That is to say, the C_v of the control valve will be approximately half its maximum rating with the stem position at 50%; approximately one-quarter its maximum rating with the stem position at 25%; and so on.

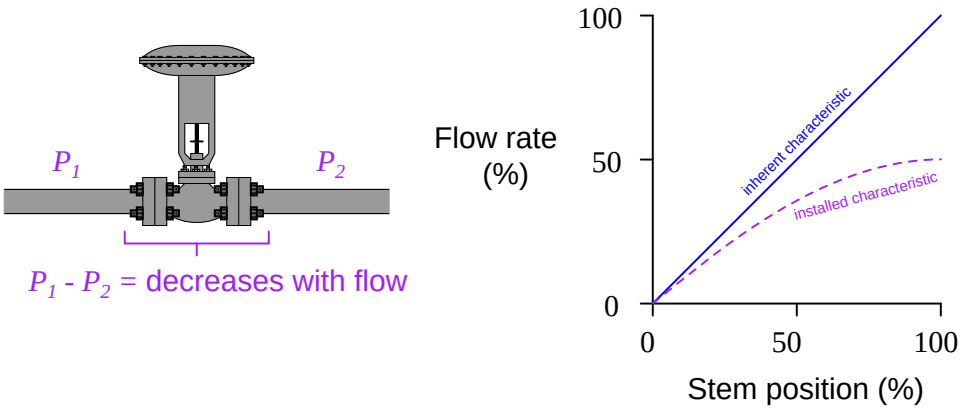
If such a valve is placed in a laboratory flow test piping system with constant differential pressure and constant fluid density, the relationship of flow rate to stem position will be linear:



However, most real installations do not place the control valve under the same conditions. Due to frictional pressure losses in piping and changes in supply/demand pressures that vary with flow rate, a typical control valve "sees" substantial changes in differential pressure as its controlled flow rate changes. Generally speaking, the pressure drop available to the control valve will *decrease* as flow rate *increases*.

The result of this pressure drop versus flow relationship is that the actual flow rate of the same valve installed in a real process will *not* linearly track valve stem position. Instead, it will "droop"

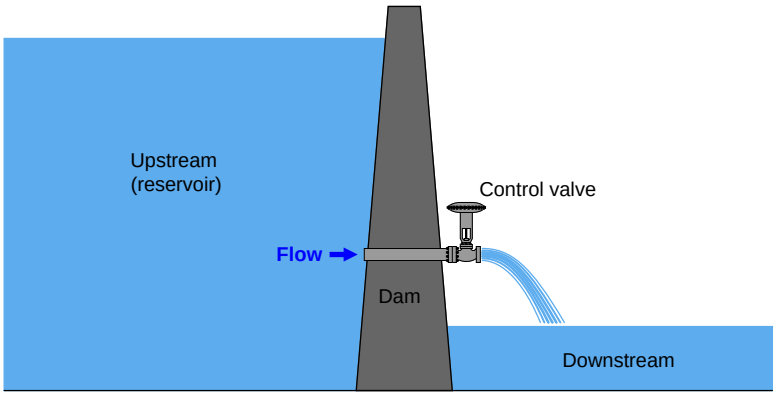
as the valve is further opened. This "drooping" graph is called the valve's *installed characteristic*, in contrast to the *inherent characteristic* exhibited in the laboratory with constant pressure drop:



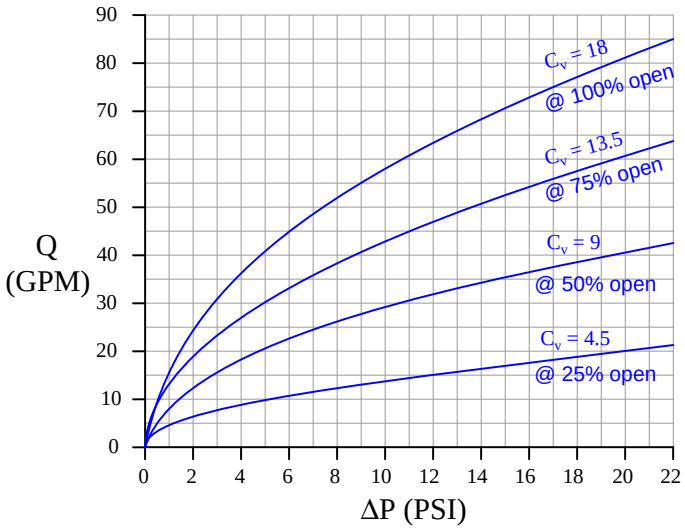
Each time the stem lifts up a bit more to open the valve trim further, flow increases, but not as much as at lower-opening positions. It is a situation of *diminishing returns*, where we still see increases in flow as the stem lifts up, but to a lesser and lesser degree.

In my years of teaching, I have found this concept of "installed characteristic" to be elusive for many students to grasp. In the interest of clarifying the concept, I wish to present a pair of contrasting scenarios using realistic numbers.

First, let us imagine a control valve installed at the base of a dam, letting water out of the reservoir. Given a constant height of water in the reservoir, the upstream (hydrostatic) pressure at the valve will likewise be constant. Let's assume this constant upstream pressure will be 20 PSI (corresponding to approximately 46 feet of water column above the valve inlet). With the valve discharging into the air, downstream pressure will essentially be zero. This set of upstream and downstream conditions guarantees a constant pressure drop of 20 PSI across our control valve at all times, for all flow conditions:



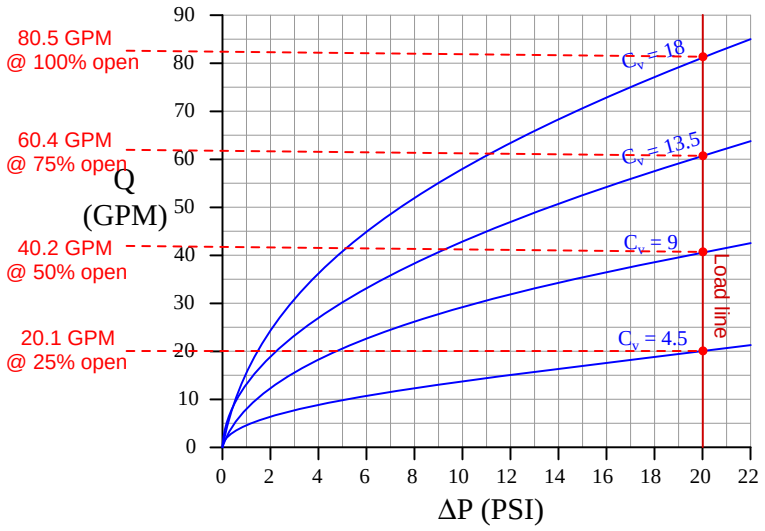
Furthermore, let us assume the control valve has a "linear" inherent characteristic and a maximum flow capacity (C_v rating) of 18. This means the valve's C_v will be 18 at 100% open, 13.5 at 75% open, 9 at 50% open, 4.5 at 25% open, and 0 at fully closed (0% open). We may plot the behavior of this control valve at these four stem positions by graphing the amount of flow through the valve for varying degrees of pressure drop across the valve. The result is a set of *characteristic curves*[28] for our hypothetical control valve:



Each curve on the graph traces the amount of flow through the valve at a constant stem position, for different amounts of applied pressure drop. For example, looking at the curve representing 50% open ($C_v = 9$), we can see the valve should flow about 42 GPM at 22 PSI, about 35 GPM at 15 PSI, about 20 GPM at 5 PSI, and so on. Of course, we can obtain these same flow figures just by evaluating the formula $Q = C_v\sqrt{\Delta P}$ (which is in fact what I used to plot these curves), but the point here is to learn how to interpret the graph.

[28] For those readers with an electronics background, the concept of "characteristic curves" for a control valve is *exactly* the same as that of characteristic curves for transistors. Instead of plotting the amount of current a transistor will pass given varying amounts of supply voltage, we are plotting how much water a valve will flow given varying amounts of supply pressure.

We may use this set of characteristic curves to determine how this valve will respond in *any* installation by superimposing another curve on the graph called a *load line*[29], describing the pressure drop available to the valve at different flow rates. Since we know our hypothetical dam applies a constant 20 PSI across the control valve for all flow conditions, the load line for the dam will be a vertical line at 20 PSI:



By noting the points of intersection[30] between the valve's characteristic curves and the load line, we may determine the flow rates from the dam at those stem positions:

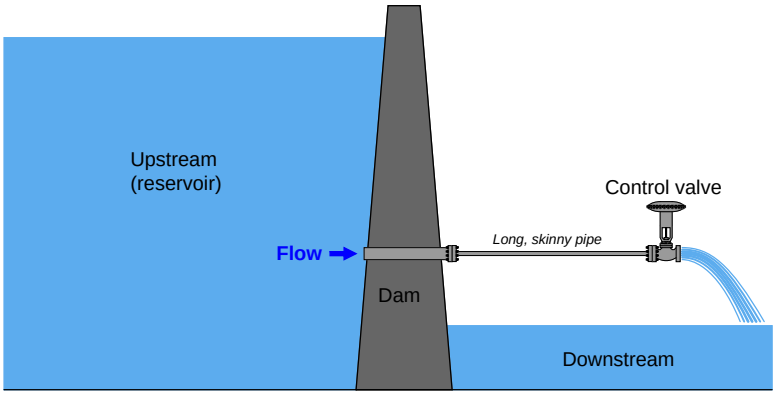
Opening (%)	C_v	Flow rate (GPM)
0	0	0
25	4.5	20.1
50	9	40.2
75	13.5	60.4
100	18	80.5

If we were to graph *this* table, plotting flow versus stem position, we would obtain a very linear graph. Note how 50% open gives us twice as much flow as 25% open, and 100% open nearly twice as much flow as 50% open. This tells us our control valve will respond linearly when pressed into service on this dam, with a constant pressure drop.

[29] Once again, the exact same concept applied in transistor circuit analysis finds application here in control valve behavior!

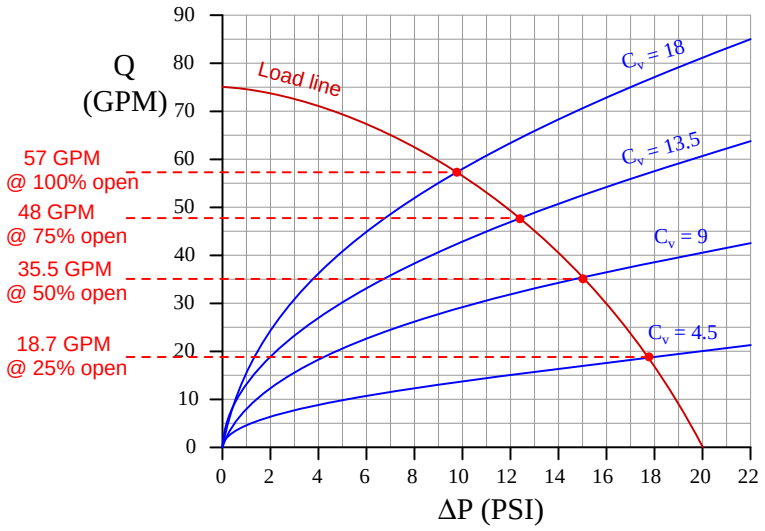
[30] Load line plots are a graphical method of solving nonlinear, simultaneous equations. Since each curve represents a set of solutions to a particular equation, the intersection of two curves represents values uniquely satisfying both equations at the same time.

But what if we alter the scenario so that the pressure drop across the valve does *not* remain constant as flow through the valve changes? Suppose the valve is not closely coupled to the dam, but rather receives water through a narrow (restrictive) pipe:



In this installation, the narrow pipe drops pressure of its own due to friction between the rushing water and the interior walls, leaving less upstream pressure at the valve with greater amounts of flow. The control valve still drains to atmosphere, so its downstream pressure is still a constant 0 PSIG, but now its upstream pressure will diminish as flow increases. How will this affect the valve's performance?

We may turn to the same set of characteristic curves for an answer to this question. All we must do is plot a new load line describing the pressure available to the valve at different flow rates, and once again look for the points of intersection between this load line and the valve's characteristic curves. For the sake of our hypothetical example, I have sketched an arbitrary "load line" (actually a load *curve*) showing how the valve's pressure falls off as flow rises:

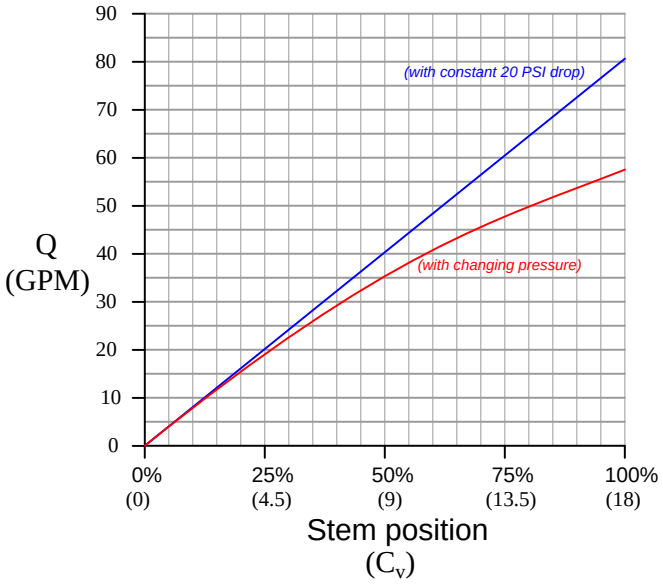


Now we see a definite nonlinearity in the control valve's behavior. No longer does a doubling of stem position (from 25% to 50%, or from 50% to 100%) result in a doubling of flow rate[31]:

Opening (%)	C_v	Flow rate (GPM)
0	0	0
25	4.5	18.7
50	9	35.5
75	13.5	48
100	18	57

[31] Not only is the response of the valve altered by this degradation of upstream pressure, but we can also see from the load line that a certain maximum flow rate has been asserted by the narrow pipe which did not previously exist: 75 GPM. Even if we unbolted the control valve from the pipe and let water gush freely into the atmosphere, the flow rate would saturate at only 75 GPM because that is the amount of flow where all 20 PSI of hydrostatic "head" is lost to friction in the pipe. Contrast this against the close-coupled scenario, where the load line was vertical on the graph, implying no theoretical limit to flow at all! With an absolutely constant upstream pressure, the only limit on flow rate was the maximum C_v of the valve (analogous to a perfect electrical voltage source with zero internal resistance, capable of sourcing any amount of current to a load).

If we plot the valve's performance in both scenarios (close-coupled to the dam, versus at the end of a restrictive pipe), we see the difference very clearly:



The "drooping" graph shows how the valve responds when it does not receive a constant pressure drop throughout the flow range. This is how the valve responds when *installed* in a non-ideal process, compared to the straight-line response it exhibits under *ideal* conditions of constant pressure. This what we mean by "installed" characteristic versus "ideal" or "inherent" characteristic.

Pressure losses due to fluid friction as it travels down pipe is just one cause of valve pressure changing with flow. Other causes exist as well, including pump curves and frictional losses in other system components such as filters and heat exchangers. Whatever the cause, any piping system that fails to provide constant pressure across a control valve will "distort" the valve's inherent characteristic in the same "drooping" manner, and this must be compensated in some way if we desire linear response from the valve.

Not only does the diminishing pressure drop across the valve mean we cannot achieve the same full-open flow rate as in the laboratory (with a constant pressure drop), but it also means the control valve responds differently at various points along its range. Note how the installed characteristic graph is relatively steep at the beginning where the valve is nearly closed, and how the graph is almost flat at the end where the valve is nearly full-open. The rate of response (rate-of-change of flow Q compared to stem position x, which may be expressed as the derivative $\frac{dQ}{dx}$) is much greater at low flow rates than it is at high flow rates, all due to diminished pressure drop at higher flow rates. This means the valve will respond more "sensitively" at the low end of its travel and more "sluggishly" at the high end of its travel.

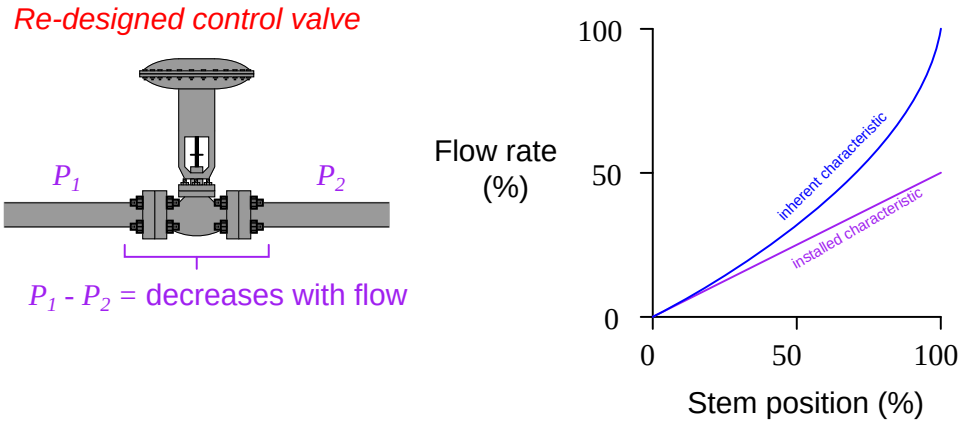
From the perspective of a flow control system, this varying valve responsiveness means the system will be unstable at low flow rates and slow-responding at high flow rates. At low flow rates, there the valve is nearly closed, any small movement of the valve stem will have a relatively large effect on flow. However, at high flow rates, a much greater stem motion will be required to effect the same

change in flow. Thus, the control system will tend to over-react at low flow rates and under-react at high flow rates. Oscillations may occur at low flow rates, and large deviations from setpoint at high flow rates as a result of this "distorted" valve behavior.

The root cause of the problem – a varying pressure drop caused by frictional losses in the piping and other factors – generally cannot be eliminated. This means there is no way to regain maximum flow capacity short of replacing the control valve with one having a greater C_v rating[32]. However, there is a clever way to flatten the valve's responsiveness to achieve a more linear characteristic, and that is to purposely design the valve such that its inherent characteristic complements the process "distortion" caused by changing pressure drop. In other words, we design the control valve trim so it opens up gradually during the initial stem travel (near the closed position), then opens up more rapidly during the final stages of stem travel (near the full-open position). With the valve made to open up in a nonlinear fashion inverse to the "droop" caused by the installed pressure changes, the two non-linearities should cancel each other and yield a more linear response.

[32] Even then, achieving the ideal maximum flow rate may be impossible. Our previous 100% flow rate for the valve was 80.5 GPM, but this goal has been rendered impossible by the narrow pipe, which according to the load line limits flow to an absolute maximum of 75 GPM (even with an infinitely large control valve).

This re-design will give the valve a nonlinear characteristic when tested in the laboratory with constant pressure drop, but the installed behavior should be more linear:

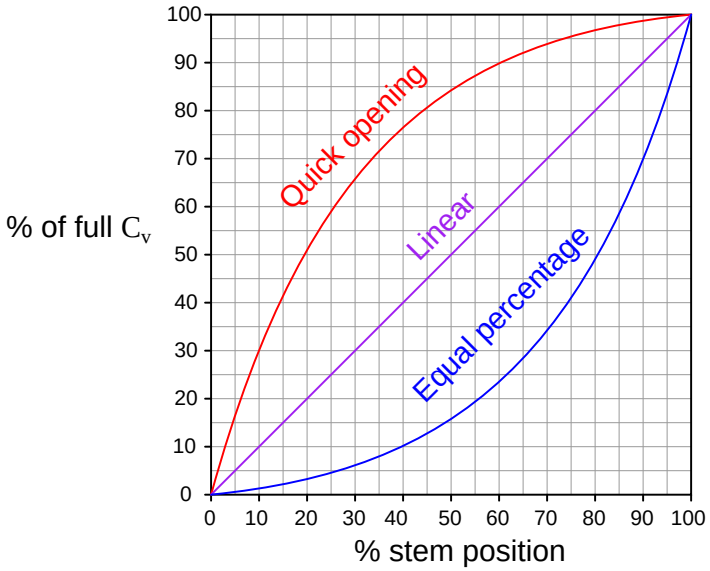


Now, control system response will be consistent at all points within the controlled flow range, which is a significant improvement over the original state of affairs.

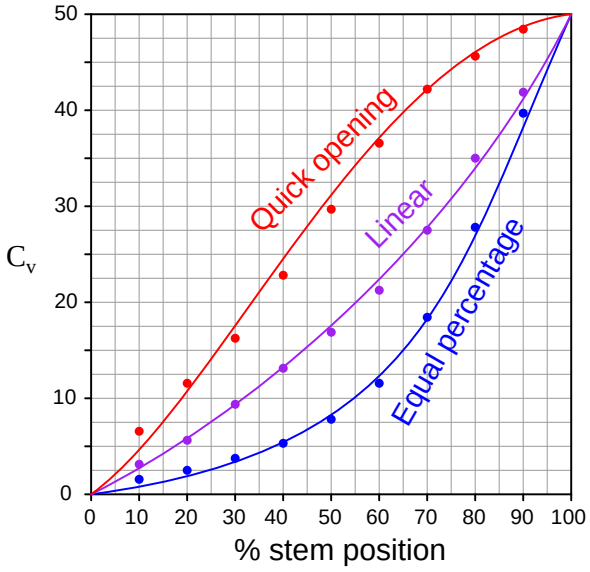
Control valve trim is manufactured in a variety of different "characteristics" to provide the desired installed behavior. The two most common inherent characteristics are *linear* and *equal percentage*. "Linear" valve trim exhibits a fairly proportional relationship between valve stem travel and flow capacity (C_v), while "equal percentage" trim is decidedly nonlinear. A control valve with "linear" trim will exhibit consistent responsiveness only with a constant pressure drop, while "equal percentage" trim is designed to counter-act the "droop" caused by changing pressure drop when installed in a process system.

Another common inherent valve characteristic available from manufacturers is *quick-opening*, where the valve's C_v increases dramatically during the initial stages of opening, but then increases at a much slower rate for the rest of the travel. Quick-opening valves are often used in pressure-relief applications, where it is important to rapidly establish flow rate during the initial portions of valve stem travel.

The standard "textbook" comparison of quick-opening, linear, and equal-percentage valve characteristics usually looks something like the following graph:

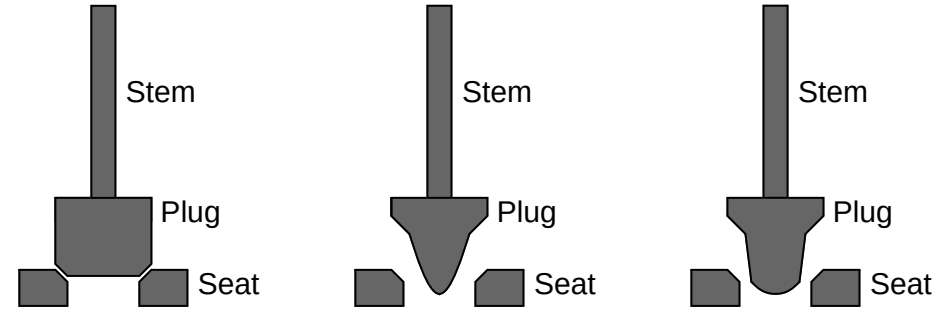


A graph showing valve characteristics taken from actual manufacturers' data on valve performance[33] shows a more moderate picture:



[33]Data for the three graphs were derived from actual C_v factors published in Fisher's ED, EAD, and EDR sliding-stem control valve product bulletin (51.1:ED). I did not copy the exact data, however; I "normalized" the data so all three valves would have the exact same full-open C_v rating of 50.

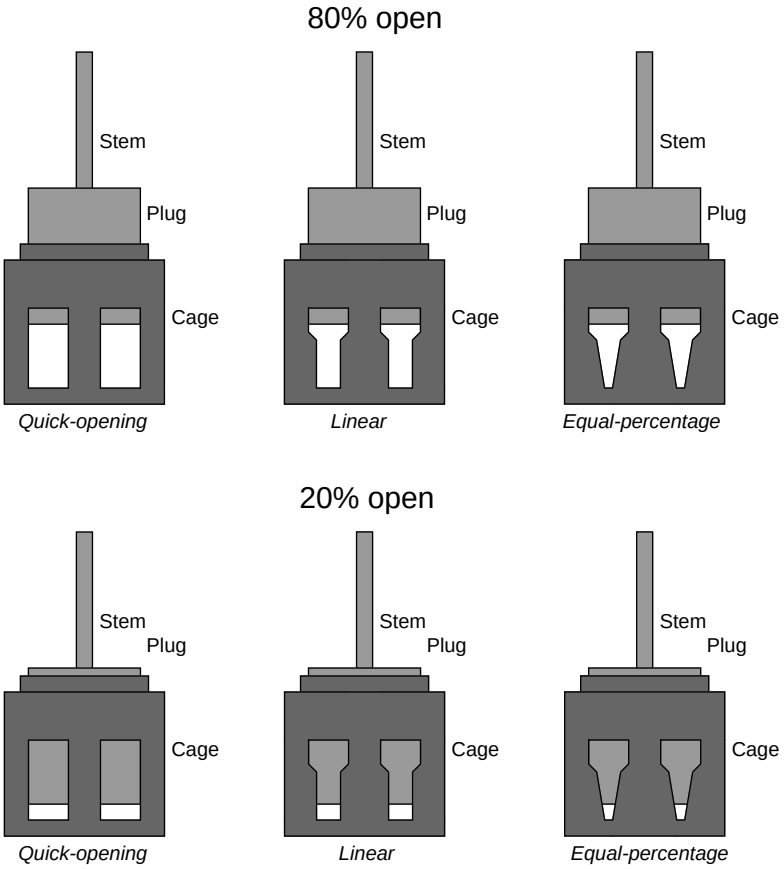
Different valve characterizations may be achieved by re-shaping the valve trim. For instance, the plug profiles of a single-ported, stem-guided globe valve may be modified to achieve the common quick-opening, linear, and equal-percentage characteristics:



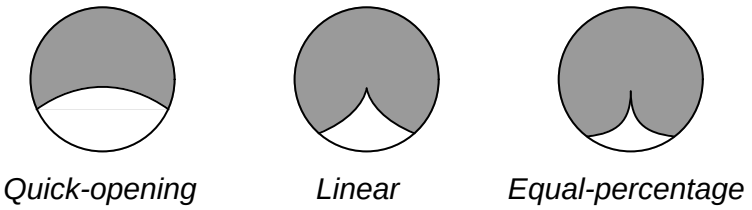
Quick-opening *Linear* *Equal-percentage*

Photographs of linear (left) and equal-percentage (right) globe valve plugs are shown side-by-side for comparison:

Cage-guided globe valve trim characteristic is a function of port shape. As the plug rises up, the amount of port area uncovered determines the shape of the characteristic graph:



Ball valve trim characteristic is a function of notch shape. As the ball rotates, the amount of notch area opened to the fluid determines the shape of the characteristic graph. All valve trim in the following illustration is shown approximately half-open (50% stem rotation):



A different approach to valve characterization is to use a non-linear positioner function instead of a non-linear trim. That is, by "programming" a valve positioner to respond in a characterized fashion to command signals, it is possible to make an inherently linear valve behave as though it

were quick-opening, equal-percentage, or anywhere in between. All the positioner does is modify the valve stem position as per the desired characteristic function instead of proportionally follow the signal as it normally would.

This approach has the distinct advantage of convenience (especially if the valve is already equipped with a positioner) over changing the actual valve trim. However, if valve stem friction ever becomes a problem, its effects will be disproportionate along the valve travel range, as the positioner must position the valve more precisely in some areas of travel than others when pressed into service as a characterizer.

25.1.14 Control valve problems

Control valves are subject to a number of common problems. This section is dedicated to an exploration of the more common control valve problems, and potential remedies.

Mechanical friction

Control valves are mechanical devices having moving parts, and as such they are subject to *friction*, primarily between the valve stem and the stem packing. Some degree of friction is inevitable in valve packing[34], and the goal is to minimize friction to a bare minimum while still maintaining a pressure-tight seal.

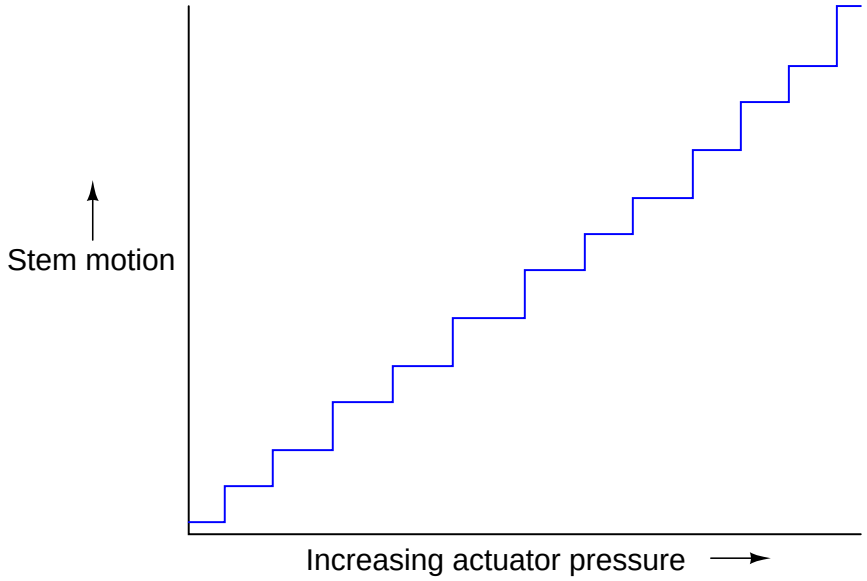
In physics, friction is classified as either *static* or *dynamic*. Static friction is defined as frictional force holding two stationary objects together. Dynamic friction is defined as frictional force impeding the motion of two objects sliding past each other. Static friction is almost always greater in magnitude than dynamic friction. Anyone who has ever pulled a sled through snow or ice knows that more force is required to "break" the sled loose from a stand-still (static friction) than is required to keep it moving (dynamic friction). The same holds true for packing friction in a control valve: the amount of force required to initially overcome static friction between the valve stem and the packing usually exceeds the amount of force required to maintain a constant speed between a moving valve stem and a stationary packing.

The presence of packing friction in a control valve increases the force necessary from the actuator to cause valve movement. If the actuator is electric or hydraulic, the only real problem with increased force is the additional energy required from the actuator to move the valve (recall that mechanical *work* is the product of force and parallel displacement). If the actuator is pneumatic, however, a more serious problem arises from the combined effects of static and dynamic friction.

A simple "thought experiment" illustrates the problem. Imagine an air-to-open, sliding-stem control valve with bench-set pressure applied to the pneumatic actuator. This should be the amount of pressure where the valve is just about to open from a fully-closed position. Now imagine slowly increasing the air pressure applied to the actuator. What should this valve do? If the spring tension is set properly, and there is negligible friction in the valve, the stem should smoothly rise from the fully-closed position as pressure increases beyond the bench-set pressure. However, what will this valve do if there is substantial friction present in the packing assembly? Instead of the stem smoothly lifting immediately as pressure exceeds the bench-set value, this valve will remain fully closed until enough *extra* pressure has accumulated in the actuator to generate a force large enough to overcome spring tension *plus* packing friction. Then, once the stem "breaks free" from static friction and begins to move, the stem will begin to accelerate because the actuator force now *exceeds* the sum of spring tension and friction, since dynamic friction is less than static friction. Compressed air trapped inside the actuator acts like a spring of its own, releasing stored energy. As the stem moves, however, the chamber volume in the diaphragm or piston actuator increases, causing pressure to drop, which causes the actuating force to decrease. When the force decreases sufficiently, the stem stops moving and static friction "grabs" it again. The stem will remain stationary until the applied pressure increases sufficiently again to overcome static friction, then the "slip-stick" cycle repeats.

[34] Bellows seals are theoretically frictionless, but in practice bellows seals are almost always combined with standard packing to prevent catastrophic blow-out in the event of the bellows rupturing, and so the theoretical advantage of low friction is never realized.

If we graph the mechanical response of a pneumatic actuator with substantial stem friction, we see something like this:



What should be a straight, smooth line is reduced to a series of "stair-steps" as the combined effect of static and dynamic friction, plus the dynamic effects of a pneumatic actuator, conspire to make precise stem positioning nearly impossible. This effect is commonly referred to as *stiction*.

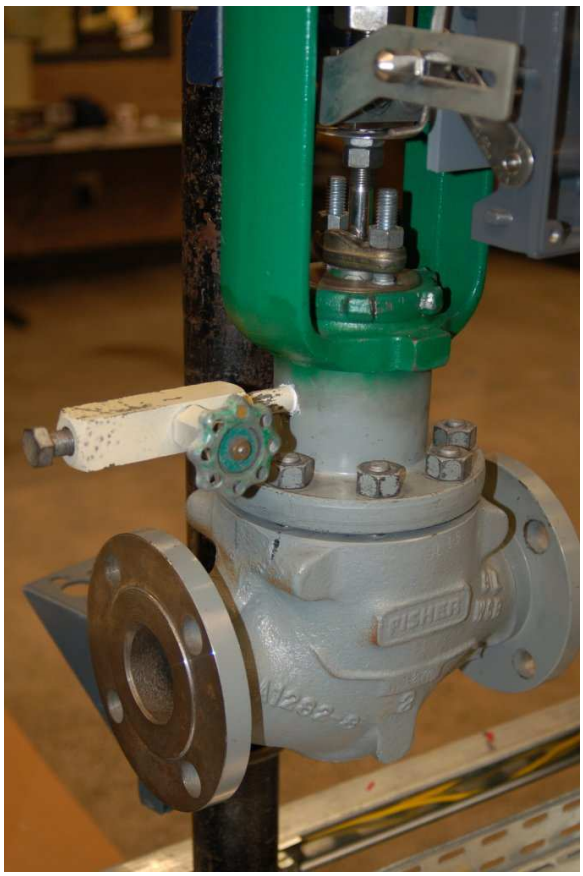
Even worse is the effect friction has on valve position when we *reverse* the direction of pressure change. Suppose that after we have reached some new valve position in the opening direction, we begin to ramp the pneumatic pressure downward. Due to static friction (again), the valve will *not* immediately respond by moving in the closed direction. Instead, it will hold still until enough pressure has been released to diminish actuator force to the point where there is enough unbalanced spring force to overcome static friction in the downward direction. Once this static friction is overcome, the stem will begin to accelerate downward because (lesser) dynamic friction will have replaced (greater) static friction. As the stem moves, however, air volume inside the actuating diaphragm or piston chamber will decrease, causing the contained air pressure to rise. Once this pressure rises enough that the stem stops moving downward, static friction will again "grab" the stem and hold it still until enough of a pressure change is applied to the actuator to overcome static friction.

What may not be immediately apparent in this second "thought experiment" is the amount of pressure change required to cause a reversal in stem motion compared to the amount of pressure change required to provoke continued stem motion in the same direction. In order to reverse the direction of stem motion, not only does the static friction have to be "relaxed" from the last movement, but additional static friction must be overcome in the opposite direction before the stem is able to move that way. To use numerical quantities, if pressure increments of 0.5 PSI are required to repeatedly overcome static friction in the upward (opening) direction, a pressure decrement of approximately twice that (1.0 PSI) will be required to make the stem go downward even just a bit.

Pressure decrements of 0.5 PSI should be sufficient to continue downward motion after the reversal, if we assume static friction to be symmetrical for the valve.

Thus, the effects of friction on a pneumatic control valve actuator are most severe (and most easily detected) by measuring the actuator pressure change required to reverse the direction of stem motion.

Short of performing a rebuild on a "sticky" control valve to replace a damaged stem and/or packing, there is not much that may be done to improve valve stiction than regular lubrication of the packing (if appropriate). Lubrication is applied to the packing by means of a special *lubricator* device threaded into the bonnet of the valve:

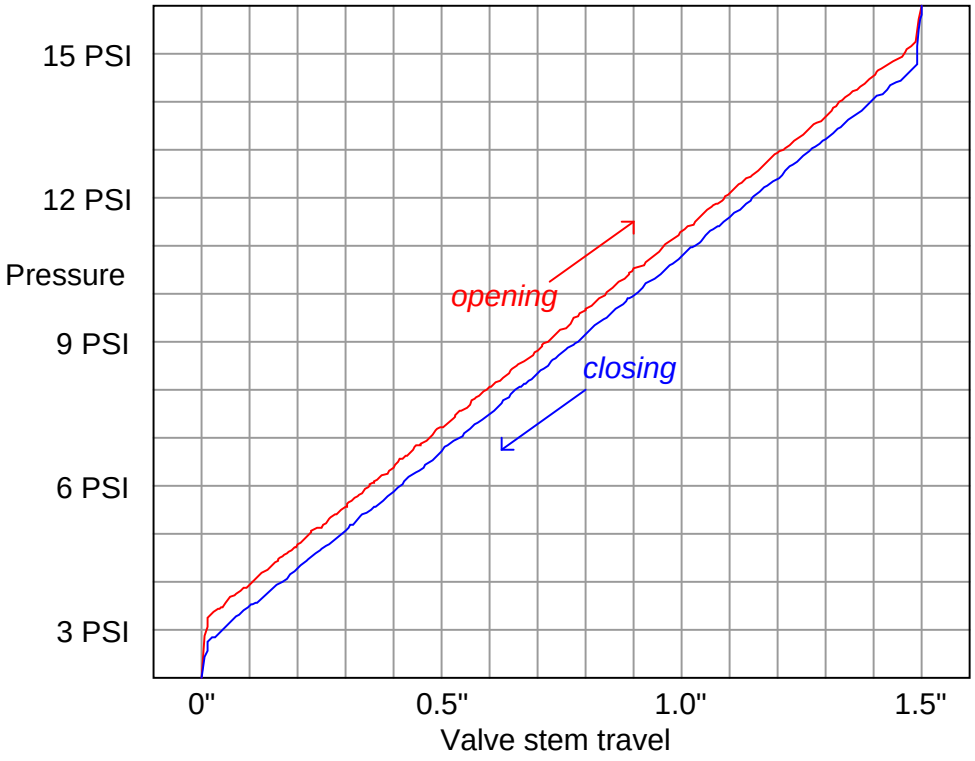


As one of the common sources of excessive packing friction is over-tightening of the packing nuts by maintenance personnel eager to prevent process fluid leaks, a great deal of trouble may be avoided simply by educating the maintenance staff as to the "care and feeding" of control valve packing for long service life.

Many modern digital valve positioners have the ability to monitor the drive force applied by an actuator on a valve stem, and correlate that force against stem motion. Consequently, it is possible

to perform highly informative diagnostic tests on a control valve's mechanical "health," at least with regard to friction[35]. For pneumatic and hydraulic actuators, actuator force is a simple and direct function of fluid pressure applied to the piston or diaphragm. For electric actuators, actuator force is an indirect function of electric motor current, or may be directly measured using load cells or springs and displacement sensors in the gear mechanism.

The following graph illustrates the kind of diagnostic "audit" that may be obtained from a digital control valve positioner based on actuator force (pneumatic air pressure) and stem motion:



This same diagnostic tool is useful for detecting trim seating problems in valve designs where there is sliding contact between the throttling element and the seat near the position of full closure (e.g. gate valves, ball valves, butterfly valves, plug valves, etc.). The force required to "seat" the valve into the fully-closed position will naturally be greater than the force required to move the throttling element during the rest of its travel, but this additional force should be smooth and consistent on the graph. A "jagged" force/travel graph near the fully-closed position indicates interference between the moving element and the stationary seat, providing information valuable for predicting the remaining service life of the valve before the next rebuild.

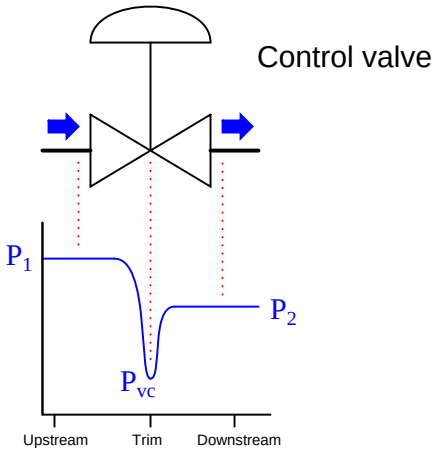
[35] Other measures of a control valve's mechanical status, such as flow capacity, flow characterization, and seat shut-off, cannot be inferred from measurements of actuator force and stem position.

Flashing

When a fluid passes through the constrictive passageways of a control valve, its average velocity increases. This is predicted by the Law of Continuity, which states that the product of fluid density (ρ), cross-sectional area of flow (A), and velocity (v) must remain constant for any flowstream:

$$\rho_1 A_1 \overline{v_1} = \rho_2 A_2 \overline{v_2}$$

This holds true for the control valves as they throttle the flow rate of a fluid by forcing it to pass through a narrow constriction. As fluid velocity increases through the constrictive passages of a control valve, the fluid molecules' kinetic energy increases. In accordance with the Law of Energy Conservation, potential energy in the form of fluid pressure must decrease correspondingly. Thus, fluid pressure decreases within the constriction of a control valve's trim as it throttles the flow, then increases (recovers) after leaving the constrictive passageways of the trim and entering the wider areas of the valve body:



If the fluid being throttled by the valve is a liquid (as opposed to a gas or vapor), and its absolute pressure ever falls below the vapor pressure of that substance, the liquid will begin to boil. This phenomenon, when it happens inside a control valve, is called *flashing*. As the graph shows, the point of lowest pressure inside the valve (called the *vena contracta* pressure, or P_{vc}) is the location where flashing will first occur, if it occurs at all.

Flashing is almost universally undesirable in control valves. The effect of boiling liquid at the point of maximum constriction is that flow through the valve becomes "choked" by the rapid expansion of liquid to vapor as it boils, severely inhibiting the total flow rate allowed through the valve. Flashing is also destructive to the valve trim, as boiling action propels tiny droplets of liquid at extremely high velocities past the plug and seat faces, eroding the metal over time.

A photograph showing a badly eroded valve plug (from a cage-guided globe valve) reveals just how destructive flashing can be:

A characteristic effect of flashing in a control valve is a "hissing" sound, reminiscent of what *sand* might sound like if it were flowing through the valve.

One of the most important performance parameters for a control valve with regard to flashing is its *pressure recovery factor*. This factor compares the valve's total pressure drop from inlet to outlet versus the pressure drop from inlet to the point of minimum pressure within the valve.

$$F_L = \sqrt{\frac{P_1 - P_2}{P_1 - P_{vc}}}$$

Where,

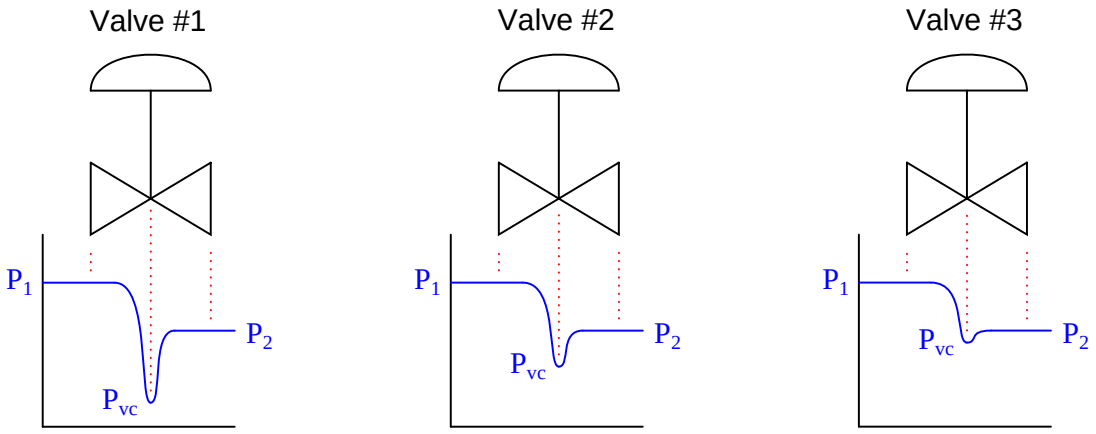
F_L = Pressure recovery factor (unitless)

P_1 = Absolute fluid pressure upstream of the valve

P_2 = Absolute fluid pressure downstream of the valve

P_{vc} = Absolute fluid pressure at the *vena contracta* (point of minimum fluid pressure within the valve)

The following set of illustrations shows three different control valves exhibiting the same permanent pressure drop ($P_1 - P_2$), but having different values of F_L:

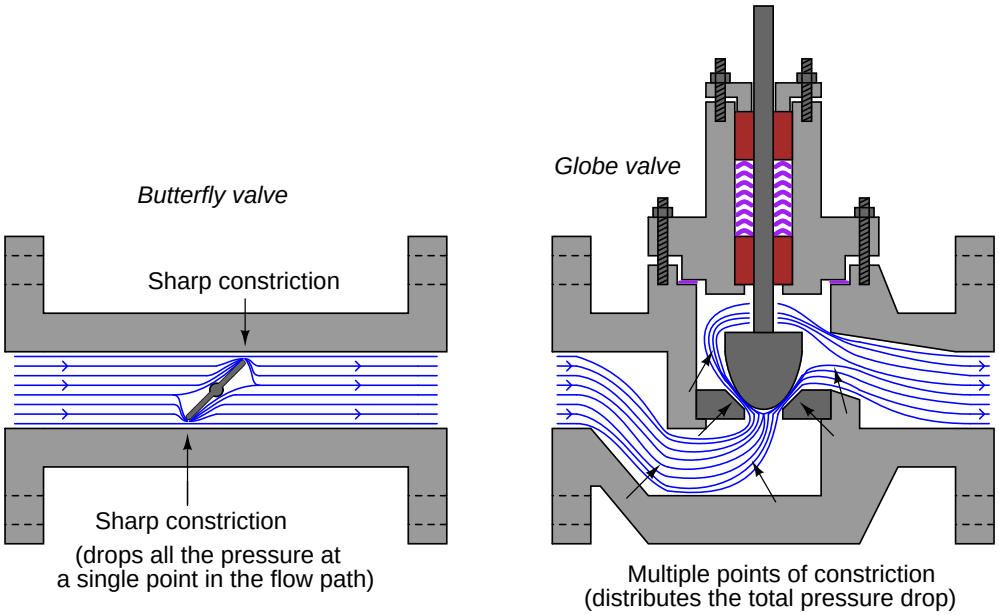


Valve #1 exhibits the greatest pressure recovery (the amount that fluid pressure *increases* from the minimum pressure at the vena contracta to the downstream pressure: $P_2 - P_{vc}$) and the lowest F_L value. It is also the valve most prone to flashing in liquid service, because the vena contracta pressure is so much lower (all other factors being equal) than in the other two valves. If any valve is going to reduce fluid pressure to the point where it spontaneously flashes to vapor, it would be valve #1.

Valve #3, by contrast, has very little pressure recovery, and an F_L value nearly equal to 1. From the perspective of avoiding flashing, it is the best of the three valves to use for liquid service.

The style of valve (ball, butterfly, globe, etc.) is very influential on pressure recovery factor. The more convoluted the path for fluid within a control valve, the more opportunities that fluid will have to dissipate energy in turbulent motion, resulting in the greatest permanent pressure drop for the least amount of restriction at any single point in the flow's path.

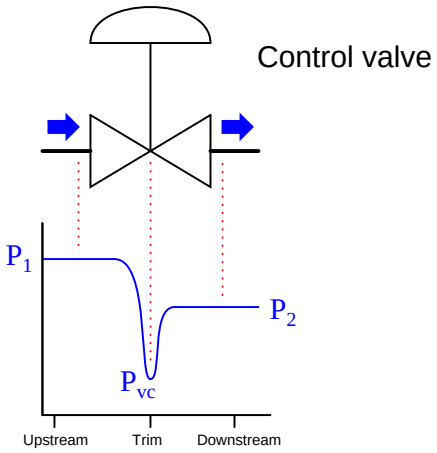
Compare these two styles of valve to see which will have lowest pressure recovery factor and therefore be most prone to flashing:



Clearly, the globe valve does a better job of evenly distributing pressure losses throughout the path of flow. By contrast, the butterfly valve can only drop pressure at the points of constriction between the disk and the valve body, because the rest of the valve body is a straight-through path for fluid offering little restriction at all. As a consequence, the butterfly valve experiences a much lower vena contracta pressure (i.e. greater pressure recovery, and a lower F_L value) than the globe valve for any given amount of permanent pressure loss, making the butterfly valve more prone to flashing than the globe valve with all other factors being equal.

Cavitation

Fluid passing through a control valve experiences changes in velocity as it enters the narrow constriction of the valve trim (increasing velocity) then enters the widening area of the valve body downstream of the trim (decreasing velocity). These changes in velocity result in the fluid molecules' kinetic energies changing as well, in accordance with the kinetic energy equation $E_k = \frac{1}{2}mv^2$. In order that energy be conserved in a moving fluid stream, any increase in kinetic energy due to increased velocity must be accompanied by a complementary decrease in potential energy, usually in the form of fluid pressure. This means the fluid's pressure will fall at the point of maximum constriction in the valve (the *vena contracta*, at the point where the trim throttles the flow) and rise again (or *recover*) downstream of the trim:



If fluid being throttled is a liquid, and the pressure at the vena contracta is less than the vapor pressure of that liquid at the flowing temperature, the liquid will spontaneously boil. This is the phenomenon of *flashing* previously described. If, however, the pressure recovers to a point greater than the vapor pressure of the liquid, the vapor will re-condense back into liquid again. This is called *cavitation*.

As destructive as flashing is to a control valve, cavitation is worse. When vapor bubbles re-condense into liquid they often do so asymmetrically, one side of the bubble collapsing before the rest of the bubble. This has the effect of translating the kinetic energy of the bubble's collapse into a high-speed "jet" of liquid in the direction of the asymmetrical collapse. These liquid "microjets" have been experimentally measured at speeds up to 100 meters per second (over 320 feet per second). What is more, the pressure applied to the surface of control valve components in the path of these microjets is immense. Each microjet strikes the valve component surface over a very small surface area, resulting in a very high pressure ($P = \frac{F}{A}$) applied to that small area. Pressure estimates as high as 1500 newtons per square millimeter (1.5 *giga*-Pascals, or about 220,000 PSI!) have been calculated for control valve applications involving water.

No substance known is able to continuously withstand this form of abuse, meaning that cavitation *will destroy* any control valve given enough time. The effect of each microjet impinging on a metal surface is to carve out a small pocket in that metal surface. Over time, the metal will begin to take on a "pock-marked" look over the area where cavitation occurs. This stands in stark contrast to the visual appearance of flashing damage, which is smooth and polished.

Photographs of a fluted valve plug and its matching seat are shown here as evidence of flashing and cavitation damage, respectively:

The plug of this valve has been severely worn by flashing and cavitation. The flashing damage is responsible for the relatively smooth wear areas seen on the plug. Cavitation damage is most prominent inside the seat, where almost all the damage is in the form of pitting. The mouth of the seat exhibits smooth wear caused by flashing, but deeper inside you can see the pock-marked surface characteristic of cavitation, where liquid microjets literally blasted away pieces of metal. This trim set belongs to a Fisher Micro-Flat Cavitation valve, designed with process liquid flow passing down instead of up (i.e. past the plug, then through the seat, rather than through the seat and up past the plug). This trim design does not prevent cavitation (as clearly evidenced by the photos), but it does "move" the area of cavitation damage down below the seat's sealing surface into a long tube extending below the seat. Although the ravages of flashing clearly took their toll on this valve's trim, the valve would have been rendered inoperable much sooner had cavitation been at work along the plug's length and at the sealing area where the plug contacts the seat.

The sound made by substantial liquid cavitation also contrasts starkly against the sound made by flashing. Whereas flashing sounds as though sand were flowing through the valve, cavitation produces a much louder "crackling" sound comprised of distinct impact pulses, reminiscent of what gravel or rocks might sound like if they were somehow forced to flow through the valve.

Sustained cavitation also has the detrimental effect of accelerating corrosion in certain process services. Bare metal surfaces are highly reactive with many chemical fluids, but become more resistant to further attack when a thin layer of reacted metal on the surface (the so-called *passivation layer*) acts as a sort of chemical barrier. Rust on steel, or the powdery-white oxide of aluminum are good examples: the initially bare metal surfaces react with their surrounding environment to form a protective outer layer, impeding further degradation of the metal beneath that layer. Cavitation works to blast away any protective layer that might otherwise accumulate, allowing corrosion to work at full speed until the entire thickness of the metal is corroded through. The complementary destructive actions of cavitation and corrosion together is sometimes referred to as *cavitation corrosion*.

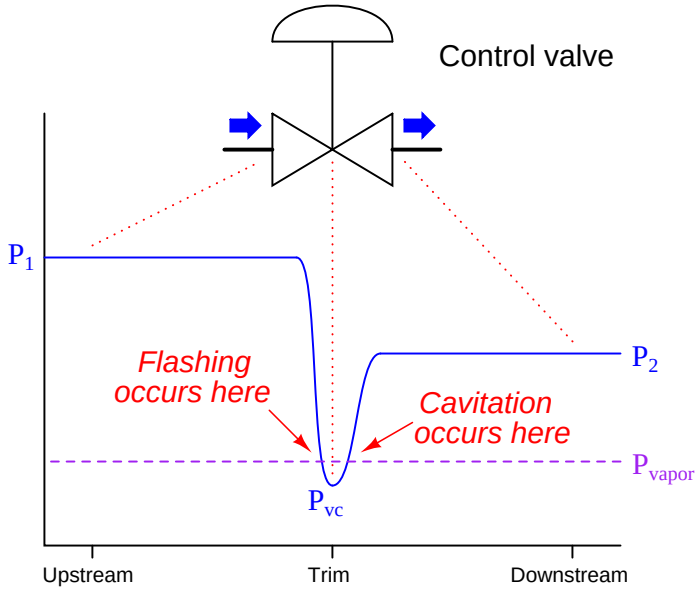
Several methods exist for abating cavitation in control valves:

1. Prevent flashing in the first place
2. Cushion with introduced gas
3. Sustain flashing action

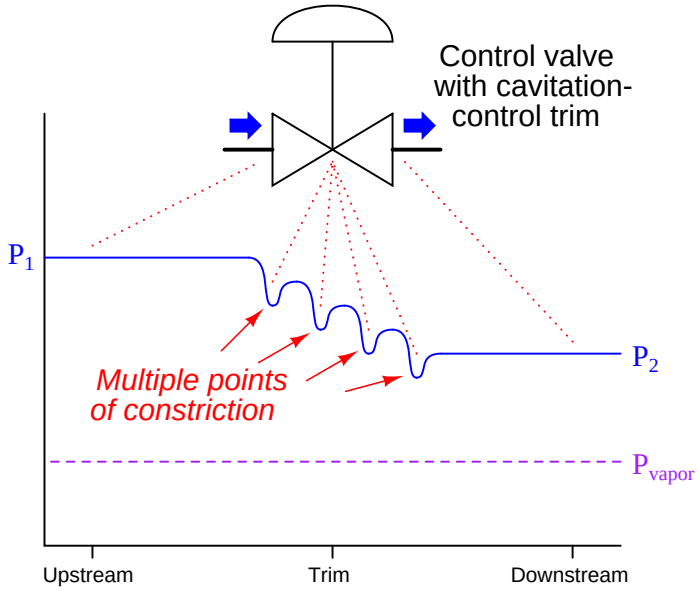
Cavitation abatement method #1 is quite simple to understand: if we prevent flashing from ever happening in a control valve, cavitation cannot follow. The key to doing this is making sure the vena contracta pressure never falls below the vapor pressure for the liquid. Several techniques exist for doing this:

- Select a control valve type having less pressure recovery (i.e. greater F_L value)
- Increase both upstream and downstream pressures by relocating the valve to a higher-pressure location in the process.
- Use multiple control valves in series to reduce the lowest pressure at either one
- Decrease the liquid's temperature (this decreases vapor pressure)
- Use cavitation-control valve trim

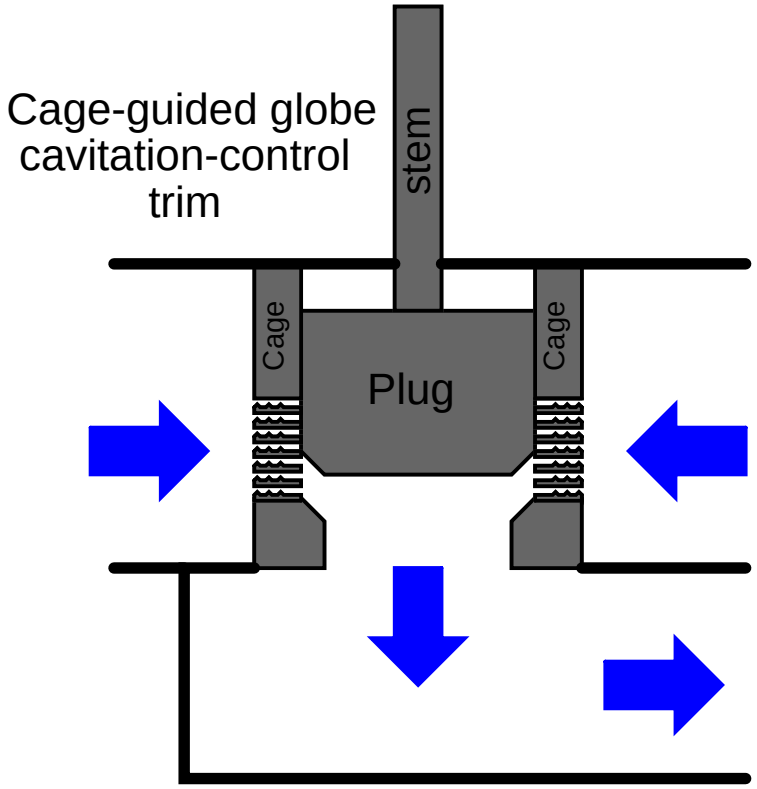
The last suggestion in this list deserves further exploration. Valve trim may be specially designed for cavitation abatement by providing multiple stages of pressure drop for the fluid as it passes through the trim. The following is a pressure versus location graph for a cavitating control valve. The liquid's vapor pressure is shown here as a dashed line marked P_{vapor}:



A valve equipped with cavitation-control trim will have a different pressure profile, with multiple *vena contracta* points where the fluid passes through a series of constrictions within the trim itself:



This way, the same final permanent pressure drop ($P_1 - P_2$) may be achieved without the lowest pressure ever falling below the liquid's vapor pressure limit. An example of cavitation-control design applied to cage-guided globe valve trim is shown in the following illustration:



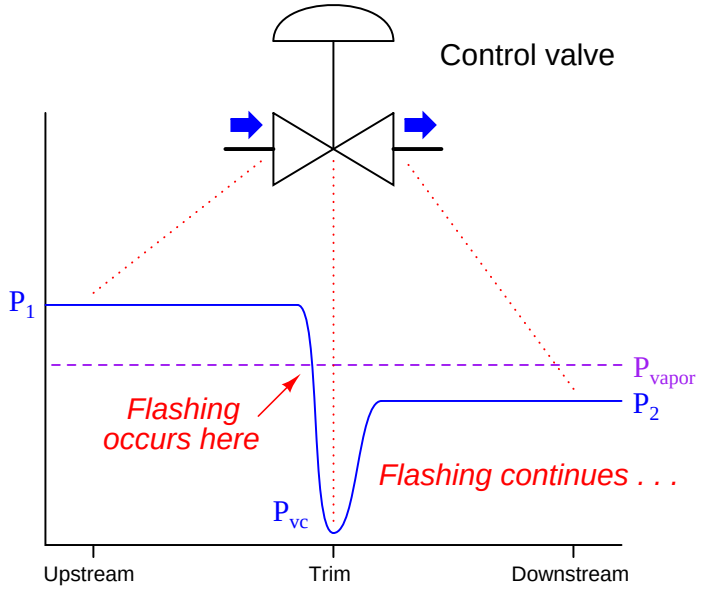
Ball-style control valves, with their relatively high pressure recovery (low pressure recovery factor F_L values) are more prone to cavitation than globe valves, all other factors being equal. Special ball trim designed to help distribute pressure drops over a longer flow path is available, an example of this shown in the next photograph:

The round (ball-shaped) portion of the trim is on the other side of this piece, with the cavitation-controlling structure visible in the foreground. Fluid flow passing through the gap between the ball's edge and the valve seat spills into this multi-chambered structure where turbulence helps develop pressure drops at several locations. In a normal ball valve, there is only one location for any substantial pressure drop to develop, and that is at the narrow gap between the ball's edge and the seat. Here, multiple regions of pressure drop exist, with the intent of avoiding the liquid's vapor pressure limit at any one location, thus eliminating flashing and consequently eliminating cavitation.

Cavitation abatement method #2 is practical only in some process applications, where a non-reacting gas may be injected into the liquid stream to provide some "cushioning" within the cavitating region. The presence of non-condensible gas bubbles in the liquid stream disturbs the microjets' pathways, helping to dissipate their energy before striking the valve body walls.

Cavitation abatement method #3 involves a strategy opposite that of method #1. If, for whatever reason, we cannot avoid falling below the vapor pressure of the liquid as the flow stream moves through the valve, we may have the option of ensuring the downstream liquid pressure never rises above the liquid's vapor pressure, at least until the fluid clears past the valuable control valve and into an area of the system where cavitation damage will not be so expensive. This avoids cavitation at the cost of guaranteed flashing within the control valve, which is generally not as destructive as cavitation.

A pressure diagram shows how this method works:



Of course, flashing is not good for a control valve either. Not only does it damage the valve over time, but it also causes problems with flow capacity, as we will explore next.

Choked flow

Both gas and liquid control valves may experience what is generally known as *choked flow*. Simply put, "choked flow" is a condition where the rate of flow through a valve does not change substantially as downstream pressure is reduced.

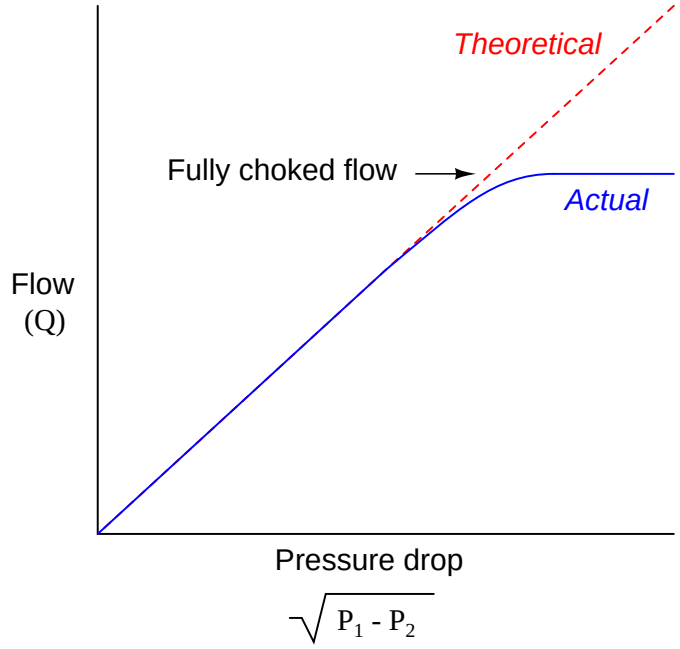
Ideally, turbulent fluid flow rate through a control valve is a simple function of valve flow capacity (C_v) and differential pressure drop ($P_1 - P_2$), as described by the basic valve flow equation:

$$Q = C_v \sqrt{\frac{P_1 - P_2}{G_f}}$$

In a gas control valve, choking occurs when the velocity of the gas reaches the speed of sound for that gas. This is often referred to as *critical* or *sonic* flow. In a liquid control valve, choking

occurs with the onset of flashing[36]. The reason sonic velocity is relevant to flow capacity for a control valve has to do with the propagation of pressure changes in fluids. Pascal's Law tells us that changes in pressure within a closed fluid system will manifest at all points in the fluid system, but this never happens instantaneously. Instead, pressure changes propagate through any fluid at the speed of sound within that fluid. If a fluid stream happens to move at or above the speed of sound, pressure changes downstream are simply not able to overcome the stream's velocity to affect anything upstream, which explains why the flow rate through a control valve experiencing sonic (critical) flow velocities does not change with changes in downstream pressure: those downstream pressure changes cannot propagate upstream against the fast-moving flow, and so will have no effect on the flow as it accelerates to sonic velocity at the point(s) of constriction.

Choked flow conditions become readily apparent if the flow-versus-pressure function of a control valve at any fixed opening value is graphed. The basic valve flow equation predicts a perfectly straight line at constant slope with flow rate (Q) as the vertical variable and the square root of pressure drop ($\sqrt{P_1 - P_2}$) as the horizontal variable. However, if we actually test a control valve by holding its upstream liquid pressure (P_1) constant and varying its downstream pressure (P_2) while maintaining a fixed stem position, we notice a point where flow reaches a maximum limit value:



In a choked flow condition, further reductions in downstream pressure achieve no greater flow of liquid through the valve. This is not to say that the valve has reached a maximum flow – we

[36] The *Control Valve Sourcebook – Power & Severe Service* on page 6-3 and the *ISA Handbook of Control Valves* on page 211 both suggest that the mechanism for choking in liquid service may be related to the speed of sound just as it is for choked flow in gas services. Normally, liquids have higher sonic velocities than gases due to their far greater densities. This makes choking due to sonic velocity very unlikely in liquid flowstreams. However, when a liquid flashes into vapor, the speed of sound *for that two-phase mixture of liquid and vapor* will be much less than it is for the liquid itself, opening up the possibility of sonic velocity choking.

may still increase flow rate through a choked valve by increasing its upstream pressure. We simply cannot coax more flow through a choked valve by decreasing its downstream pressure.

An approximate predictor of choked flow conditions for gas valve service is the upstream-to-minimum absolute pressure ratio. When the vena contracta pressure is less than one-half the upstream pressure, both measured in absolute pressure units, choked flow is virtually guaranteed. One should bear in mind that this is merely an approximation and not a precise prediction for choked flow. A lot more information would have to be known about the valve design, the particular process gas, and other factors in order to more precisely predict the presence of choking.

Choked flow in liquid services is predicted when the vena contracta pressure equals the liquid's vapor pressure, since choking is a function of flashing for liquid flowstreams.

No attempt will be made in this book to explain sizing procedures for control valves in choked-flow service, due to the complexity of the subject.

An interesting and beneficial application of choked flow in gases is a device called a *critical velocity nozzle*. This is a nozzle designed to allow a fixed flow rate of gas through it given a known upstream pressure, and a downstream pressure that is sufficiently low to allow sonic velocities in the nozzle throat. One practical use for critical velocity nozzles is in the flow testing of compressed air systems. One or more of these nozzles are connected to the main header line of an air compressor system and allowed to vent to atmosphere. So long as the compressor(s) are able to maintain constant header pressure, the flow rate of air through the nozzles(s) is guaranteed to be fixed, allowing a technician to monitor compressor parameters under precisely known load conditions.

Valve noise

A troublesome phenomenon in severe services is audible noise produced by control valves. Noise output is worse for gas services experiencing sonic (critical) flow and for liquid services experiencing cavitation, although it is possible for a control valve to produce substantial noise even when avoiding these operating conditions.

One way to reduce noise output is to use special valve trim resembling the trim used to mitigate cavitation. A common cage-guided globe valve trim design for noise reduction uses a special cage designed with numerous, small holes for process gas to flow through. These small holes do not in themselves reduce aerodynamic noise, but they do shift the *frequency* of that noise up. This increase in frequency places the sound outside the range where the human ear is most sensitive to noise, and it also helps to reduce noise coupling to the piping, confining most of the noise "power" to the internal volume of the process fluid rather than radiating outward into the air.

Fisher manufactures a series of noise-abatement trim for process gas service called *Whisper* trim. A "Whisper" plug and cage set is shown in this next photograph:

In some versions, the holes are merely straight through the cage wall. In more sophisticated versions of Whisper trim (particularly the "WhisperFlo"), the small holes lead to a labyrinth of passages designed to dissipate energy by forcing the fluid to take several sharp turns as it passes through the wall of the cage. This allows a pressure drop to develop across the valve without necessarily generating high fluid velocities, which is the primary causal factor for noise in control valves.

Erosion

A problem common to control valves used in *slurry* service (where the process fluid is a liquid containing a substantial quantity of hard, solid particles) is *erosion*, where the valve trim and body are worn by the passage of solid particles. Erosion produces some of the most striking examples of valve damage, as shown by the following photographs[37]:

Here we see large holes worn in a globe valve plug, and substantial damage done to the seat as well. The process service in this case was water with "coke fines" (small particles of coke, a solid petroleum product). Even ceramic valve seat components are not immune to damage from slurry service, as revealed by this photograph of a valve seat with a notch worn by slurry flow:

There really is no good way to reduce the effects of erosion damage from slurry flows, other than to use exceptionally hard valve trim materials. Even then, the control valve must be considered a fast-wearing component (along with pumps and any other components in harm's way of the slurry flowstream), rebuilt or replaced at regular intervals.

[37] A colleague of mine humorously refers to these valve trim samples as "shock and awe," because they so dramatically reveal the damaging nature of certain process fluid services.

Another cause of erosion in control valves is *wet steam*, where steam contains droplets of liquid water propelled at high velocity by the steam flow. A dramatic example of wet steam damage appears in this next photograph, where the cage from a Fisher valve has been literally cut in half from the flow:

Steam may also "cut" other parts of a valve if allowed to leak past. Here, we see a valve bonnet with considerable damage caused by steam leaking past the *outside* of the packing, between the packing rings and the bonnet's bore:

Any fluid with sufficient velocity may cause extensive damage to valve components. Small holes developing in the body of a valve may become large holes over time, as fluid rushes through the hole. The initial cause of the hole may be a manufacturing defect (such as *porosity* in the metal casting) or damage inflicted by the user (e.g. a crack in the valve body caused by some traumatic event). An example of such a hole in a valve body becoming worse over time appears in this next photograph, taken of a control valve removed after *40 years* of continuous service:

A close-up photograph of this same hole shows the leak path worn larger by passing fluid, allowing fluid to flow by the seat even with the valve in the fully-closed position:

In more severe process services, such holes rapidly grow in size. This next photograph shows a rather extreme example of a hole near the seat of a valve body, enlargened to the point where the valve is hardly capable of restricting fluid flow at all because the hole provides a bypass route for flow around the valve plug and seat:

Chemical attack

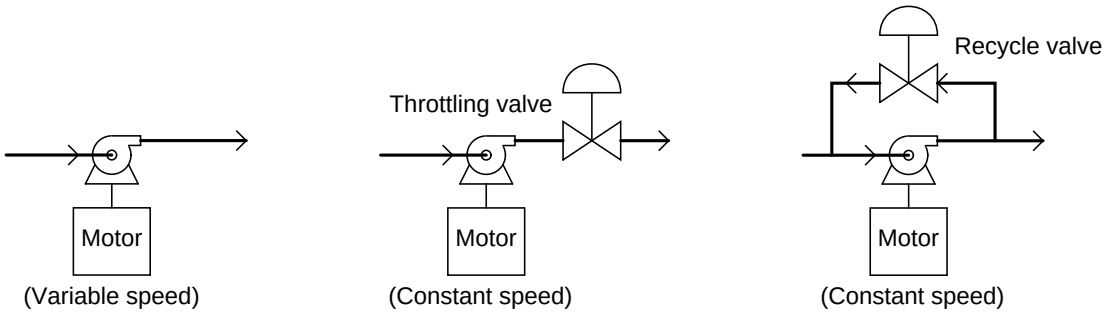
Corrosive chemicals may attack the metal components of control valves if those components are not carefully selected for the proper service. A close-up photograph of a chemically-pitted valve plug shows pitting characteristic of chemical attack:

As mentioned previously in this chapter, the effects of corrosion are multiplied when combined with the effects of cavitation. Most metals develop what is known as a *passivation layer* in response to chemical attack. The outer layer of metal corrodes, but the byproduct of that corrosion is a relatively inert compound acting to shield the rest of the metal from further attack. Rust on steel, or aluminum oxide on aluminum, are both common examples of passivation layers in response to oxidation of the metal. When cavitation happens inside a valve, however, the extremely high pressures caused by the liquid microjets will blast away any protection afforded by the passivation layer, allowing chemical attack to begin anew. The result is rapid degradation of the valve components.

25.2 Variable-speed motor controls

An alternative method of flow control in lieu of control valves is to vary the speed of the machine(s) motivating fluid to flow. In the case of liquid flow control, this would take the form of variable-speed pumps. In the case of gas flow control, it would mean varying the speed of compressors or blowers.

Flow control by machine speed control makes a lot of sense for some process applications. It is certainly more energy-efficient to vary the speed of the machine pushing fluid to control flow, as opposed to letting the machine run at full speed all the time and adjusting flow rate by throttling the machine's discharge (outlet) or recycling fluid back to the machine's suction (inlet). The fact that the system has one less component in it (no control valve) also reduces capital investment and potentially increases system reliability:



Further advantages of machine speed control include the ability to "soft-start" the machine instead of always accelerating rapidly from a full stop to full speed, reduced wear on machines due to less motion over time, and reduced vibration. In applications such as conveyor belt control, robotic machine motion control, and electric vehicle propulsion, variable-speed technology makes perfect sense because the prime mover device is already (in most cases) an electric motor, with precise speed control of that motor providing many practical benefits. In some applications, *regenerative braking* may be of benefit, where the motor is used as an electrical generator to slow down the machine on command. Regenerative braking transfers kinetic energy within the machine to the power grid where it may be gainfully used in other processes, saving energy and reducing wear on any mechanical (friction) brakes already installed in the machine.

With all these advantages inherent to variable-speed pumps, fans, and compressors (as opposed to using dissipative control valves), one might wonder, "Why would anyone *ever* use a control valve to regulate flow? Why not just control all fluid flows using variable-speed pumping machines?" Several good answers exist to this question:

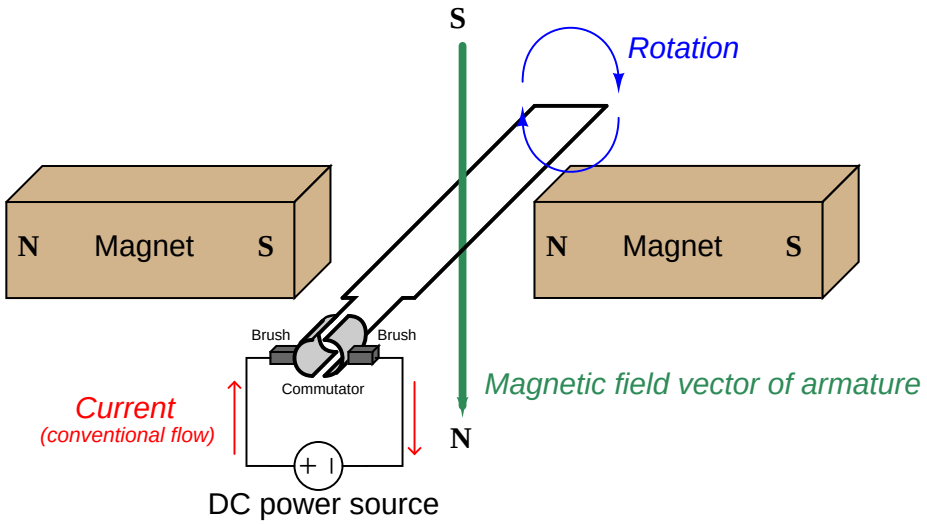
- Variable-speed machines often cannot respond as rapidly as control valves
- Control valves have the ability to positively halt flow; a stopped pump or blower will not necessarily prevent flow from going through
- Some process applications *must* contain a dissipative element in order for the system to function (e.g. let-down valves in closed refrigeration systems)
- Split-ranging may be difficult or impossible to achieve with multiple machine speed control

- Limited options for fail-safe status
- In many cases, there is no machine dedicated to a particular flow path (e.g. a pressure release valve)

25.2.1 DC motor speed control

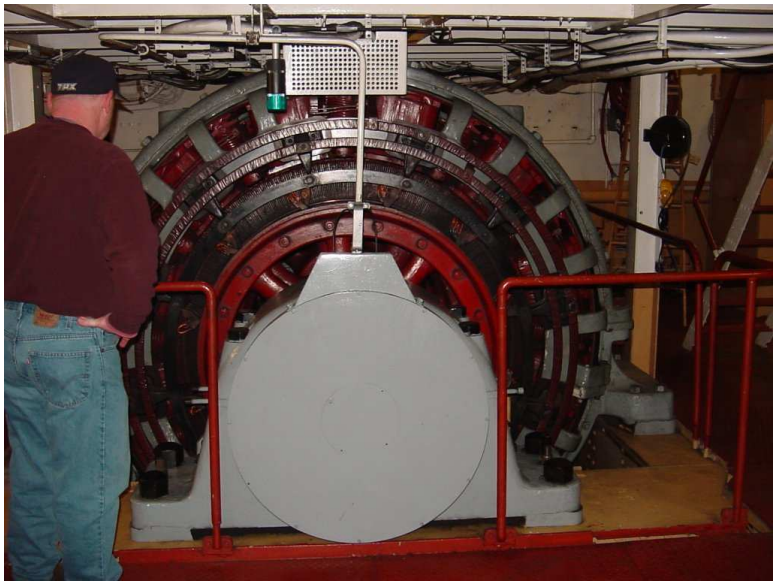
DC electric motors generate torque by a reaction between two magnetic fields: one field established by stationary "field" windings (coils), and the other by windings in the rotating armature. Some DC motors lack field windings, substituting large permanent magnets in their place so that the stationary magnetic field is constant for all operating conditions.

In any case, the operating principle of a DC electric motor is that current passed through the armature creates a magnetic field that tries to align with the stationary magnetic field. This causes the armature to rotate:

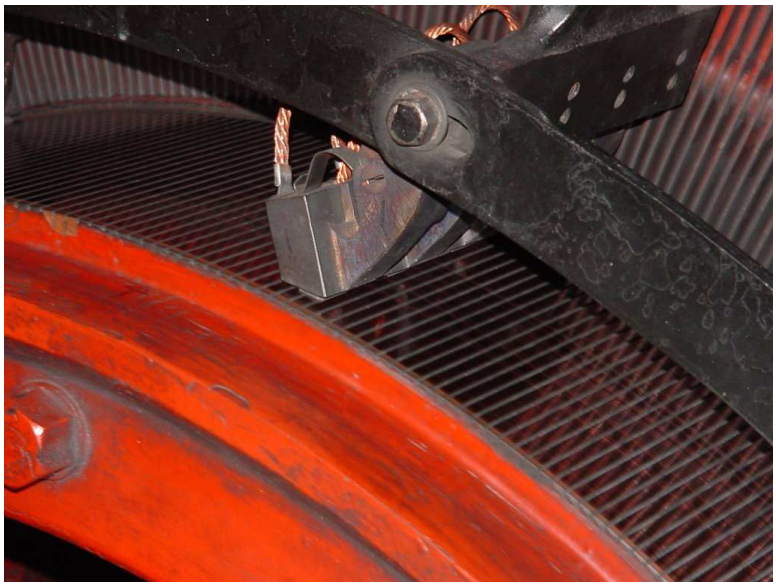


However, a set of segmented copper strips called a *commutator* breaks electrical contact with the now-aligned coil and energizes another coil (or in the simple example shown above, it re-energizes the same loop of wire in the opposite direction) to create another out-of-alignment magnetic field that continues to rotate the armature. Electrical contact between the rotating commutator segments and the stationary power source is made through carbon *brushes*. These brushes wear over time (as does the commutator itself), and must be periodically replaced.

Most industrial DC motors are built with multiple armature coils, not just one as shown in the simplified illustration above. A photograph of a large (1250 horsepower) DC motor used to propel a ferry ship is shown here, with the field and armature poles clearly seen (appearing much like spokes in a wheel):



A close-up of one brush assembly on this large motor shows both the carbon brush, the brush's spring-loaded holder, and the myriad of commutator bars the brush makes contact with as the armature rotates:



DC motors exhibit the following relationships between mechanical and electrical quantities:

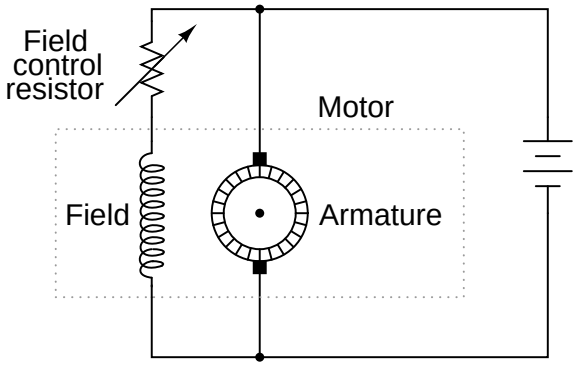
Torque:

- Torque is directly proportional to armature magnetic field strength, which in turn is directly proportional to current through the armature windings
- Torque is also directly proportional to the stationary pole magnetic field strength, which in turn is directly proportional to current through the field windings (in a motor with non-permanent field magnets)

Speed:

- Speed is limited by the counter-EMF generated by the armature as it spins through the stationary magnetic field. This counter-EMF is directly proportional to armature speed, and also directly proportional to stationary pole magnetic field strength (which is directly proportional to field winding current in a motor that is not permanent-magnet)
- Thus, speed is directly proportional to armature voltage
- Speed is also inversely proportional to stationary magnetic field strength, which is directly proportional to current through the field windings (in a motor with non-permanent field magnets)

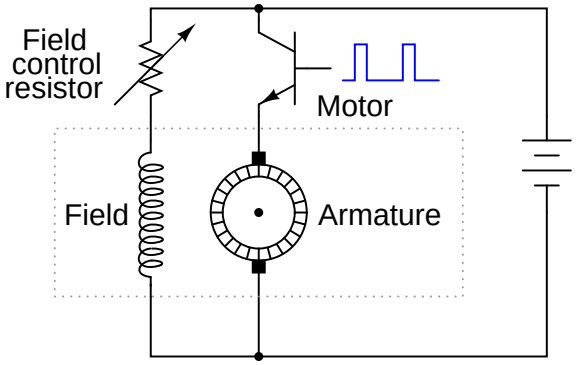
A very simple method for controlling the speed and torque characteristics of a wound-field (non-permanent magnet) DC motor is to control the amount of current through the field winding:



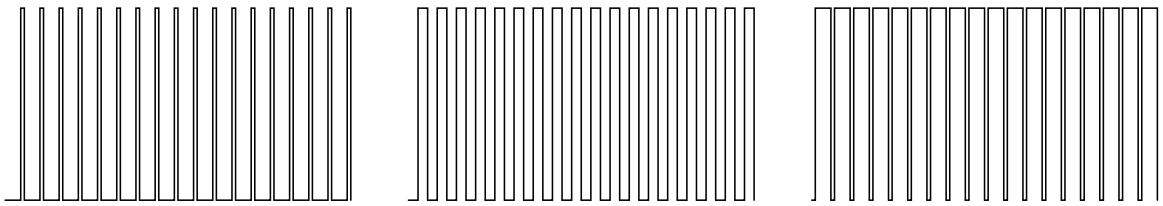
Decreasing the field control resistor's resistance allows more current through the field winding, strengthening its magnetic field. This will have two effects on the motor's operation: first, the motor will generate more torque than it did before (for the same amount of armature current) because there is now a stronger magnetic field for the armature to react against; second, the motor's speed will decrease because more counter-EMF will be generated by the spinning armature for the same rotational speed, and this counter-EMF naturally attempts to equalize with the applied DC source voltage. Conversely, we may increase a DC motor's speed (and reduce its torque output) by increasing the field control resistor's resistance, weakening the stationary magnetic field through which the armature spins.

Regulating field current may alter the balance between speed and torque, but it does little to control total motor *power*. In order to control the power output of a DC motor, we must also regulate armature voltage and current. Variable resistors may also be used for this task, but this is generally frowned upon in modern times because of the wasted power.

A better solution is to have an electronic power control circuit very rapidly switch transistors on and off, switching power to the motor armature. This is called *pulse-width modulation*, or *PWM*.

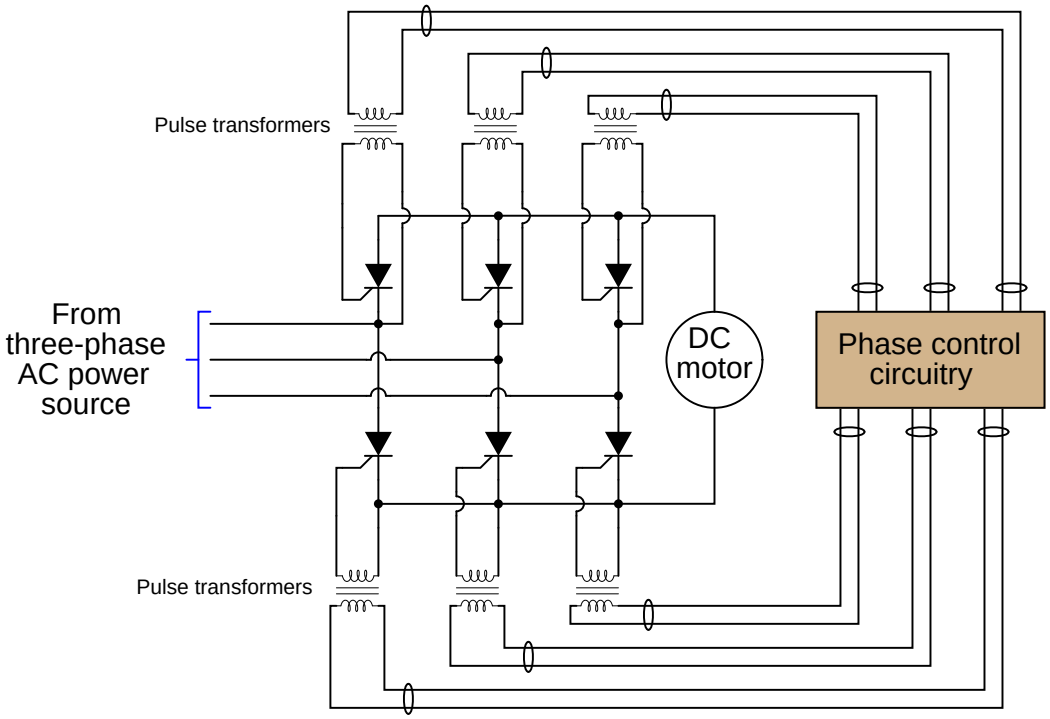


The *duty cycle* (on time versus on+off time) of the pulse waveform will determine the fraction of total power delivered to the motor:



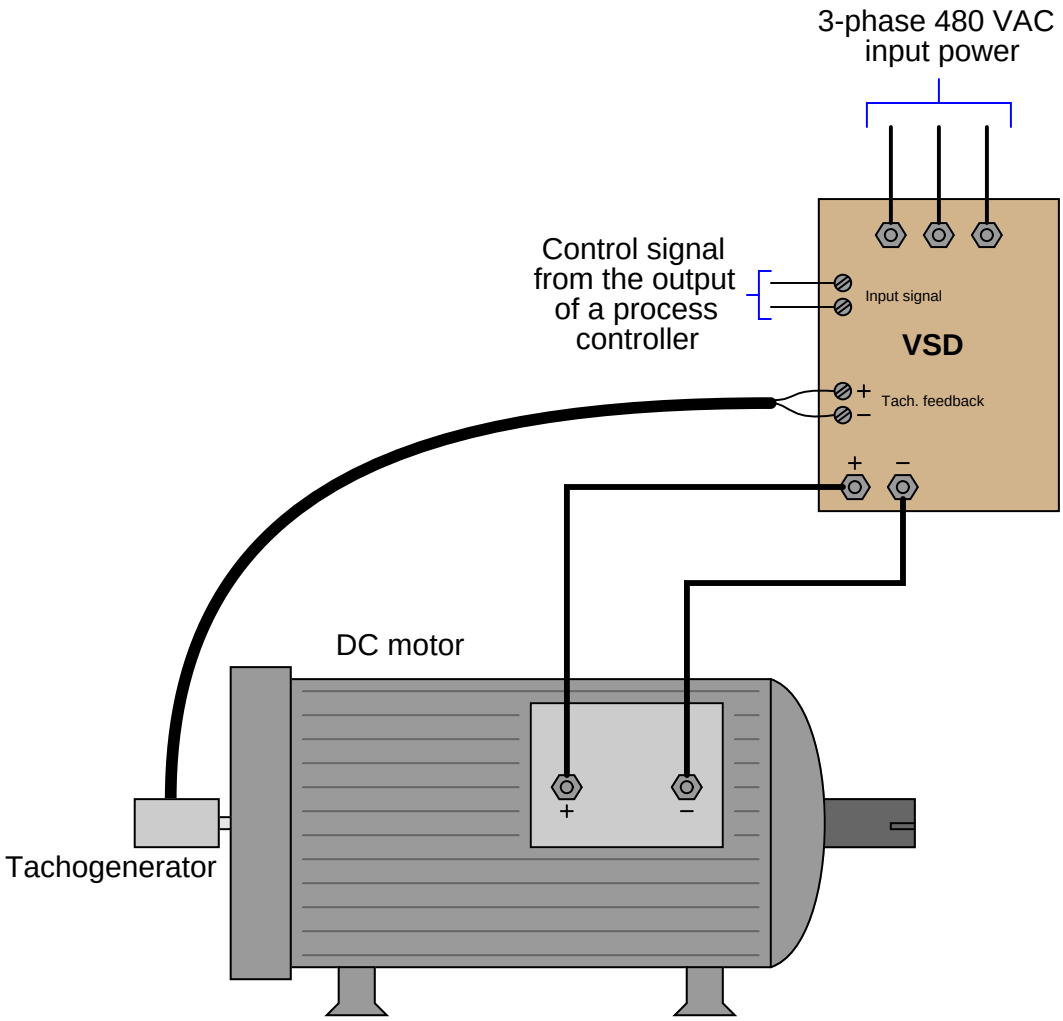
Such an electronic power-control circuit is generally referred to as a *drive*. Thus, a *variable-speed drive* or *VSD* is a high-power circuit used to control the speed of a DC motor. Motor drives may be manually set to run a motor at a set speed, or accept an electronic control signal to vary the motor speed in the same manner an electronic signal commands a control valve to move. When equipped with remote control signaling, a motor drive functions just like any other final control element: following the command of a process controller in order to stabilize some process variable at setpoint.

An older technology for pulsing power to a DC motor is to use a *controlled rectifier* circuit, using SCRs instead of regular rectifying diodes to convert AC to DC. Since the main power source of most industrial DC motors is AC anyway, and that AC must be converted into DC at some point in the system, it makes sense to integrate control right at the point of rectification:



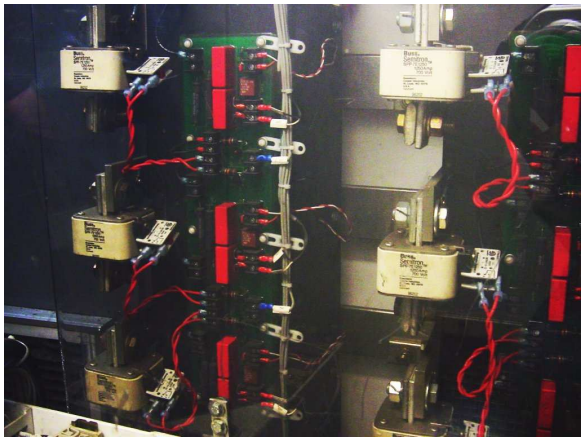
Controlled rectifier circuits work on the principle of varying the "trigger" pulse times relative to the AC waveform pulses. The earlier the AC cycle each SCR is triggered on, the longer it will be on to pass current to the motor. The "phase control" circuitry handles all this pulse timing and generation.

A DC motor drive that simply varied power to the motor according to a control signal would be crude and difficult to apply to the control of most processes. What is ideally desired from a variable-speed drive is precise command over the motor's *speed*. For this reason, most VSDs are designed to receive feedback from a tachometer mechanically connected to the motor shaft, so the VSD "knows" how fast the motor is turning. The tachometer is typically a small DC generator, producing a DC voltage directly proportional to its shaft speed (0 to 10 volts is a common scale). With this information, the VSD may throttle electrical power to the motor as necessary to achieve whatever speed is being commanded by the control signal. Having a speed-control feedback loop built into the drive makes the VSD a "slave controller" in a cascade control system, the drive receiving a speed setpoint signal from whatever process controller is sending an output signal to it:



A photograph of the tachogenerators (dual, for redundancy) mechanically coupled to that large 1250 horsepower ferry ship propulsion motor appears here:

The SCRs switching power to this motor may be seen here, connected via twisted-pair wires to control boards issuing "firing" pulses to each SCR at the appropriate times:



The integrity of the tachogenerator feedback signal to the VSD is extremely important for safety reasons. If the tachogenerator becomes disconnected – whether mechanically or electrically (it doesn't matter) – from the drive, the drive will "think" the motor is not turning. In its capacity as a speed *controller*, the drive will then send full power to the DC motor in an attempt to get it up to speed. Thus, loss of tachogenerator feedback causes the motor to immediately "run away" to full speed. This is undesirable at best, and likely dangerous in the case of motors as large as the one powering this ship.

As with all forms of electric power control based on pulse durations and duty cycles, there is a lot of electrical "noise" cast by VSD circuits. Square-edged pulse waveforms created by the rapid on-and-off switching of the semiconductor power devices are equivalent to infinite series of high-

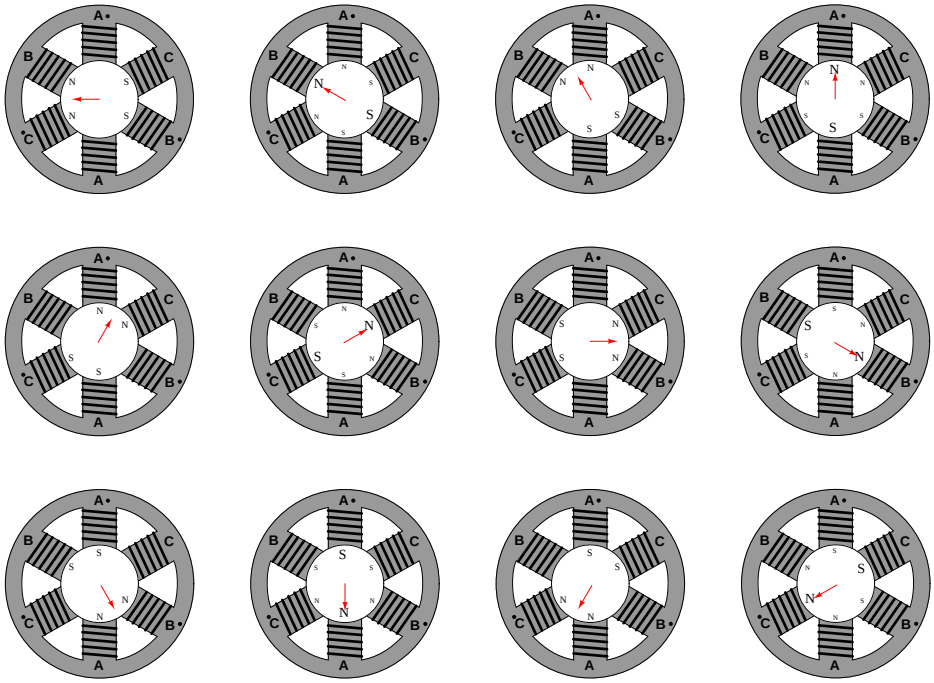
frequency sine waves[38], some of which may be of high enough frequency to self-propagate through space as electromagnetic waves. This *radio-frequency interference* or *RFI* may be quite severe given the high power levels of industrial motor drive circuits. For this reason, it is *imperative* that neither the motor power conductors nor the conductors feeding AC power to the drive circuit be routed anywhere near small-signal or control wiring, because the induced noise *will* wreak havoc with whatever systems utilize those low-level signals.

RFI noise on the AC power conductors may be mitigated by routing the AC power through *filter* circuits placed near the drive. The filter circuits block high-frequency noise from propagating back to the rest of the AC power distribution wiring where it may influence other electronic equipment. However, there is little that may be done about the RFI noise between the drive and the motor other than to shield the conductors in well-grounded metallic conduit.

[38]This equivalence was mathematically proven by Jean Baptiste Joseph Fourier (1768-1830), and is known as a *Fourier series*.

25.2.2 AC motor speed control

AC induction motors are based on the principle of a *rotating magnetic field* produced by a set of stationary windings (called *stator* windings) energized by AC power of different phases. The effect is not unlike a series of blinking light bulbs which appear to "move" in one direction due to intentional sequencing of the blinking. If sets of wire coils (windings) are energized in a like manner – each coil reaching its peak field strength at a different time from its adjacent neighbor – the effect will be a magnetic field that "appears" to move in one direction. If these windings are oriented around the circumference of a circle, the moving magnetic field rotates about the center of the circle, as illustrated by this sequence of images (read left-to-right, top-to-bottom, as if you were reading words in a sentence):



Any magnetized object placed in the center of this circle will attempt to spin at the same rotational speed as the rotating magnetic field. *Synchronous* AC motors use this principle, where a magnetized rotor precisely follows the magnetic field's speed.

Any electrically conductive object placed in the center of the circle will experience *induction* as the magnetic field direction changes around the conductor. This will induce electric currents within the conductive object, which in turn will react against the rotating magnetic field in such a way that the object will be "dragged along" by the field, always lagging a bit in speed. *Induction* AC motors use this principle, where a non-magnetized (but electrically conductive) rotor rotates at a speed slightly less[39] than the synchronous speed of the rotating magnetic field.

[39] The difference between the synchronous speed and the rotor's actual speed is called the motor's *slip speed*.

The rotational speed of this magnetic field is directly proportional to the frequency of the AC power, and inversely proportional to the number of poles in the stator:

$$S = \frac{120f}{n}$$

Where,

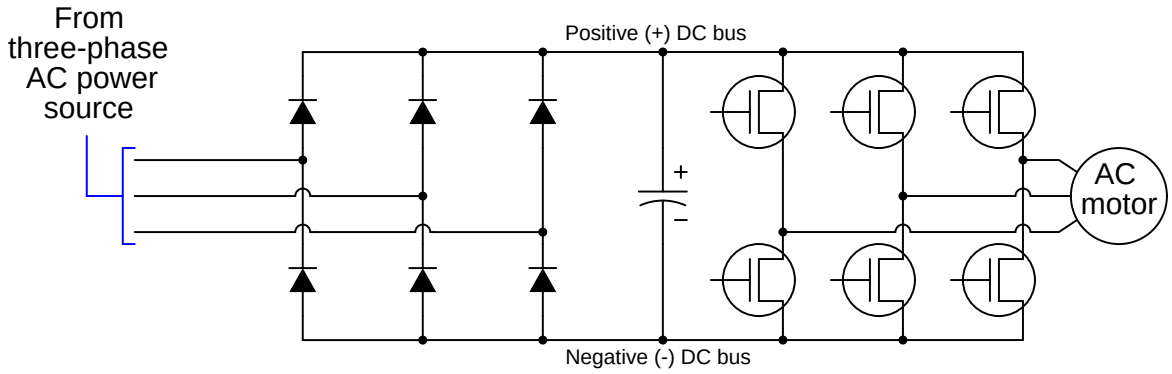
S = Synchronous speed of rotating magnetic field, in revolutions per minute (RPM)

f = Frequency, in cycles per second (Hz)

n = Total number of stator poles per phase (the simplest possible AC motor design will have 2 poles)

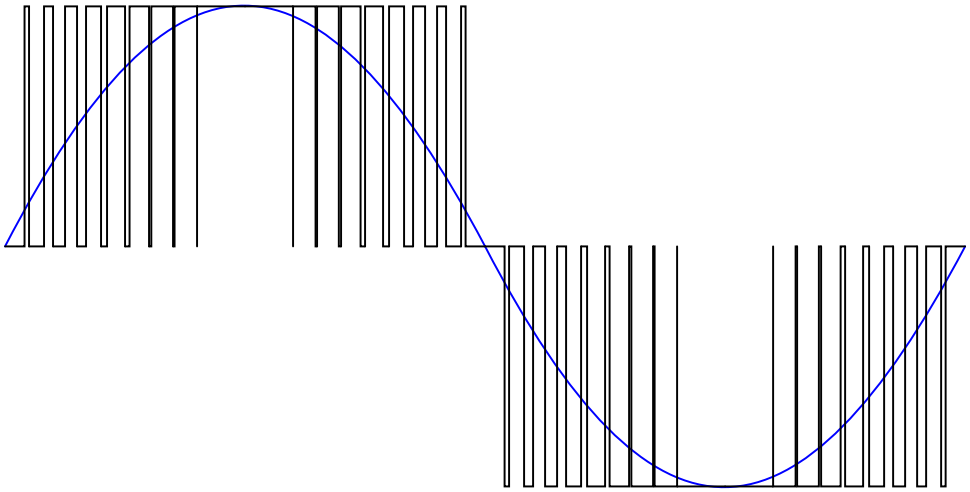
While the number of poles in the motor's stator is a quantity fixed at the time of the motor's manufacture, the frequency of power we apply may be adjusted with the proper electronic circuitry. A high-power circuit designed to produce varying frequencies for an AC motor to run on is called a *variable-frequency drive*, or *VFD*.

A simplified schematic diagram for a VFD is shown here, with a rectifier section on the left (to convert AC input power into DC), a filter capacitor to "smooth" the rectified DC power, and a transistor "bridge" to switch DC into AC at whatever frequency is desired to power the motor. The transistor control circuitry has been omitted from this diagram for the sake of simplicity:



As with DC motor drives (VSDs), the power transistors in an AC drive (VFD) switch on and off very rapidly with a varying duty cycle. Unlike DC drives, however, the duty cycle of an AC drive's power transistors must vary rapidly in order to synthesize an AC waveform from the DC "bus" voltage following the rectifier. A DC drive circuit's PWM duty cycle controls motor power, and so it will remain at a constant value when the desired motor power is constant. Not so for an AC motor drive circuit: its duty cycle must vary from zero to maximum and back to zero repeatedly in order to generate an AC waveform for the motor to run on.

The equivalence between a rapidly-varied pulse-width modulation (PWM) waveform and a sine wave is shown in the following illustration:



This concept of rapid PWM transistor switching allows the drive to "carve" any arbitrary waveform out of the filtered DC voltage it receives from the rectifier. Virtually any frequency may be synthesized (up to a maximum limited by the frequency of the PWM pulsing), and any voltage (up to a maximum peak established by the DC bus voltage), giving the VFD the ability to power an induction motor over a wide range of speeds.

While frequency control is the key to synchronous and induction AC motor speed control, it is generally not enough on its own. While the speed of an AC motor is a direct function of frequency (controlling how fast the rotating magnetic field rotates around the circumference of the stator), torque is approximately proportional to stator winding current. Since the stator windings are inductors by nature, their reactance varies with frequency as described by the formula $X_L = 2\pi fL$. Thus, as frequency is increased, winding reactance increases right along with it. This increase in reactance results in a decreased stator current (assuming the AC voltage is held constant as frequency is increased). This can cause undue torque loss at high speeds, and excessive torque (as well as excessive stator heat!) at low speeds. For this reason, the AC voltage applied to a motor by a VFD is usually made to vary in direct proportion to the applied frequency, so that the stator current will remain within good operating limits throughout the speed range of the VFD. This correspondence is called the *voltage-to-frequency ratio*, or V/F ratio.

Variable-frequency motor drives are manufactured for industrial motor control in a wide range of sizes and horsepower capabilities. Some VFDs are small enough to hold in your hand, while others are large enough to require a freight train for transport. The following photograph shows a pair of moderately-sized Allen-Bradley VFDs (about 100 horsepower each, standing about 4 feet high), used to control pumps at a wastewater treatment plant:

Variable-frequency AC motor drives do not require motor speed feedback the way variable-speed DC motor drives do. The reason for this is quite simple: the controlled variable in an AC drive is the frequency of power sent to the motor, and rotating-magnetic-field AC motors are *frequency-controlled* machines by their very nature. For example, a 4-pole AC induction motor powered by 60 Hz has a base speed of 1728 RPM (assuming 4% slip). If a VFD sends 30 Hz AC power to this motor, its speed will be approximately half its base-speed value, or 864 RPM. There is really no need for speed-sensing feedback in an AC drive, because the motor's real speed will always be limited by the drive's output frequency. To control frequency *is* to control motor speed for AC synchronous and induction motors, so no tachogenerator feedback is necessary for an AC drive to "know" approximately[40] how fast the motor is turning. The non-necessity of speed feedback for AC drives eliminates a potential safety hazard common to DC drives: the possibility of a "runaway" event where the drive loses its speed feedback signal and sends full power to the motor.

As with DC motor drives, there is a lot of electrical "noise" cast by VFD circuits. Square-edged pulse waveforms created by the rapid on-and-off switching of the power transistors are equivalent to infinite series of high-frequency sine waves[41], some of which may be of high enough frequency to self-propagate through space as electromagnetic waves. This *radio-frequency interference* or *RFI* may be quite severe given the high power levels of industrial motor drive circuits. For this reason, it is *imperative* that neither the motor power conductors nor the conductors feeding AC power to

[40] For more precise control of AC motor speed (especially at low speeds where slip speed becomes a greater percentage of actual speed), speed sensors may indeed be necessary.

[41] This equivalence was mathematically proven by Jean Baptiste Joseph Fourier (1768-1830), and is known as a *Fourier series*.

the drive circuit be routed anywhere near small-signal or control wiring, because the induced noise *will* wreak havoc with whatever systems utilize those low-level signals.

RFI noise on the AC power conductors may be mitigated by routing the AC power through *filter* circuits placed near the drive. The filter circuits block high-frequency noise from propagating back to the rest of the AC power distribution wiring where it may influence other electronic equipment. However, there is little that may be done about the RFI noise between the drive and the motor other than to shield the conductors in well-grounded metallic conduit.

25.2.3 Motor drive features

Modern DC and AC motor drives provide features useful when using electric motors as final control elements. Some common features seen in both VSDs and VFDs are listed here:

- Speed limiting
- Torque limiting
- Torque profile curves (used to regulate the amount of torque available at different motor speeds)
- Acceleration (speed rate-of-change) limiting
- Deceleration (speed rate-of-change) limiting
- Dynamic braking (turning the motor into an electromagnetic brake[42])
- Plugging (applying reverse-direction power to a motor to *quickly* stop it)
- Regenerative braking (turning the motor into a generator to recover kinetic energy)
- Overcurrent monitoring and automatic shut-down
- Overvoltage monitoring and automatic shut-down
- PWM frequency adjustment (may be helpful in reducing electromagnetic interference with some equipment)

Not only are some of these limiting parameters useful in extending the life of the motor, but they may also help extend the operating life of the mechanical equipment powered by the motor. It is certainly advantageous, for example, to have torque limiting on a conveyor belt motor, so that the motor does not apply full rated torque (i.e. stretching force) to the belt during start-up.

If a motor drive is equipped with digital network communication capability (e.g. Modbus), it is usually possible for a host system such as a PLC or DCS to update these control parameters as the motor is running.

[42] This is accomplished in very different ways for DC versus AC motors. To dynamically brake a DC motor, the field winding must be kept energized while a high-power load resistor is connected to the armature. As the motor turns, the armature will push current through the resistor, generating a braking torque as it does. To dynamically brake an AC motor, a relatively small DC current is passed through the stator windings, causing large braking currents to be induced in the rotor.

25.2.4 Metering pumps

A very common method for directly controlling low flow rates of fluids is to use a device known as a *metering pump*. A "metering pump" is a pump mechanism, motor, and drive electronics contained in a monolithic package. Simply supply 120 VAC power and a control signal to a metering pump, and it is ready to use.

Metering pumps are commonly used in water treatment processes to inject small quantities of treatment chemicals (e.g. coagulants, disinfectants, acid or caustic liquids for pH neutralization, corrosion-control chemicals) into the water flowstream, as is the Milton-Roy unit shown in this photograph:

Adjustment knobs on the front of the pump establish the maximum flow rate at a control signal value of 100%:

While some metering pumps use rotary motor and pump mechanisms, many use a "plunger" style mechanism operated by a solenoid at variable intervals. Thus, the latter type of metering pump does not provide continuous flow control, but rather a flow consisting of discrete pulse events distributed over a period of time. The "plunger" metering pumps are quite simple and reliable, and are entirely appropriate if non-continuous flow is permissible for the process.

References

Baumann, Hans D., *Control Valve Primer, A User's Guide*, Second Edition, Instrument Society of America, Research Triangle Park, NC, 1994.

"Cavitation in Control Valves", document L351 EN, Samson AG, Frankfurt, Germany.

Control Valve Handbook, Third Edition, Fisher Controls International, Inc., Marshalltown, IA, 1999.

Control Valve Sourcebook – Power & Severe Service, Fisher Controls International, Inc., Marshalltown, IA, 1988.

"Design ED, EAD, and EDR Sliding-Stem Control Valves", Product Bulletin 51.1:ED, Fisher, Marshalltown, IA, 2006.

Grumstrup, Bruce, *Considerations in the Design and Selection of Bellows Seal Equipment Valves*, Technical Monograph 37, Fisher Controls International Inc., Marshalltown, IA, 1991.

Hutchison, J.W., *ISA Handbook of Control Valves*, Second Edition, Instrument Society of America, Research Triangle Park, NC, 1976.

Irwin, J. David, *The Industrial Electronics Handbook*, CRC Press, Boca Raton, FL, 1997.

Jury, Floyd D., *Fundamentals of Aerodynamic Noise in Control Valves*, Technical Monograph 43, Fisher Controls International Inc., Marshalltown, IA, 1999.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Control Volume II*, Third Edition, CRC Press, Boca Raton, FL, 1999.

"Micro Trims for Globe and Angle Valve Applications", Product Bulletin 80.4:010, Emerson Process Management, Marshalltown, IA, 2005.

"Packing Selection Guidelines for Sliding-Stem Valves", Product Bulletin 59.1:062, Emerson Process Management, Marshalltown, IA, 2007.

"Pipeline Accident Report – Pipeline Rupture and Subsequent Fire in Bellingham, Washington June 10, 1999", NTSB/PAR-02/02, PB2002-916502, Notation 7264A, National Transportation Safety Board, Washington DC, 2002.

Richardson, Jonathan W., *Primary Seat Shutoff*, Technical Monograph 47, Fisher Controls International LLC, Marshalltown, IA, 2005.

Riveland, Marc, *Fundamentals of Valve Sizing for Liquids*, Technical Monograph 30, Fisher Controls International Inc., Marshalltown, IA, 1985.

Schafbuch, Paul, *Fundamentals of Flow Characterization*, Technical Monograph 29, Fisher Controls International Inc., Marshalltown, IA, 1985.

Warnett, Chris, *Using Valve Actuators as Predictive Maintenance Tools for MOVs*, Rotork Controls, Inc., Rochester, NY, 2000.

"Valve Sizing Technical Bulletin", document MS-06-84-E, revision 3, Swagelok Company, MI, 2002.

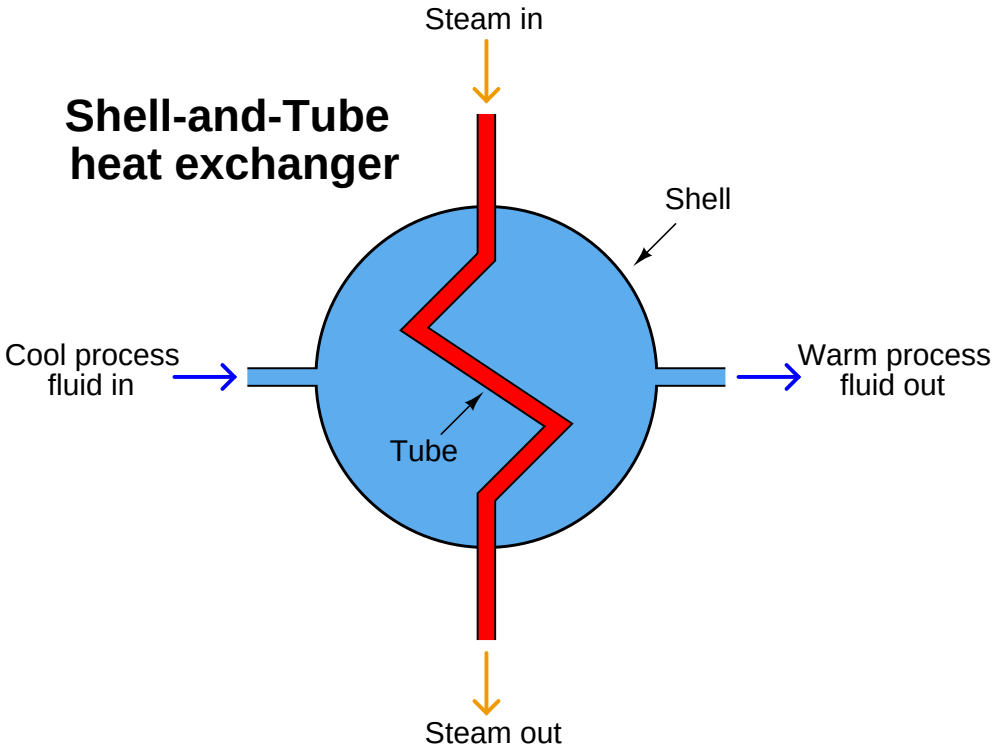
Chapter 26

Principles of feedback control

Instrumentation is the science of automated measurement and control. Applications of this science abound in modern research, industry, and everyday living. From automobile engine control systems to home thermostats to aircraft autopilots to the manufacture of pharmaceutical drugs, automation surrounds us. This chapter explains some of the fundamental principles of automatic process control.

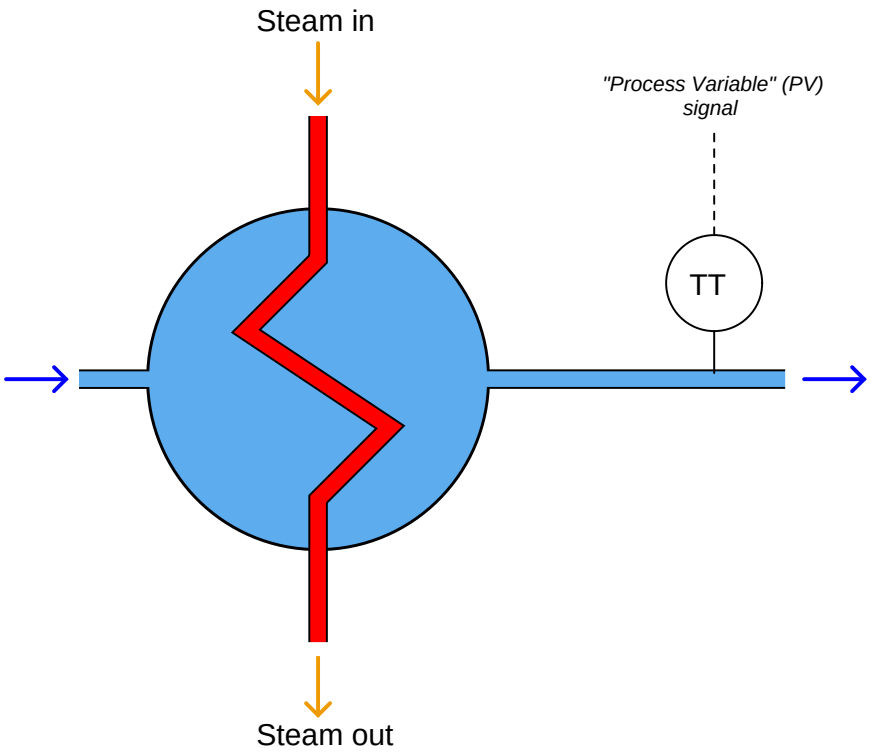
26.1 Basic feedback control principles

Before we begin our discussion on process control, we must define a few key terms. First, we have what is known as the *process*. This is the physical system we wish to monitor and control. For the sake of illustration, consider a heat exchanger that uses high-temperature steam to transfer heat to a lower-temperature liquid. Heat exchangers are used frequently in the chemical industries to maintain the necessary temperature of a chemical solution, so the desired blending, separation, or reactions can occur. A very common design of heat exchanger is the "shell-and-tube" style, where a metal shell serves as a conduit for the chemical solution to flow through, while a network of smaller tubes runs through the heating space, carrying steam or some other heating medium. The hotter steam flowing through the tubes transfers heat energy to the cooler process fluid surrounding the tubes, inside the shell of the heat exchanger:



In this case, the *process* is the entire heating system, consisting of the fluid we wish to heat, the heat exchanger, and the steam delivering the required heat energy. In order to maintain steady control of the process fluid's exiting temperature, we must find a way to measure it and represent that measurement in signal form so it may be interpreted by other instruments taking some form of control action. In instrumentation terms, the measuring device is known as a *transmitter*, because it *transmits* the process measurement in the form of a signal.

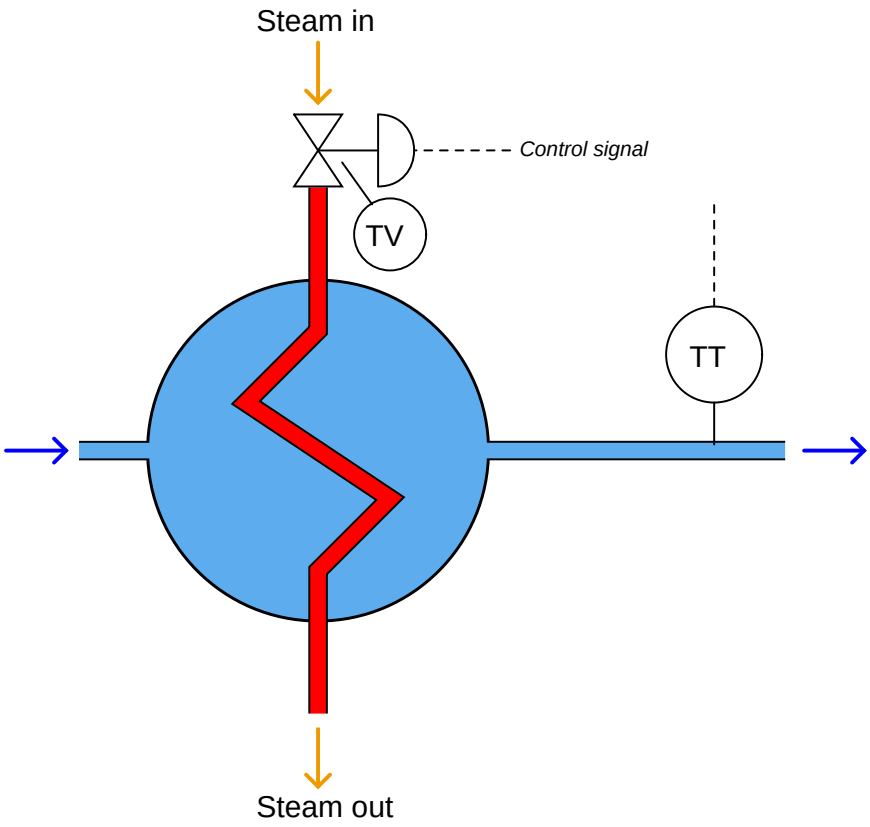
Transmitters are represented in process diagrams by small circles with identifying letters inside, in this case, "TT," which stands for **T**emperature **T**ransmitter:



The signal coming from the transmitter (shown in the illustration by the dashed line), representing the heated fluid's exiting temperature, is called the *process variable*. Like a variable in a mathematical equation that represents some story-problem quantity, this signal represents the measured quantity we wish to control in the process.

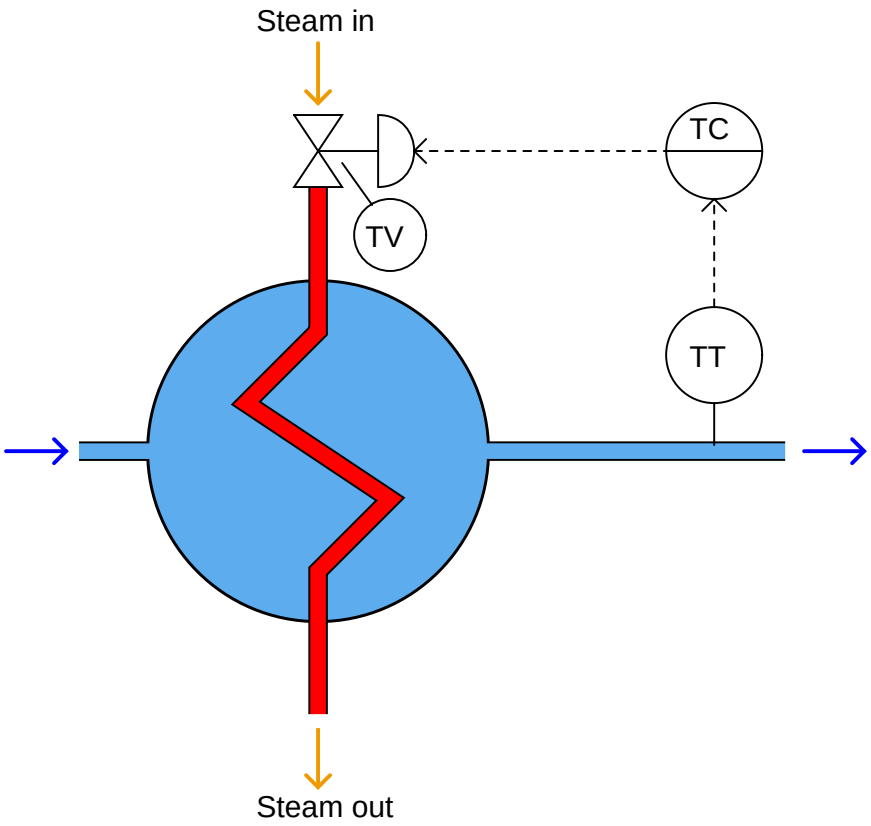
In order to exert control over the process variable, we must have some way of altering fluid flow through the heat exchanger, either of the process fluid, the steam, or both. Generally, it makes more sense to alter the flow of the heating medium (the steam), and let the process fluid flow rate be dictated by the demands of the larger process. If this heat exchanger were part of an oil refinery unit, for example, it would be far better to throttle steam flow to control oil temperature rather than to throttle the oil flow itself, since altering the oil's flow will undoubtedly affect other processes upstream and downstream of the exchanger. Ideally, the exchanger will act as a device that provides even, consistent temperature oil out, for any given temperature and flow-rate of oil in.

One convenient way to throttle steam flow into the heat exchanger is to use a control valve (labeled "TV" because it is a **T**emperature **V**alve). In general terms, a control valve is known as a *final control element*. Other types of final control elements exist (servo motors, variable-flow pumps, and other mechanical devices used to vary some physical quantity at will), but valves are the most common, and probably the simplest to understand. With a final control element in place, the steam flow becomes known as the *manipulated variable*, because it is the quantity we will manipulate in order to gain control over the process variable:



Valves come in a wide variety of sizes and styles. Some valves are hand-operated: that is, they have a "wheel" or other form of manual control that may be moved to "pinch off" or "open up" the flow passage through the pipe. Other valves come equipped with signal receivers and positioner devices, which move the valve mechanism to various positions at the command of a signal (usually an electrical signal, like the type output by transmitter instruments). This feature allows for remote control, so a human operator or computer device may exert control over the manipulated variable from a distance.

This brings us to the final, and most critical, component of the heat exchanger temperature control system: the *controller*. This is a device designed to interpret the transmitter's process variable signal and decide how far open the control valve needs to be in order to maintain that process variable at the desired value.



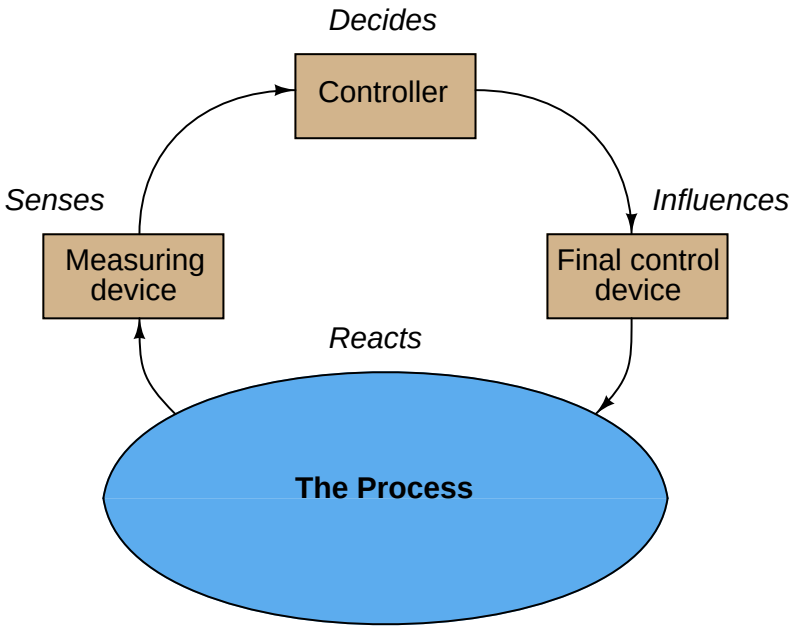
Here, the circle with the letters "TC" in the center represents the controller. Those letters stand for **T**emperature **C**ontroller, since the process variable being controlled is the process fluid's *temperature*. Usually, the controller consists of a computer making automatic decisions to open and close the valve as necessary to stabilize the process variable at some predetermined *setpoint*.

Note that the controller's circle has a solid line going through the center of it, while the transmitter and control valve circles are open. An open circle represents a field-mounted device according to the ISA standard for instrumentation symbols, and a single solid line through the middle of a circle tells us the device is located on the front of a control panel in a main control room location. So, even though the diagram might appear as though these three instruments are located close to one another, they in fact may be quite far apart. Both the transmitter and the valve must be located near the heat exchanger (out in the "field" area rather than inside a building), but the controller may be located a long distance away where human operators can adjust the setpoint from inside a safe and secure control room.

These elements comprise the essentials of a *feedback control system*: the *process* (the system

to be controlled), the *process variable* (the specific quantity to be measured and controlled), the *transmitter* (the device used to measure the process variable and output a corresponding signal), the *controller* (the device that decides what to do to bring the process variable as close to setpoint as possible), the *final control element* (the device that directly exerts control over the process), and the *manipulated variable* (the quantity to be directly altered to effect control over the process variable).

Feedback control may be viewed as a sort of information "loop," from the transmitter (measuring the process variable), to the controller, to the final control element, and through the process itself, back to the transmitter. Ideally, a process control "loop" not only holds the process variable at a steady level (the setpoint), but also maintains control over the process variable given changes in setpoint, and even changes in other variables of the process:



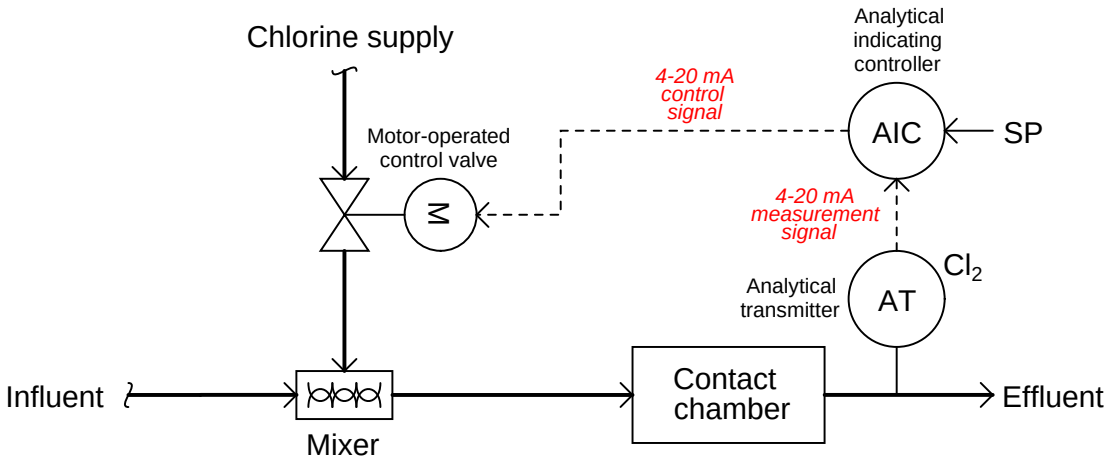
For example, if we were to raise the temperature setpoint in the heat exchanger process, the controller would automatically call for more steam flow by opening the control valve, thus introducing more heat energy into the process, thus raising the temperature to the new setpoint level. If the process fluid flow rate (an uncontrolled, or *wild* variable) were to suddenly increase, the heat exchanger outlet temperature would fall due to the physics of heat transfer, but once this drop was detected by the transmitter and reported to the controller, the controller would automatically call for additional steam flow to compensate for the temperature drop, thus bringing the process variable back in agreement with the setpoint. Ideally, a well-designed and well-tuned control loop will sense and compensate for *any* change in the process or in the setpoint, the end result being a process variable value that always holds steady at the setpoint value.

Many types of processes lend themselves to feedback control. Consider an aircraft autopilot system, keeping an airplane on a steady course heading: reading the plane's heading (process

variable) from an electronic compass and using the rudder as a final control element to change the plane's "yaw." An automobile's "cruise control" is another example of a feedback control system, with the process variable being the car's velocity, and the final control element being the engine's throttle. Steam boilers with automatic pressure controls, electrical generators with automatic voltage and frequency controls, and water pumping systems with automatic flow controls are further examples of how feedback may be used to maintain control over certain process variables.

Modern technology makes it possible to control nearly anything that may be measured in an industrial process. This extends beyond the pale of simple pressure, level, temperature, and flow variables to include even certain chemical properties.

In municipal water and wastewater treatment systems, numerous chemical quantities must be measured and controlled automatically to ensure maximum health and minimum environmental impact. Take for instance the chlorination of treated wastewater, before it leaves the wastewater treatment facility into a large body of water such as a river, bay, or ocean. Chlorine is added to the water to kill any residual bacteria so they do not consume oxygen in the body of water they are released to. Too little chlorine added, and not enough bacteria are killed, resulting in a high *biological oxygen demand* or *BOD* in the water which will asphyxiate the fish swimming in it. Too much chlorine added, and the chlorine itself poses a hazard to marine life. Thus, the chlorine content must be carefully controlled at a particular setpoint, and the control system must take aggressive action if the dissolved chlorine concentration strays too low or too high:



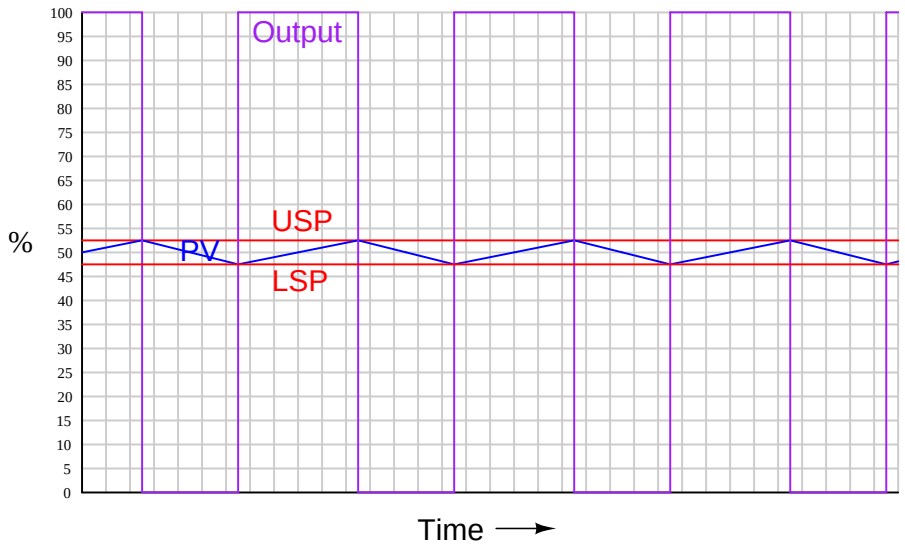
Now that we have seen the basic elements of a feedback control system, we will concentrate on the *algorithms* used in the controller to maintain a process variable at setpoint. For the scope of this topic, an "algorithm" is a mathematical relationship between the process variable and setpoint inputs of a controller, and the output (manipulated variable). Control algorithms determine *how* the manipulated variable quantity is deduced from PV and SP inputs, and range from the elementary to the very complex. In the most common form of control algorithm, the so-called "PID" algorithm, calculus is used to determine the proper final control element action for any combination of input signals.

26.2 On/off control

Once while working as an instrument technician in a large manufacturing facility, a mechanic asked me what it was that I did. I began to explain my job, which was essentially to calibrate, maintain, troubleshoot, document, and modify (as needed) all automatic control systems in the facility. The mechanic seemed puzzled as I explained the task of "tuning" loop controllers, especially those controllers used to maintain the temperature of large, gas-fired industrial furnaces holding many tons of molten metal. "Why does a controller have to be 'tuned'?" he asked. "All a controller does is turn the burner on when the metal's too cold, and turn it off when it becomes too hot!"

In its most basic form, the mechanic's assessment of the control system was correct: to turn the burner on when the process variable (molten metal temperature) drops below setpoint, and turn it off when it rises above setpoint. However, the actual algorithm is much more complex than that, finely adjusting the burner intensity according to the amount of *error* between PV and SP, the amount of time the error has accumulated, and the rate-of-change of the error over time. In his limited observation of the furnace controllers, though, he had noticed nothing more than the full-on/full-off action of the controller.

The technical term for a control algorithm that merely checks for the process variable exceeding or falling below setpoint is *on/off control*. In colloquial terms, it is known as *bang-bang* control, since the manipulated variable output of the controller rapidly switches between fully "on" and fully "off" with no intermediate state. Control systems this crude usually provide very imprecise control of the process variable. Consider our example of the shell-and-tube heat exchanger, if we were to implement simple on/off control[1]:



As you can see, the degree of control is rather poor. The process variable "cycles" between the upper and lower setpoints (USP and LSP) without ever stabilizing at the setpoint, because that

[1] To be precise, this form of on/off control is known as *differential gap* because there are two setpoints with a gap in between. While on/off control is possible with a single setpoint (FCE on when below setpoint and off when above), it is usually not practical due to the frequent cycling of the final control element.

would require the steam valve to be position somewhere *between* fully closed and fully open.

This simple control algorithm may be adequate for temperature control in a house, but not for a sensitive chemical process! Can you imagine what it would be like if an automobile's cruise control system relied on this algorithm? Not only is the lack of precision a problem, but the frequent cycling of the final control element may contribute to premature failure due to mechanical wear. In the heat exchanger scenario, thermal cycling (hot-cold-hot-cold) will cause metal fatigue in the tubes, resulting in a shortened service life. Furthermore, every excursion of the process variable above setpoint is wasted energy, because the process fluid is being heated to a greater temperature than what is necessary.

Clearly, the only practical answer to this dilemma is a control algorithm able to *proportion* the final control element rather than just operate it at zero or full effect (the control valve fully closed or fully open). This, in its simplest form, is called *proportional control*.

26.3 Proportional-only control

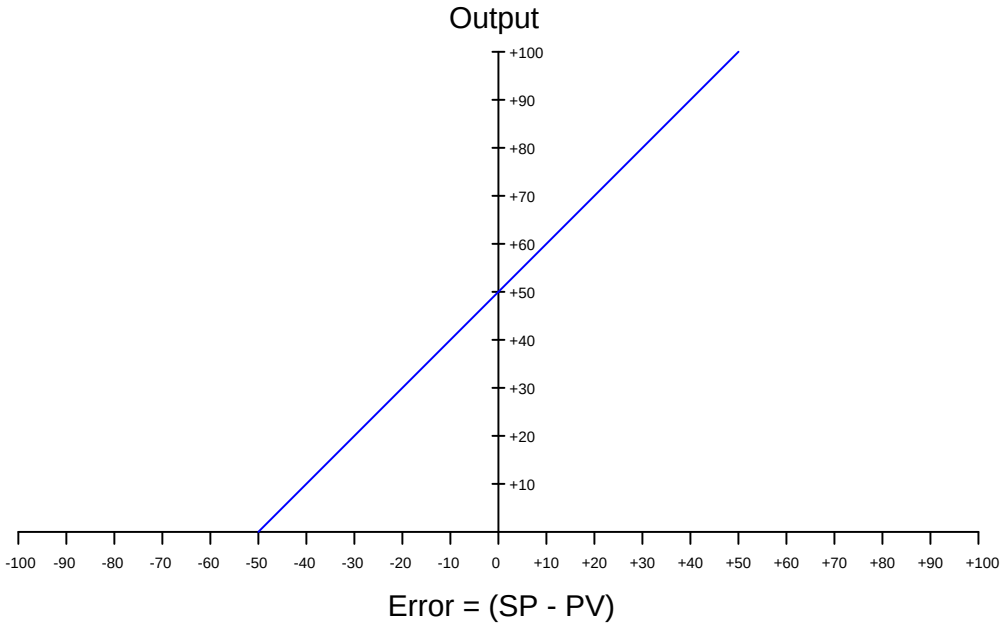
Here is where math starts to enter the algorithm: a proportional controller calculates the difference between the process variable signal and the setpoint signal, and calls it the *error*. This is a measure of how far off the process is deviating from its setpoint, and may be calculated as SP−PV or as PV−SP, depending on whether or not the controller has to produce an *increasing* output signal to cause an increase in the process variable, or output a *decreasing* signal to do the same thing. This choice in how we subtract determines whether the controller will be *reverse-acting* or *direct-acting*. The direction of action required of the controller is determined by the nature of the process, transmitter, and final control element. In this case, we are assuming that an increasing output signal sent to the valve results in increased steam flow, and consequently higher temperature, so our algorithm will need to be reverse-acting (i.e. an increase in measured temperature results in a decrease in output signal; error calculated as SP−PV). This error signal is then multiplied by a constant value called the *gain*, which is programmed into the controller. The resulting figure, plus a "bias" quantity, becomes the output signal sent to the valve to proportion it:

$$m = K_p e + b$$

Where,

m = Controller output
e = Error (difference between PV and SP)
K_p = Proportional gain
b = Bias

If this equation appears to resemble the standard slope-intercept form of linear equation ($y = mx + b$), it is more than coincidence. Often, the response of a proportional controller is shown graphically as a line, the slope of the line representing gain and the y-intercept of the line representing the output bias point, or what value the output signal will be when there is zero error (PV precisely equals SP):



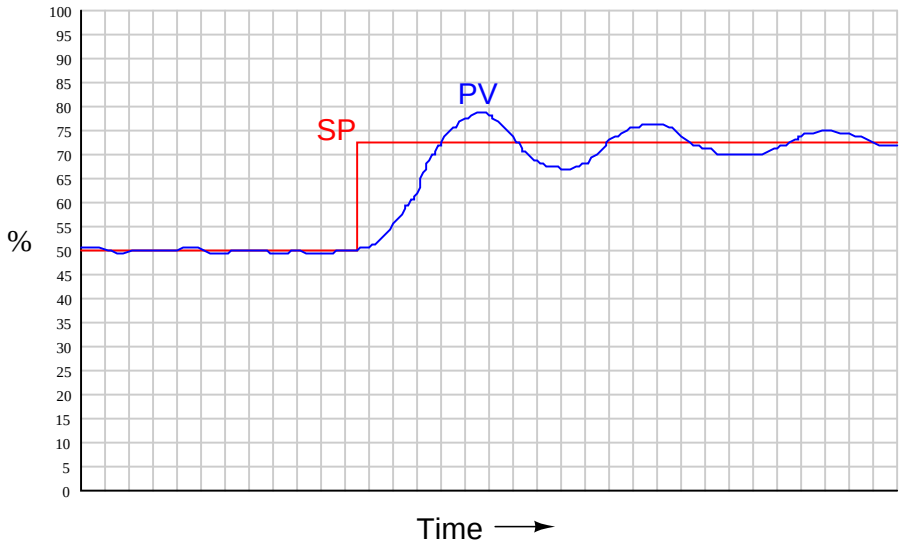
In this graph the bias value is 50% and the gain of the controller is 1.

Proportional controllers give us a choice as to how "sensitive" we want the controller to be to changes in process variable (PV) and setpoint (SP). With the simple on/off ("bang-bang") approach, there was no adjustment. Here, though, we get to program the controller for any desired level of aggressiveness. The gain value (K_p) of a controller is something which may be altered by a technician or engineer. In pneumatic controllers, this takes the form of a lever or valve adjustment; in analog electronic controllers, a potentiometer adjustment; in digital control systems, a programmable parameter.

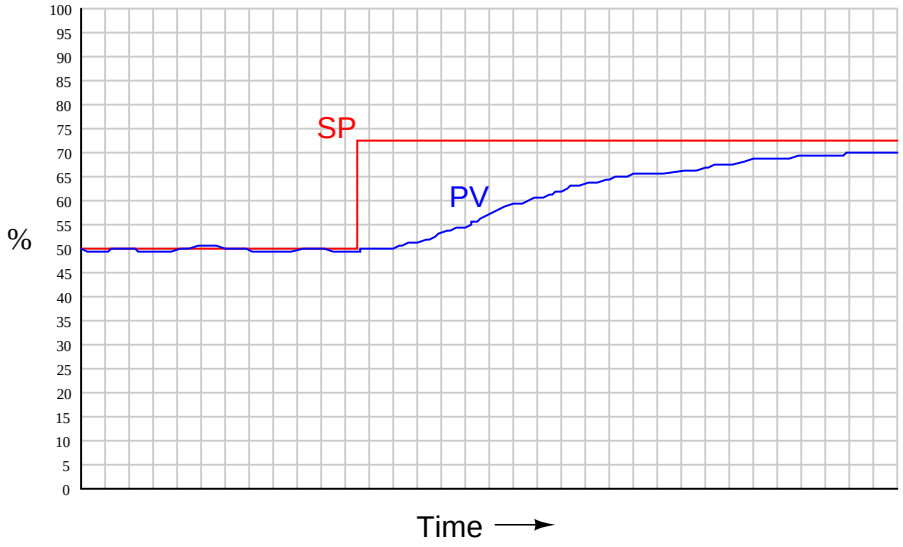
If the controller could be configured for infinite gain, its response would duplicate on/off control. That is, *any* amount of error will result in the output signal becoming "saturated" at either 0% or 100%, and the final control element will simply turn on fully when the process variable drops below setpoint and turn off fully when the process variable rises above setpoint. Conversely, if the controller is set for zero gain, it will become completely unresponsive to changes in either process variable *or* setpoint: the valve will hold its position at the bias point no matter what happens to the process.

Obviously, then, we must set the gain somewhere between infinity and zero in order for this algorithm to function any better than on/off control. Just how much gain a controller needs to have depends on the process and all the other instruments in the control loop.

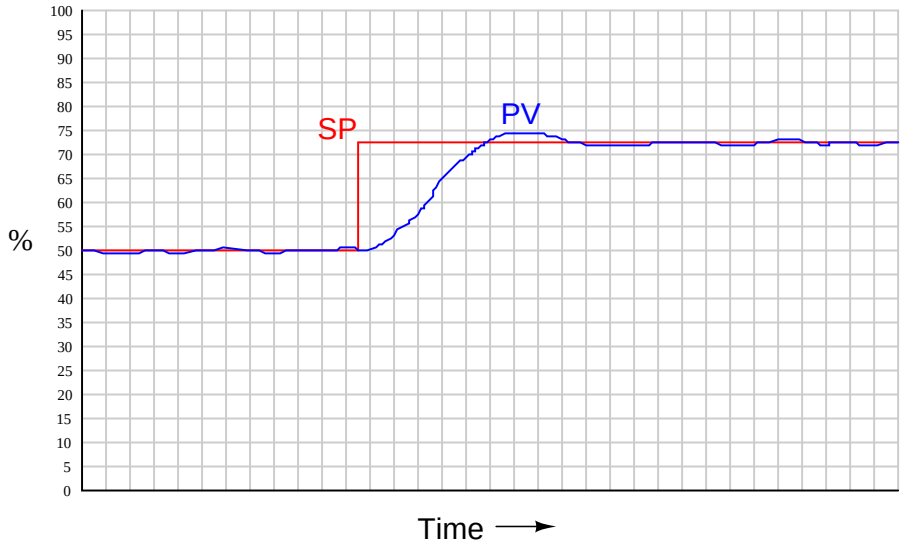
If the gain is set too high, there will be oscillations as the PV converges on a new setpoint value:



If the gain is set too low, the process response will be stable under steady-state conditions, but "sluggish" to changes in setpoint because the controller does not take aggressive enough action to cause quick changes in the process:



With proportional-only control, the only way to obtain fast-acting response to setpoint changes or "upsets" in the process is to set the gain constant high enough that some "overshoot" results:



As with on/off control, instances of overshoot (the process variable rising above setpoint) and undershoot (drifting below setpoint) are generally undesirable, and for the same reasons. Ideally, the controller will be able to respond in such a way that the process variable is made equal to setpoint as quickly as the process dynamics will allow, yet with no substantial overshoot or undershoot. With plain proportional control, however, this ideal goal is nearly impossible.

An unnecessarily confusing aspect of proportional control is the existence of two completely different ways to express the "aggressiveness" of proportional action. In the proportional-only equation shown earlier, the degree of proportional action was specified by the constant K_p, called *gain*. However, there is another way to express the sensitivity of proportional action, and that is to state the percentage of error change necessary to make the output (m) change by 100%. Mathematically, this is the inverse of gain, and it is called *proportional band* (PB):

$$K_p = \frac{1}{\text{PB}} \qquad \text{PB} = \frac{1}{K_p}$$

Gain is always specified as a unitless value[2], whereas proportional band is always specified as a percentage. For example, a gain value of 2.5 is equivalent to a proportional band value of 40%.

[2] In electronics, the unit of *decibels* is commonly used to express gains. Thankfully, the world of process control was spared the introduction of decibels as a unit of measurement for controller gain. The last thing we need is a *third* way to express the degree of proportional action in a controller!

Due to the existence of these two completely opposite conventions for specifying proportional action, you may see the proportional term of the control equation written differently depending on whether the author assumes the use of gain or the use of proportional band:

$$K_p = \text{gain} \qquad \text{PB} = \text{proportional band}$$

$$K_p e \qquad\qquad \frac{1}{\text{PB}} e$$

Many modern digital electronic controllers allow the user to conveniently select the unit they wish to use for proportional action. However, even with this ability, anyone tasked with adjusting a controller's "tuning" values may be required to translate between gain and proportional band, especially if certain values are documented in a way that does not match the unit configured for the controller.

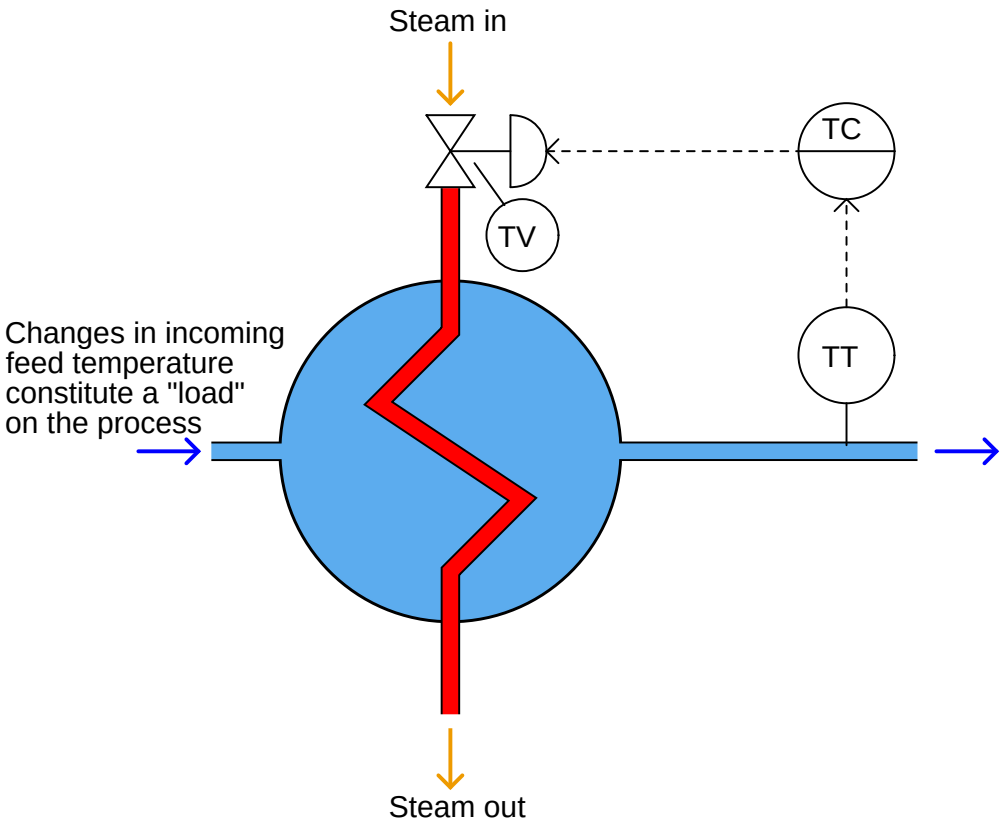
When you communicate the proportional action setting of a process controller, you should always be careful to specify either "gain" or "proportional band" to avoid ambiguity. *Never* simply say something like, "The proportional setting is twenty," for this could mean either:

- Gain = 20; Proportional band = 5% *. . . or . . .*
- Proportional band = 20%; Gain = 5

26.4 Proportional-only offset

A fundamental limitation of proportional control has to do with its response to changes in setpoint and changes in process *load*. A "load" in a controlled process is any variable subject to change which has an impact on the variable being controlled (the process variable), but is not subject to correction by the controller. In other words, a "load" is any variable in the process we cannot or do not control, yet affects the process variable we are trying to control.

In our hypothetical heat exchanger system, the temperature of the incoming process fluid is an example of a load:



If the incoming fluid temperature were to suddenly decrease, the immediate effect this would have on the process would be to decrease the outlet temperature (which is the temperature we are trying to maintain at a steady value). It should make intuitive sense that a colder incoming fluid will require more heat input to raise it to the same outlet temperature as before. If the heat input remains the same (at least in the immediate future), this colder incoming flow must make the outlet flow colder than it was before. Thus, incoming feed temperature has an impact on the outlet temperature whether we like it or not, and the control system has no way to regulate how warm or cold the process fluid is before it enters the heat exchanger. This is precisely the definition of a "load."

Of course, it is the job of the controller to counteract any tendency for the outlet temperature to stray from setpoint, but as we shall soon see this cannot be perfectly achieved with proportional control alone.

Let us carefully analyze the scenario of sudden inlet fluid temperature decrease to see how a proportional controller would respond. Imagine that previous to this sudden drop in feed temperature, the controller was controlling outlet temperature exactly at setpoint (PV = SP) and everything was stable. Recall that the equation for a proportional controller is as follows:

$$m = K_p e + b$$

Where,

m = Controller output
e = Error (difference between PV and SP)
K_p = Proportional gain
b = Bias

We know that a decrease in feed temperature will result in a decrease of outlet temperature with all other factors remaining the same. From the equation we can see that a decrease in process variable (PV) will cause the Output value in the proportional controller equation to increase. This means a wider-open steam valve, admitting more heating steam into the heat exchanger. All this is good, as we would expect the controller to call for more steam as the outlet temperature drops. But will this action be enough to bring the outlet temperature back up to setpoint where it was prior to the load change? Unfortunately it will not, although the reason for this may not be evident upon first inspection.

In order to prove that the PV will never go back to SP as long as the incoming feed temperature has dropped, let us imagine for a moment that somehow it did. According to the proportional controller equation, this would mean that the steam valve would resume its old pre-load-change position, only letting through the original flow rate of steam to heat the process fluid. Obviously, if the incoming process fluid is colder than before, and the flow rate is unchanged, the same amount of heat input (from steam) will be inadequate to maintain the outlet temperature at setpoint. If it were adequate, the outlet temperature never would have decreased and the controller never would have had to adjust the steam valve position at all. In other words, if the steam valve goes back to its old position, the outlet temperature will fall just as it did when the incoming flow suddenly became colder. This tells us the controller *cannot* bring the outlet temperature exactly to setpoint by proportional action alone.

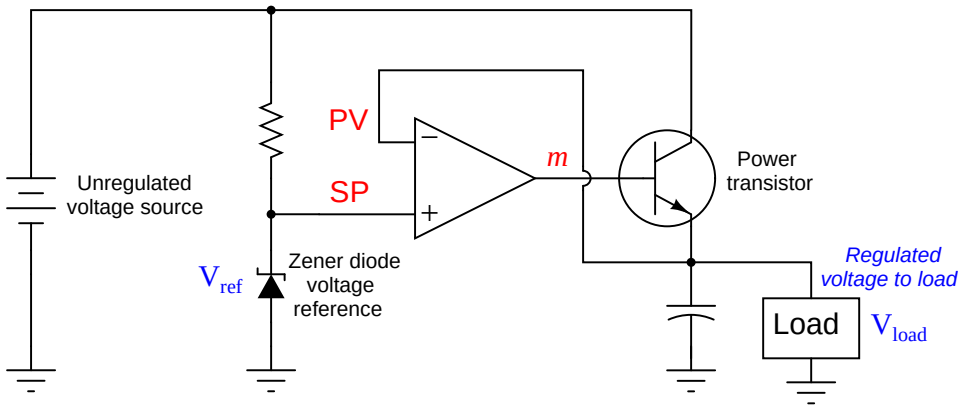
What *will* happen is that the controller's output will increase with falling outlet temperature, until there is enough steam flow admitted to the heat exchanger to prevent the temperature from falling any further. But in order to maintain this greater flow rate of steam (for greater heating effect), an error must develop between PV and SP. In other words, the process variable (temperature) *must deviate from setpoint* in order for the controller to call for more steam, in order that the process variable does not fall any further than this.

This necessary error between PV and SP is called *proportional-only offset*, sometimes less formally known as *droop*. The amount of droop depends on how severe the load change is, and how aggressive the controller responds (i.e. how much gain it has). The term "droop" is very misleading, as it is possible for the error to develop the other way (i.e. the PV might rise above SP due to a load change!). Imagine the opposite load-change scenario, where the incoming feed temperature

suddenly *rises* instead of falls. If the controller was controlling exactly at setpoint before this upset, the final result will be an outlet temperature that settles at some point above setpoint, enough so the controller is able to pinch the steam valve far enough closed to stop any further rise in temperature.

Proportional-only offset also occurs as a result of setpoint changes. We could easily imagine the same sort of effect following an operator's increase of setpoint for the temperature controller on the heat exchanger. After increasing the setpoint, the controller immediately increases the output signal, sending more steam to the heat exchanger. As temperature rises, though, the proportional algorithm causes the output signal to decrease. When the rate of heat energy input by the steam equals the rate of heat energy carried away from the heat exchanger by the heated fluid (a condition of *energy balance*, the temperature stops rising. This new equilibrium temperature will not be at setpoint, assuming the temperature was holding at setpoint prior to the human operator's setpoint increase. The new equilibrium temperature indeed *cannot* ever achieve any setpoint value higher than the one it did in the past, for if the error ever returned to zero (PV = SP), the steam valve would return to its old position, which we know would be insufficient to raise the temperature of the heated fluid.

An example of proportional-only control in the context of electronic power supply circuits is the following op-amp voltage regulator, used to stabilize voltage to a load with power supplied by an unregulated voltage source:



In this circuit, a zener diode establishes a "reference" voltage (which may be thought of as a "setpoint" for the controlling op-amp to follow). The operational amplifier acts as the proportional-only controller, sensing voltage at the load (PV), and sending a driving output voltage to the base of the power transistor to keep load voltage constant despite changes in the supply voltage or changes in load current (both "loads" in the process-control sense of the word, since they tend to influence voltage at the load circuit without being under the control of the op-amp).

If everything functions properly in this voltage regulator circuit, the load's voltage will be stable over a wide range of supply voltages and load currents. However, the load voltage cannot ever *precisely* equal the reference voltage established by the zener diode, even if the operational amplifier (the "controller") is without defect. The reason for this incapacity to perfectly maintain

"setpoint" is the simple fact that in order for the op-amp to generate any output signal at all, there *absolutely must be* a differential voltage between the two input terminals for the amplifier to amplify. Operational amplifiers (ideally) generate an output voltage equal to the enormously high gain value (A_V) multiplied by the difference in input voltages (in this case, $V_{ref} - V_{load}$). If V_{load} (the "process variable") were to ever achieve equality with V_{ref} (the "setpoint"), the operational amplifier would experience absolutely no differential input voltage to amplify, and its output signal driving the power transistor would fall to zero. Therefore, there must always exist some *offset* between V_{load} and V_{ref} (between process variable and setpoint) in order to give the amplifier some input voltage to amplify.

The amount of offset is ridiculously small in such a circuit, owing to the enormous gain of the operational amplifier. If we take the op-amp's transfer function to be $V_{out} = A_V(V_{(+)} - V_{(-)})$, then we may set up an equation predicting the load voltage as a function of reference voltage (assuming a constant 0.7 volt drop between the base and emitter terminals of the transistor):

$$V_{out} = A_V(V_{(+)} - V_{(-)})$$

$$V_{out} = A_V(V_{ref} - V_{load})$$

$$V_{load} + 0.7 = A_V(V_{ref} - V_{load})$$

$$V_{load} + 0.7 = A_V V_{ref} - A_V V_{load}$$

$$(A_V + 1)V_{load} + 0.7 = A_V V_{ref}$$

$$(A_V + 1)V_{load} = A_V V_{ref} - 0.7$$

$$V_{load} = \frac{A_V V_{ref} - 0.7}{A_V + 1}$$

If, for example, our zener diode produced a reference voltage of 5.00000 volts and the operational amplifier had an open-loop voltage gain of 250,000, the load voltage would settle at a theoretical value of 4.9999772 volts: just barely below the reference voltage value. If the op-amp's open-loop voltage gain were much less – say only 100 – the load voltage would only be 4.94356 volts. This still is quite close to the reference voltage, but definitely not as close as it would be with a greater op-amp gain!

Clearly, then, we can minimize proportional-only offset by increasing the gain of the process controller gain (i.e. decreasing its proportional band). This makes the controller more "aggressive" so it will move the control valve further for any given change in PV or SP. Thus, not as much error needs to develop between PV and SP to move the valve to any new position it needs to go. However, too much controller gain and the control system will become unstable: at best it will exhibit residual oscillations after setpoint and load changes, and at worst it will oscillate out of control altogether. Extremely high gains work well to minimize offset in operational amplifier circuits, only because time delays are negligible between output and input. In applications where large physical processes

are being controlled (e.g. furnace temperatures, tank levels, gas pressures, etc.) rather than voltages across small electronic loads, such high controller gains would be met with debilitating oscillations.

If we are limited in how much gain we can program in to the controller, how do we minimize this offset? One way is for a human operator to periodically place the controller in manual mode and move the control valve just a little bit more so the PV once again reaches SP, then place the controller back into automatic mode. In essence this technique adjusts the "Bias" term of the controller equation. The disadvantage of this technique is rather obvious: it requires frequent human intervention. What's the point of having an automation system that needs periodic human intervention to maintain setpoint?

A more sophisticated method for eliminating proportional-only offset is to add a different control action to the controller: one that takes action based on the amount of error between PV and SP and the amount of time that error has existed. We call this control mode *integral*, or *reset*. This will be the subject of the next section.

26.5 Integral (reset) control

Integration is a calculus principle, but don't let the word "calculus" scare you. You are probably already familiar with the concept of numerical integration even though you may have never heard of the term before.

Calculus is a form of mathematics dealing with *changing* variables, and how rates of change relate between different variables. When we "integrate" a variable with respect to time, what we are doing is *accumulating* that variable's value as time progresses. Perhaps the simplest example of this is a vehicle odometer, accumulating the total distance traveled by the vehicle over a certain time period. This stands in contrast to a speedometer, indicating how far the vehicle travels *per* unit of time.

Imagine a car moving along at exactly 30 miles per hour. How far will this vehicle travel after 1 hour of driving this speed? Obviously, it will travel 30 miles. Now, how far will this vehicle travel if it continues for another 2 hours at the exact same speed? Obviously, it will travel 60 more miles, for a total distance of 90 miles since it began moving. If the car's speed is a constant, calculating total distance traveled is a simple matter of multiplying that speed by the travel time.

The odometer mechanism that keeps track of the mileage traveled by the car may be thought of as *integrating* the speed of the car with respect to time. In essence, it is multiplying speed times time continuously to keep a running total of how far the car has gone. When the car is traveling at a high speed, the odometer "integrates" at a faster rate. When the car is traveling slowly, the odometer "integrates" slowly.

If the car travels in reverse, the odometer will decrement (count down) rather than increment (count up) because it sees a negative quantity for speed[3]. The rate at which the odometer decrements depends on how fast the car travels in reverse. When the car is stopped (zero speed), the odometer holds its reading and neither increments nor decrements.

Now imagine how this concept might apply to a process controller. Integration is provided either by a mechanism (in the case of a pneumatic controller), an op-amp circuit (in the case of an analog electronic controller), or by a microprocessor executing a digital integration algorithm. The variable being integrated is *error* (the difference between PV and SP). Thus the integral mode of the controller ramps the output either up or down over time, the direction of ramping determined by the sign of the error (PV greater or less than SP), and the rate of ramping determined by the magnitude of the error (how far away PV is from SP).

If proportional action is where the error tells the output how *far* to move, integral action is where the error tells the output how *fast* to move. One might think of integral as being how "impatient" the controller is, with integral action constantly ramping the output as far as it needs to go in order to eliminate error. Once the error is zero (PV = SP), of course, the integral action stops ramping, leaving the controller output (valve position) at its last value just like a stopped car's odometer holds a constant value.

[3] At least the old-fashioned mechanical odometers would. Some new cars use a pulse detector on the driveshaft which cannot tell the difference between forward and reverse, and therefore their odometers always increment. Shades of the movie *Ferris Bueller's Day Off*.

If we add an integral term to the controller equation, we get something that looks like this[4]:

$$m = K_p e + \frac{1}{\tau_i} \int e \, dt + b$$

Where,

m = Controller output
e = Error (difference between PV and SP)
K_p = Proportional gain
τ_i = Integral time constant (minutes)
t = Time
b = Bias

The most confusing portion of this equation for those new to calculus is the part that says "$\int e \, dt$". The integration symbol (looks like an elongated letter "S") tells us the controller will accumulate ("sum") multiple products of error (e) over tiny slices of time (dt). Quite literally, the controller multiplies error by time (for very short segments of time, dt) and continuously adds up all those products to contribute to the output signal which then drives the control valve (or other final control element). The integral time constant (τ_i) is a value set by the technician or engineer configuring the controller, proportioning this cumulative action to make it more or less aggressive over time.

To see how this works in a practical sense, let's imagine how a proportional + integral controller would respond to the scenario of a heat exchanger whose inlet temperature suddenly dropped. As we saw with proportional-only control, an inevitable offset occurs between PV and SP with changes in load, because an error *must* develop if the controller is to generate the different output signal value necessary to halt further change in PV. We called this effect *proportional-only offset*.

Once this error develops, though, integral action begins to work. Over time, a larger and larger quantity accumulates in the integral mechanism (or register) of the controller because an error persists over time. That accumulated value adds to the controller's output, driving the steam control valve further and further open. This, of course, adds heat at a faster rate to the heat exchanger, which causes the outlet temperature to rise. As the temperature re-approaches setpoint, the error becomes smaller and thus the integral action proceeds at a slower rate (like a car's odometer ticking by at a slower rate when the car's speed decreases). So long as the PV is below SP (the outlet temperature is still too cool), the controller will continue to integrate upwards, driving the control valve further and further open. Only when the PV rises to exactly meet SP does integral action finally rest, holding the valve at a steady position. Integral action ceaselessly works to eliminate any offset between PV and SP, thus neatly eliminating the offset problem experienced with proportional-only control action.

As with proportional action, there are (unfortunately) two completely opposite ways to specify the degree of integral action offered by a controller. One way is to specify integral action in terms of *minutes* or *minutes per repeat*. A large value of "minutes" for a controller's integral action means a less aggressive integral action over time, just as a large value for proportional band means a less aggressive proportional action. The other way to specify integral action is the inverse: how many

[4]The equation for a proportional + integral controller is often written without the bias term (b), because the presence of integral action makes it unnecessary.

repeats per minute, equivalent to specifying proportional action in terms of gain (large value means aggressive action). For this reason, you will sometimes see the integral term of a PID equation written differently:

$$\tau_i = \text{minutes per repeat} \qquad K_i = \text{repeats per minute}$$

$$\frac{1}{\tau_i} \int e\, dt \qquad K_i \int e\, dt$$

Many modern digital electronic controllers allow the user to select the unit they wish to use for integral action, just as they allow a choice between specifying proportional action as gain or as proportional band.

Integral is a highly effective mode of process control. In fact, some processes respond so well to integral controller action that it is possible to operate the control loop on integral action alone, without proportional. Typically, though, process controllers are designed to operate as proportional-only (P), proportional plus integral (PI).

Just as too much proportional gain will cause a process control system to oscillate, too much integral action (i.e. an integral time constant that is too short) will also cause oscillation. If the integration happens at too fast a rate, the controller's output will "saturate" either high or low before the process variable can make it back to setpoint. Once this happens, the only condition that will "unwind" the accumulated integral quantity is for an error to develop of the opposite sign, and remain that way long enough for a canceling quantity to accumulate. Thus, the PV must cross over the SP, guaranteeing at least another half-cycle of oscillation.

A similar problem called *reset windup* (or *integral windup*) happens when external conditions make it impossible for the controller to hold the process variable equal to setpoint. Imagine what would happen in the heat exchanger system if the steam boiler suddenly stopped producing steam. As outlet temperature dropped, the controller's proportional action would open up the control valve in a futile effort to raise temperature. If and when steam service is restored, proportional action would just move the valve back to its original position as the process variable returned to its original value (before the boiler died). This is how a proportional-only controller would respond to a steam "outage": nice and predictably. If the controller had integral action, however, a much worse condition would result. All the time spent with the outlet temperature below setpoint causes the controller's integral term to "wind up" in a futile attempt to admit more steam to the heat exchanger. This accumulated quantity can only be un-done by the process variable rising above setpoint for an equal error-time product, which means when the steam supply resumes, the temperature will rise well above setpoint until the integral action finally "unwinds" and brings the control valve back to a sane position again.

Various techniques exist to manage integral windup. Controllers may be built with limits to restrict how far the integral term can accumulate under adverse conditions. In some controllers, integral action may be turned off completely if the error exceeds a certain value. The surest fix for integral windup is human operator intervention, by placing the controller in manual mode. This typically resets the integral accumulator to a value of zero and loads a new value into the bias term of the equation to set the valve position wherever the operator decides. Operators usually wait until the process variable has returned at or near setpoint before releasing the controller into automatic mode again.

While it might appear that operator intervention is again a problem to be avoided (as it was in the case of having to correct for proportional-only offset), it is noteworthy to consider that

the conditions leading to integral windup usually occur only during shut-down conditions. It is customary for human operators to run the process manually anyway during a shutdown, and so the switch to manual mode is something they would do anyway and the potential problem of windup often never manifests itself.

26.6 Derivative (rate) control

The final facet of PID control is the "D" term, which stands for *derivative*. This is a calculus concept like integral, except most people consider it easier to understand. Simply put, derivative is the expression of a variable's *rate-of-change* with respect to another variable. Finding the derivative of a function (differentiation) is the inverse operation of integration. With integration, we calculated accumulated value of some variable's product with time. With derivative, we calculate the ratio of a variable's change per unit of time. Whereas integration is fundamentally a multiplicative operation (products), differentiation always involves division (ratios).

A controller with derivative (or *rate*) action looks at how fast the process variable changes per unit of time, and takes action proportional to that rate of change. In contrast to integral (reset) action which represents the "impatience" of the controller, derivative (rate) action represents the "cautious" side of the controller.

If the process variable starts to change at a high rate of speed, the job of derivative action is to move the control valve in such a direction as to counteract this rapid change, and thereby moderate the speed at which the process variable changes.

What this will do is make the controller "cautious" with regard to rapid changes in process variable. If the process variable is headed toward the setpoint value at a rapid rate, the derivative term of the equation will diminish the output signal, thus slowing tempering the control response and slowing the process variable's approach toward setpoint. To use an automotive analogy, it is as if a driver, driving a very heavy vehicle, preemptively applies the brakes to slow the vehicle's approach to an intersection, knowing that the vehicle doesn't "stop on a dime." The heavier the vehicle, the sooner a wise driver will apply the brakes, to avoid "overshoot" beyond the stop sign and into the intersection.

If we modify the controller equation to incorporate differentiation, it will look something like this:

$$m = K_p e + \frac{1}{\tau_i} \int e \, dt + \tau_d \frac{de}{dt} + b$$

Where,

m = Controller output
e = Error (difference between PV and SP)
K_p = Proportional gain
τ_i = Integral time constant (minutes)
τ_d = Derivative time constant (minutes)
t = Time
b = Bias

The $\frac{de}{dt}$ term of the equation expresses the rate of change of error (e) over time (t). The lower-case letter "d" symbols represent the calculus concept of *differentials* which may be thought of in this context as very tiny increments of the following variables. In other words, $\frac{de}{dt}$ refers to the ratio of a very small change in error (de) over a very small increment of time (dt). On a graph, this is interpreted as the slope of a curve at a specific point (slope being defined as *rise over run*).

It is also possible to build a controller with proportional and derivative actions, but lacking integral action. These are most commonly used in applications prone to wind-up, and where the elimination of offset is not critical:

$$m = K_p e + \tau_d \frac{de}{dt} + b$$

Many PID controllers offer the option of calculating derivative response based on rates of change for the process variable (PV) only, rather than the error (PV − SP or SP − PV). This avoids huge "spikes" in the output of the controller if ever a human operator makes a sudden change in setpoint[5]. The mathematical expression for such a controller would look like this:

$$m = K_p e + \frac{1}{\tau_i} \int e \, dt + \tau_d \frac{d\text{PV}}{dt} + b$$

It should be mentioned that derivative mode should be used with caution. Since it acts on rates of change, derivative action will "go crazy" if it sees substantial noise in the PV signal. Even small amounts of noise possess extremely large rates of change (defined as percent PV change per minute of time) owing to the relatively high frequency of noise compared to the timescale of physical process changes.

Ziegler and Nichols, the engineers who wrote the ground-breaking paper entitled "Optimum Settings for Automatic Controllers" had these words to say regarding "pre-act" control (page 762 of the November 1942 *Transactions of the A.S.M.E.*):

> The latest control effect made its appearance under the trade name "Pre-Act." On some control applications, the addition of pre-act response made such a remarkable improvement that it appeared to be in embodiment of mythical "anticipatory" controllers. On other applications it appeared to be worse than useless. Only the difficulty of predicting the usefulness and adjustment of this response has kept it from being more widely used.

26.7 Summary of PID control terms

PID control can be a confusing concept to understand. Here, a brief summary of each term within PID (P. I, and D) is presented for your learning benefit.

[5] It should not be assumed that such spikes are always undesirable. In processes characterized by long lag times, such a response may be quite helpful in overcoming that lag for the purpose of rapidly achieving new setpoint values. Slave (secondary) controllers in cascaded systems – where the controller receives its setpoint signal from the output of another (primary, or master) controller – may similarly benefit from derivative action calculated on error instead of just PV. As usual, the specific needs of the application dictate the ideal controller configuration.

26.7.1 Proportional control mode (P)

Proportional – sometimes called *gain* or *sensitivity* – is a control action reproducing changes in input as changes in output. Proportional controller action responds to present changes in input by generating immediate and commensurate changes in output. When you think of "proportional action" (P), think *punctual*: this control action works immediately (never too soon or too late) to match changes in the input signal.

Mathematically defined, proportional action is the ratio of output change to input change. This may be expressed as a quotient of differences, or as a derivative (a rate of change, using calculus notation):

$$\text{Gain value} = \frac{\Delta\text{Output}}{\Delta\text{Input}}$$

$$\text{Gain value} = \frac{d\text{Output}}{d\text{Input}} = \frac{dm}{de}$$

For example, if the PV input of a proportional-only process controller with a gain of 2 suddenly changes ("steps") by 5 percent, and the output will immediately jump by 10 percent ($\Delta\text{Output} = \text{Gain} \times \Delta\text{Input}$). The direction of this output jump in relation to the direction of the input jump depends on whether the controller is configured for direct or reverse action.

A legacy term used to express this same concept is *proportional band*: the mathematical reciprocal of gain. "Proportional band" is defined as the amount of input change necessary to evoke full-scale (100%) output change in a proportional controller. Incidentally, it is always expressed as a percentage, never as fraction or as a decimal:

$$\text{Proportional Band value} = \frac{\Delta\text{Input}}{\Delta\text{Output}}$$

$$\text{Proportional Band value} = \frac{d\text{Input}}{d\text{Output}} = \frac{de}{dm}$$

Using the same example of a proportional controller exhibiting an output "step" of 10% in response to a PV "step" of 5%, the proportional band would be 50%: the reciprocal of its gain ($\frac{1}{2} = 50\%$). Another way of saying this is that a 50% input "step" would be required to change the output of this controller by a full 100%, since its gain is set to a value of 2.

26.7.2 Integral control mode (I)

Integral – sometimes called *reset* or *floating control* – is a control action causing the output signal to change over time at a rate proportional to the amount of error (the difference between PV and SP values). Integral controller action responds to error accumulated over time, ramping the output signal are far as it needs to go to completely eliminate error. If proportional (P) action tells the output how *far* to go when an error appears, integral (I) action tells the output how *fast* to move when an error appears. If proportional (P) action acts on the *present*, integral (I) action acts on the *past*. Thus, how far the output signal gets driven by integral action depends on the *history* of the error over time: how much error existed, and for how long. When you think of "integral action" (I), think *impatience*: this control action drives the output further and further the longer PV fails to match SP.

Mathematically defined, integral action is the ratio of output *velocity* to input error:

$$\text{Integral value (repeats per minute)} = \frac{\text{Output velocity}}{\text{Input error}}$$

$$\text{Integral value (repeats per minute)} = \frac{\frac{dm}{dt}}{e}$$

An alternate way to express integral action is to use the reciprocal unit of "minutes per repeat." If we define integral action in these terms, the defining equations must be reciprocated:

$$\text{Integral time constant (minutes per repeat)} = \tau_i = \frac{\text{Input error}}{\text{Output velocity}}$$

$$\text{Integral time constant (minutes per repeat)} = \tau_i = \frac{e}{\frac{dm}{dt}}$$

For example, if an error of 5% appears between PV and SP on an integral-only process controller with an integral value of 3 repeats per minute (i.e. an integral time constant of 0.333 minutes per repeat), the output will begin ramping at a rate of 15% per minute ($\frac{dm}{dt}$ = Integral_value $\times$ e, or $\frac{dm}{dt} = \frac{e}{\tau_i}$). In most PI and PID controllers, integral response is also multiplied by proportional gain, so the same conditions applied to a PI controller that happened to also have a gain of 2 would result in an output ramping rate of 30% per minute ($\frac{dm}{dt}$ = Gain_value $\times$ Integral_value $\times$ e, or $\frac{dm}{dt}$ = Gain_value $\times$ $\frac{e}{\tau_i}$). The direction of this ramping in relation to the direction (sign) of the error depends on whether the controller is configured for direct or reverse action.

26.7.3 Derivative control mode (D)

Derivative – sometimes called *rate* or *pre-act* – is a control action causing the output signal to be offset by an amount proportional to the rate at which the input is changing. Derivative controller action responds to how quickly the input changes over time, biasing the output signal commensurate with that rate of input change. If proportional (P) action tells the output how *far* to go when an error appears, derivative (D) action tells the output how far to go when the input *ramps*. If proportional (P) action acts on the *present* and integral (I) action acts on the *past*, derivative (D) action acts on the *future*: it effectively "anticipates" overshoot by tempering the output response according to how fast the process variable is rising or falling. When you think of "derivative action" (D), think *discretion*: this control action is cautious and prudent, working against change.

Mathematically defined, derivative action is the ratio of output offset to input *velocity*:

$$\text{Derivative time constant (minutes)} = \tau_d = \frac{\text{Output offset}}{\text{Input velocity}}$$

$$\text{Derivative time constant (minutes)} = \tau_d = \frac{\Delta\text{Output}}{\frac{de}{dt}}$$

For example, if the PV signal begins to ramp at a rate of 5% per minute on a process controller with a derivative time constant of 4 minutes, the output will immediately become offset by 20% (ΔOutput = Derivative_value $\times$ $\frac{de}{dt}$). In most PD and PID controllers, derivative response is also multiplied by proportional gain, so the same conditions applied to a PD controller that happened to also have a gain of 2 would result in an immediate offset of 40% (ΔOutput = Gain_value $\times$ Derivative_value $\times$ $\frac{de}{dt}$). The direction (sign) of this offset in relation to the direction of the input ramping depends on whether the controller is configured for direct or reverse action.

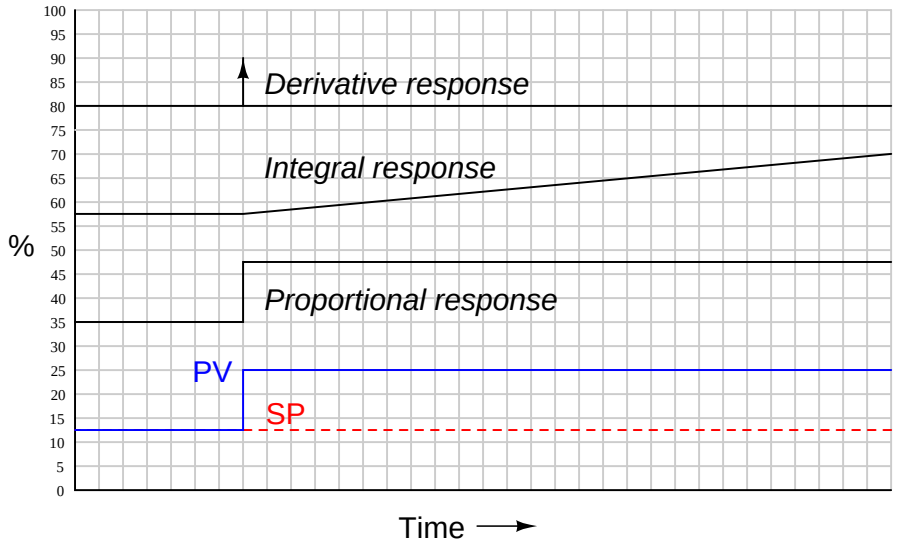
26.8 P, I, and D responses graphed

A very helpful method for understanding the operation of proportional, integral, and derivative control terms is to analyze their respective responses to the same input conditions over time. This section is divided into subsections showing P, I, and D responses for several different input conditions, in the form of graphs. In each graph, the controller is assumed to be *direct-acting* (i.e. an increase in process variable results in an increase in output).

It should be noted that these graphic illustrations are all qualitative, not quantitative. There is too little information given in each case to plot exact responses. The illustrations of P, I, and D actions focus only on the *shapes* of the responses, not their exact numerical values.

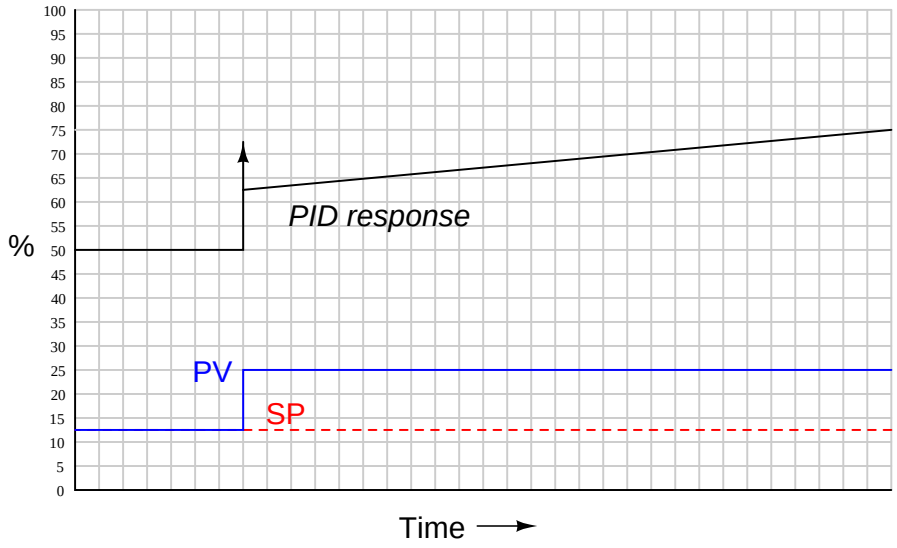
In order to *quantitatively* predict PID controller responses, one would have to know the values of all PID settings, as well as the original starting value of the output before an input change occurred and a time index of when the change(s) occurred.

26.8.1 Responses to a single step-change

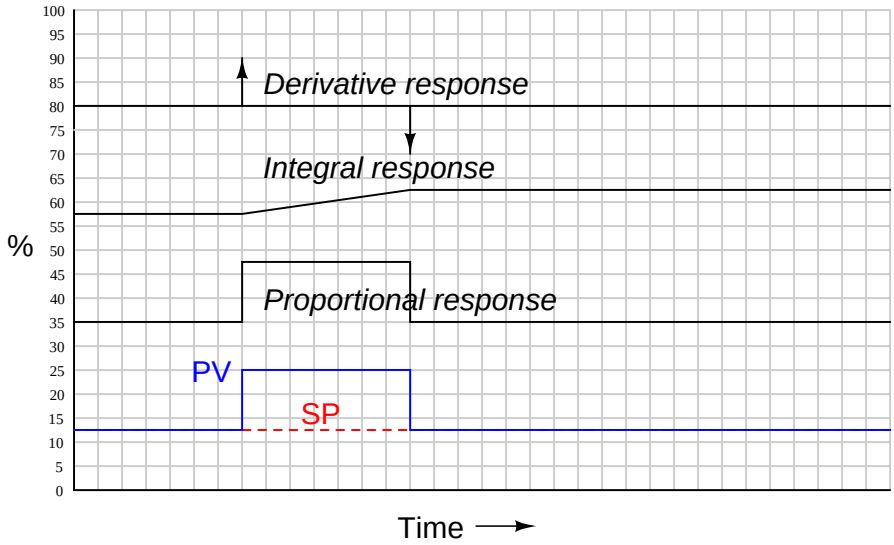


Proportional action directly mimics the shape of the input change (a step). Integral action ramps at a rate proportional to the magnitude of the input step. Since the input step holds a constant value, the integral action ramps at a constant rate (a constant *slope*). Derivative action interprets the step as an *infinite* rate of change, and so generates a "spike" driving the output to saturation.

When combined into one PID output, the three actions produce this response:



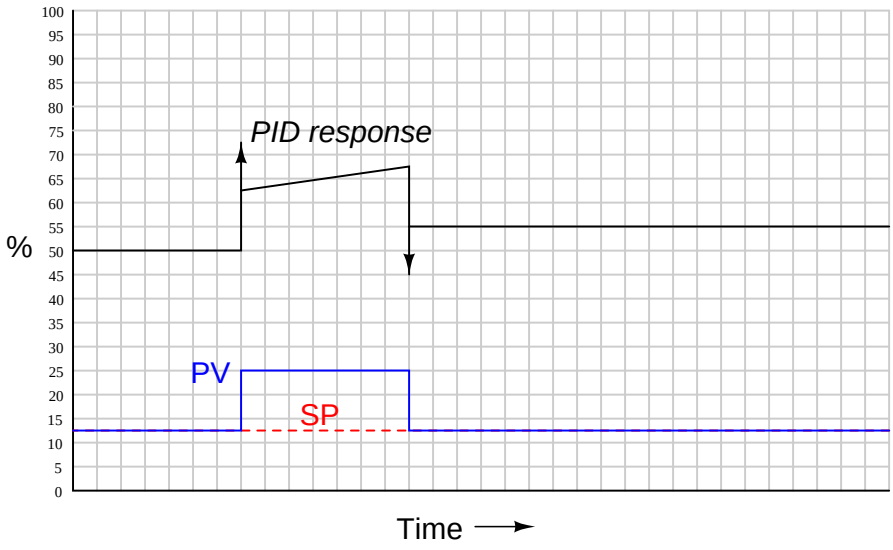
26.8.2 Responses to a momentary step-and-return



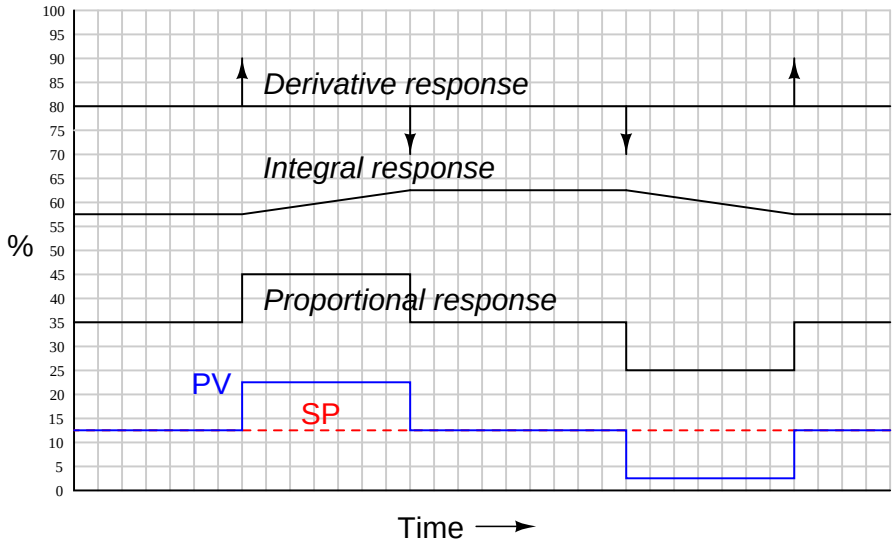
Proportional action directly mimics the shape of the input change (an up-and-down step). Integral action ramps at a rate proportional to the magnitude of the input step, for as long as the PV is unequal to the SP. Once PV = SP again, integral action stops ramping and simply holds the last value[6]. Derivative action interprets both steps as *infinite* rates of change, and so generates a "spike" at the leading and at the trailing edges of the step. Note how the leading (rising) edge causes derivative action to saturate high, while the trailing (falling) edge causes it to saturate low.

[6] This is a good example of how integral controller action represents the *history* of the PV − SP error. The continued offset of integral action from its starting point "remembers" the area accumulated under the rectangular "step" between PV and SP. This offset will go away only if a *negative* error appears having the same percent-minute product (area) as the positive error step.

When combined into one PID output, the three actions produce this response:

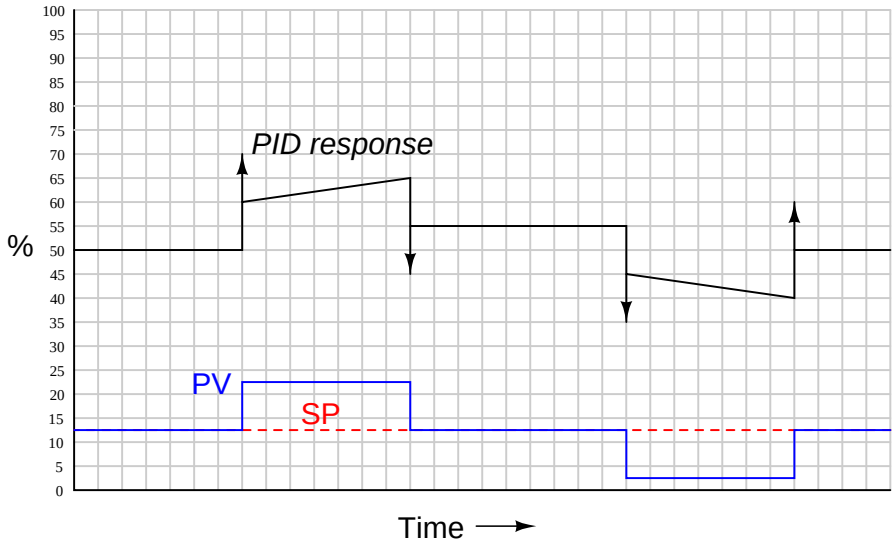


26.8.3 Responses to two momentary steps-and-returns

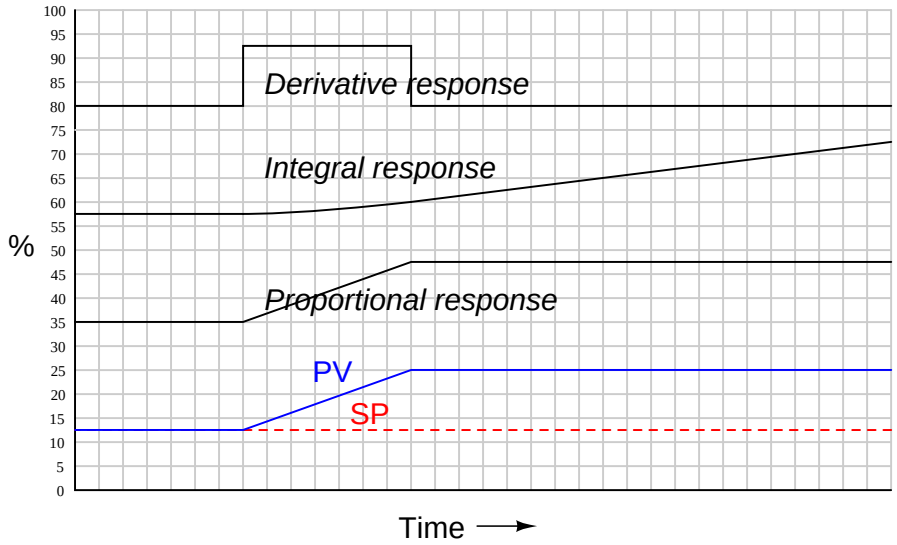


Proportional action directly mimics the shape of all input changes. Integral action ramps at a rate proportional to the magnitude of the input step, for as long as the PV is unequal to the SP. Once PV = SP again, integral action stops ramping and simply holds the last value. Derivative action interprets each step as an *infinite* rate of change, and so generates a "spike" at the leading and at the trailing edges of each step. Note how a leading (rising) edge causes derivative action to saturate high, while a trailing (falling) edge causes it to saturate low.

When combined into one PID output, the three actions produce this response:

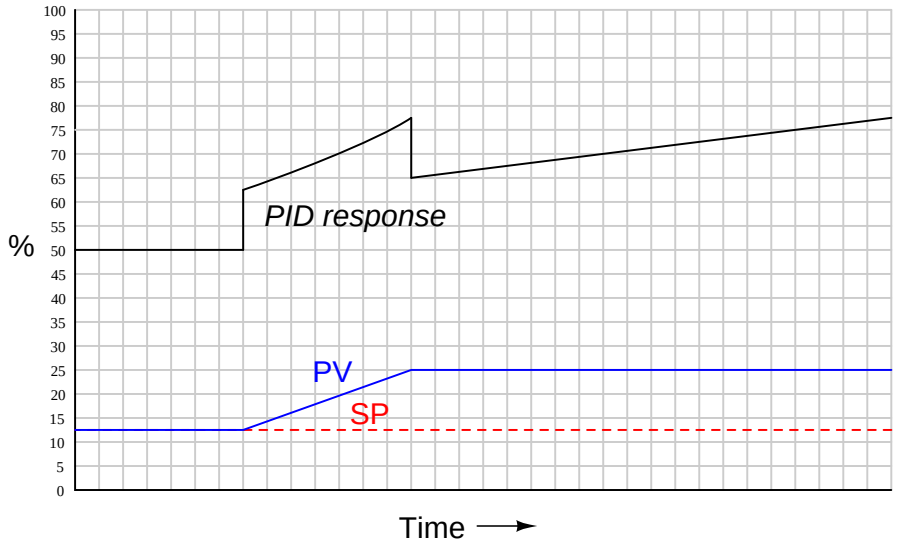


26.8.4 Responses to a ramp-and-hold

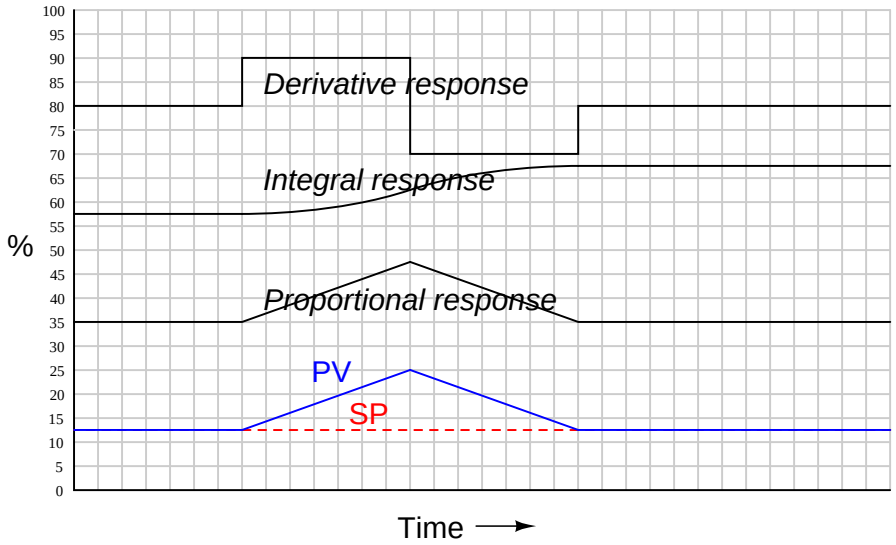


Proportional action directly mimics the ramp-and-hold shape of the input. Integral action ramps slowly at first (when the error is small) but increases ramping rate as error increases. When error stabilizes, integral rate likewise stabilizes. Derivative action offsets the output according to the input's ramping rate.

When combined into one PID output, the three actions produce this response:

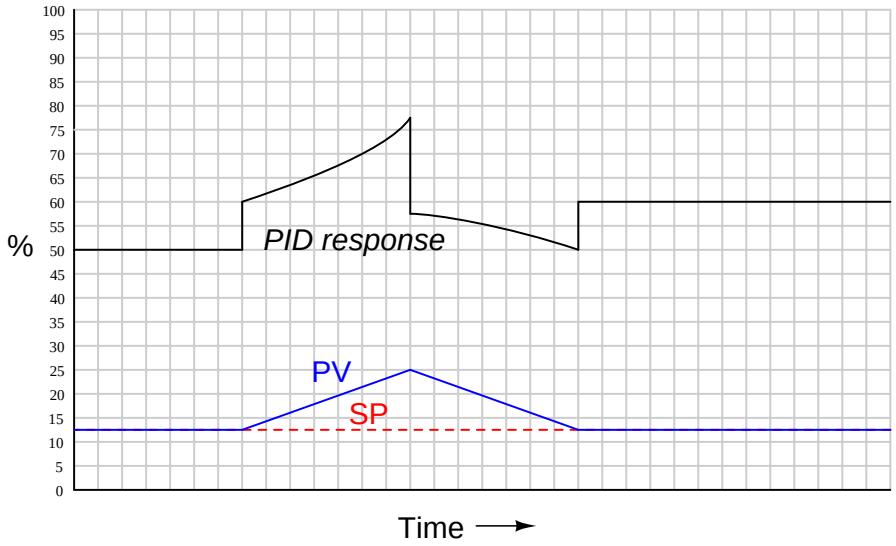


26.8.5 Responses to an up-and-down ramp

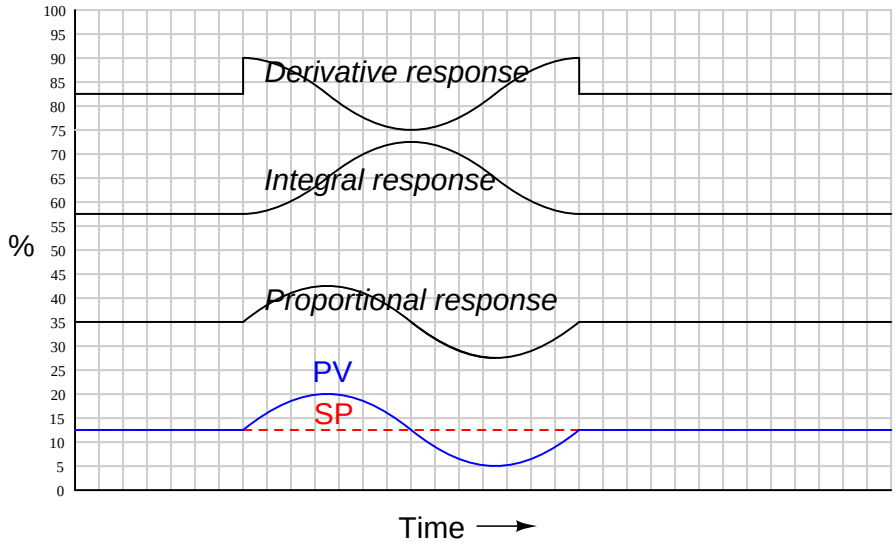


Proportional action directly mimics the up-and-down ramp shape of the input. Integral action ramps slowly at first (when the error is small) but increases ramping rate as error increases, then ramps slower as error decreases back to zero. Once PV = SP again, integral action stops ramping and simply holds the last value. Derivative action offsets the output according to the input's ramping rate: first positive then negative.

When combined into one PID output, the three actions produce this response:



26.8.6 Responses to a sine wavelet



As always, proportional action directly mimics the shape of the input. The 90^o phase shift seen in the integral and derivative responses, compared to the PV wavelet, is no accident or coincidence. The derivative of a sinusoidal function is *always* a cosine function, which is mathematically identical to a sine function with the angle advanced by 90^o:

$$\frac{d}{dx}(\sin x) = \cos x = \sin(x + 90^o)$$

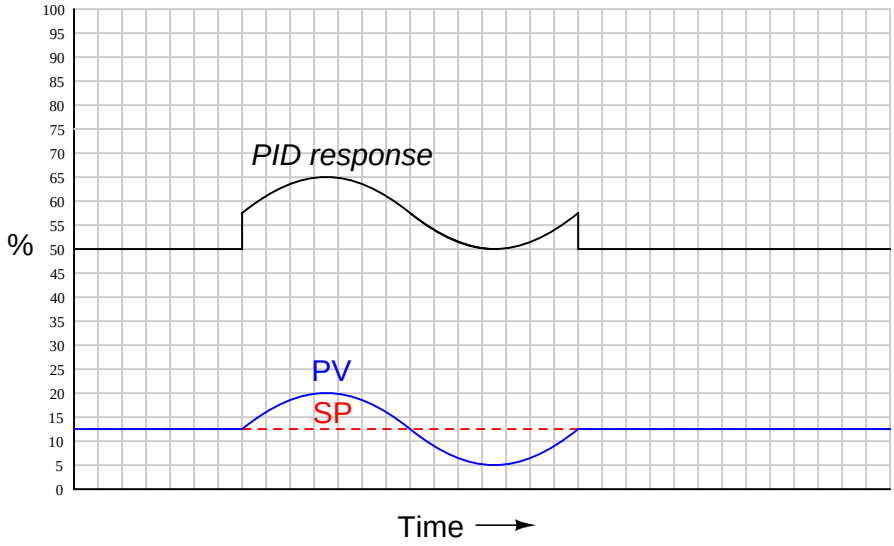
Conversely, the integral of a sine function is *always* a negative cosine function[7], which is mathematically identical to a sine function with the angle retarded by 90^o:

$$\int \sin x \, dx = -\cos x = \sin(x - 90^o)$$

In summary, the derivative operation always adds a positive (leading) phase shift to a sinusoidal input waveform, while the integral operation always adds a negative (lagging) phase shift to a sinusoidal input waveform.

[7] In this example, I have omitted the constant of integration (C) to keep things simple. The actual integral is as such: $\int \sin x \, dx = -\cos x + C = \sin(x - 90^o) + C$. This constant value is essential to explaining why the integral response does not immediately "step" like the derivative response does at the beginning of the PV sine wavelet.

When combined into one PID output, these particular integral and derivative actions mostly cancel, since they happen to be sinusoidal wavelets of equal amplitude and opposite phase. Thus, the only way that the final (PID) output differs from proportional-only action in this particular case is the "steps" caused by derivative action responding to the input's sudden rise at the beginning and end of the wavelet:



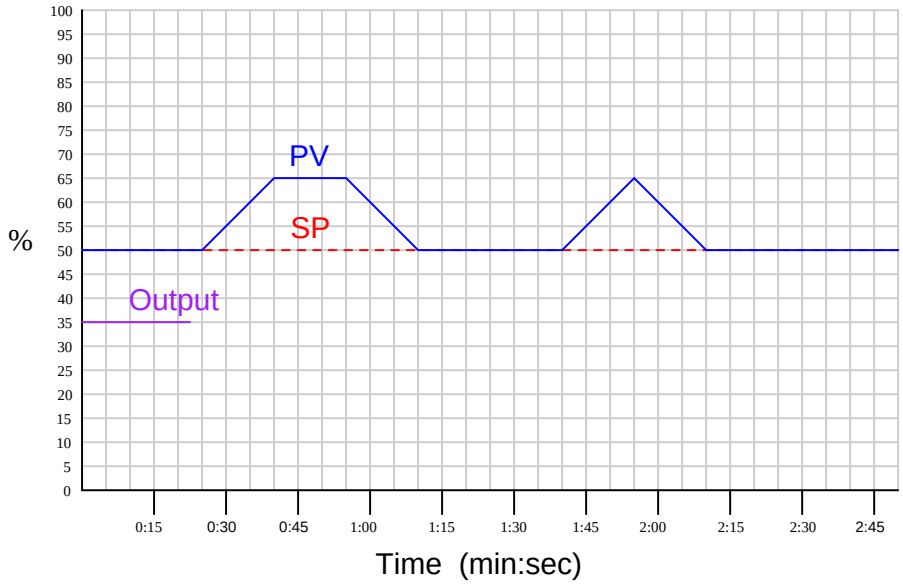
If the I and D tuning parameters were such that the integral and derivative responses were *not* equal in amplitude, their effects would not completely cancel. Rather, the resultant of P, I, and D actions would be a sine wavelet having a phase shift somewhere between -90^o and +90^o exclusive, depending on the relative strengths of the P, I, and D actions.

The 90 degree phase shifts associated with the integral and derivative operations are useful to understand when tuning PID controllers. If one is familiar with these phase shift relationships, it is relatively easy to analyze the response of a PID controller to a sinusoidal input (such as when a process oscillates following a sudden load or setpoint change) to determine if the controller's response is dominated by any one of the three actions. This may be helpful in "de-tuning" an over-tuned (overly aggressive) PID controller, if an excess of P, I, or D action may be identified from a phase comparison of PV and output waveforms.

26.8.7 Note to students regarding quantitative graphing

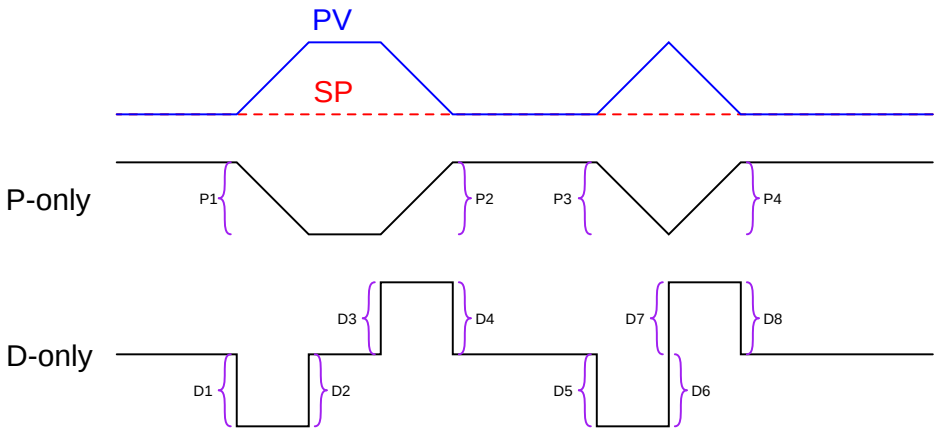
A common exercise for students learning the function of PID controllers is to practice graphing a controller's output given input (PV and SP) conditions, either qualitatively or quantitatively. This can be a frustrating experience for some students, as they struggle to accurately combine the effects of P, I, and/or D responses into a single output trend. Here, I will present a way to ease the pain.

Suppose for example you were tasked with graphing the response of a PD (proportional + derivative) controller to the following PV and SP inputs over time. You are told the controller has a gain of 1, a derivative time constant of 0.3 minutes, and is reverse-acting:

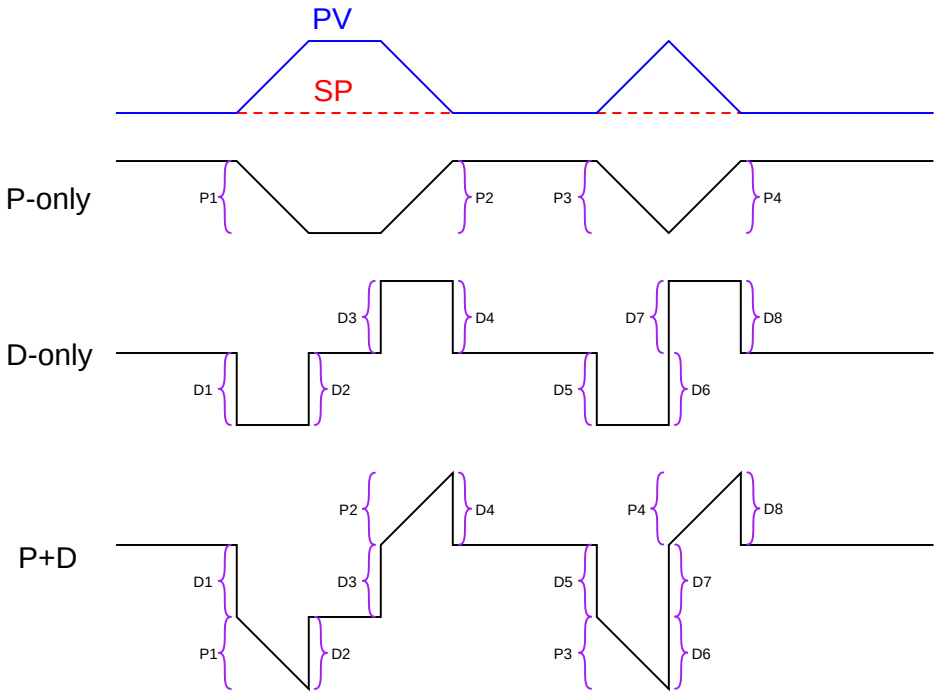


My first recommendation is to *qualitatively* sketch the individual P and D responses. Simply draw two different trends, each one right above or below the given PV/SP trends, showing the shapes of each response over time. You might even find it easier to do if you re-draw the original PV and SP trends on a piece of non-graph paper with the qualitative P and D trends also sketched on the same piece of non-graph paper. The purpose of the qualitative sketches is to separate the task of determining shapes from the task of determining numerical values, in order to simplify the process.

After sketching the separate P and D trends, label each one of the "features" (changes either up or down) in these qualitative trends. This will allow you to more easily combine the effects into one output trend later:



Now, you may qualitatively sketch an output trend combining each of these "features" into one graph. Be sure to label each ramp or step originating with the separate P or D trends, so you know where each "feature" of the combined output graph originates from:

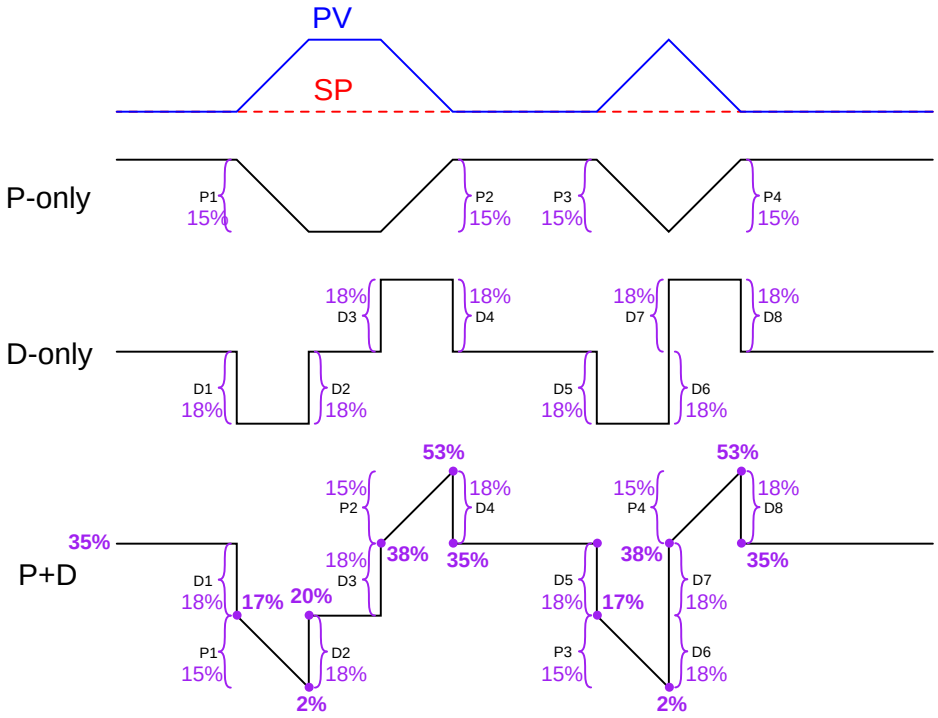


Once the general shape of the output has been qualitatively determined, you may go back to the separate P and D trends to calculate numerical values for each of the labeled "features."

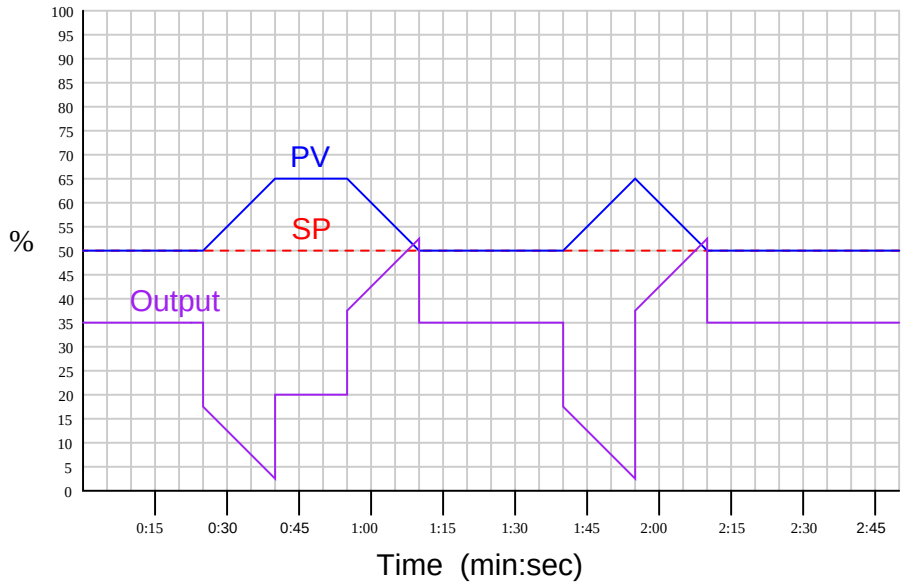
Note that each of the PV ramps is 15% in height, over a time of 15 seconds (one-quarter of a minute). With a controller gain of 1, the proportional response to each of these ramps will also be a ramp that is 15% in height.

Taking our given derivative time constant of 0.3 minutes and multiplying that by the PV's rate-of-change ($\frac{d\text{PV}}{dt}$) during each of its ramping periods (15% per one-quarter minute, or 60% per minute) yields a derivative response of 18% during each of the ramping periods. Thus, each derivative response "step" will be 18% in height.

Going back to the qualitative sketches of P and D actions, and to the combined (qualitative) output sketch, we may apply the calculated values of 15% for each proportional ramp and 18% for each derivative step to the labeled "features." We may also label the starting value of the output trend as given in the original problem (35%), to calculate actual output values at different points in time. Calculating output values at specific points in the graph becomes as easy as cumulatively adding and subtracting the P and D "feature" values to the starting output value:



Now that we know the output values at all the critical points, we may quantitatively sketch the output trend on the original graph:



26.9 Different PID equations

The equation used to describe PID control so far in this chapter is the simplest form, sometimes called the *parallel* equation, because each action (P, I, and D) occurs in separate terms of the equation, with the combined effect being a simple sum:

$$m = K_p e + \frac{1}{\tau_i} \int e\, dt + \tau_d \frac{de}{dt} + b \qquad \textbf{Parallel PID equation}$$

In the parallel equation, each action parameter (K_p, τ_i, τ_d) is independent of the others. At first, this may seem to be an advantage, for it means each adjustment made to the controller should only affect one aspect of its action. However, there are times when it is better to have the gain parameter affect all three control actions (P, I, and D).

An alternate version of the PID equation exists to provide this very functionality. This version is called the *Ideal* or *ISA* equation:

$$m = K_p \left(e + \frac{1}{\tau_i} \int e\, dt + \tau_d \frac{de}{dt} \right) + b \qquad \textbf{Ideal or ISA PID equation}$$

Here, the gain constant (K_p) is distributed to all terms within the parentheses, equally affecting all three control actions. Increasing K_p in this style of PID controller makes the P, the I, *and* the D actions equally more aggressive.

A third version, with origins in the peculiarities of pneumatic and analog electronic circuits, is called the *Series* or *Interacting* equation:

$$m = K_p \left(e + \frac{1}{\tau_i} \int e\, dt \right) \left(1 + \tau_d \frac{d}{dt} \right) + b \qquad \textbf{Series or Interacting PID equation}$$

Here, the gain constant (K_p) affects all three actions (P, I, and D) just as with the "ideal" equation. The difference, though, is the fact that both the integral and derivative constants have an effect on proportional action as well! That is to say, adjusting either τ_i or τ_d does not merely adjust those actions, but also influences the aggressiveness of proportional action.

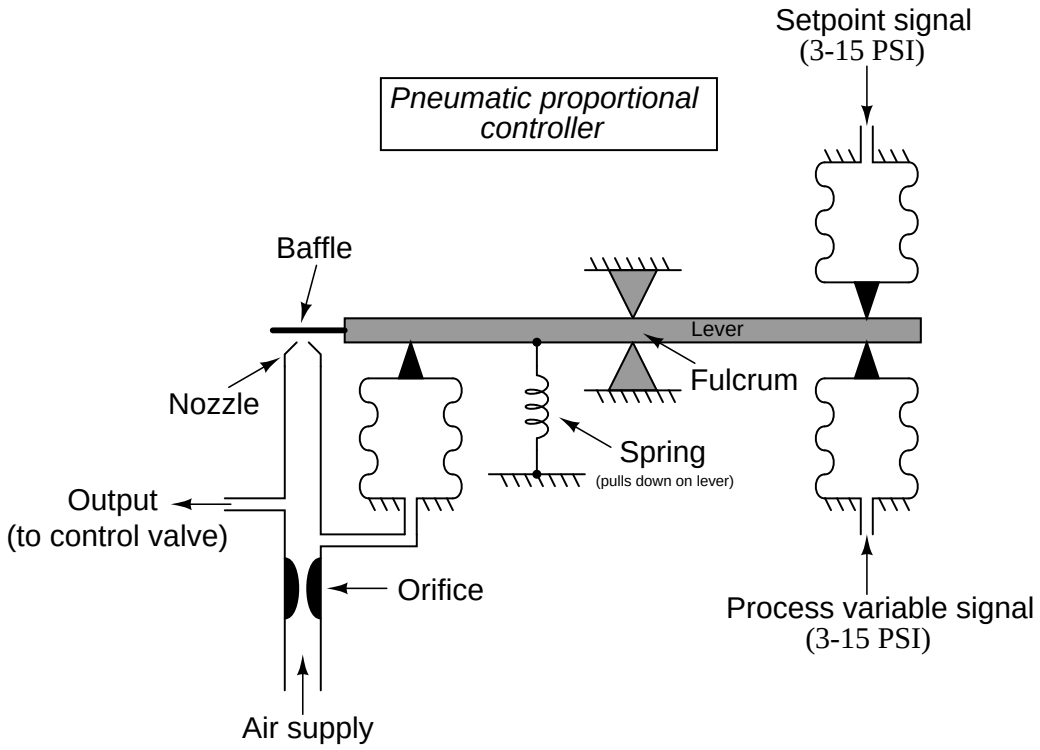
This "interacting" equation was an artifact of certain pneumatic and electronic controller designs. Back when these were the dominant technologies, and PID controllers were modularly designed such that integral and derivative actions were separate hardware modules included in a controller at additional cost beyond proportional-only action, the easiest way to implement the integral and derivative actions was in a way that just happened to have an interactive effect on controller gain. In other words, this odd equation form was a sort of compromise made for the purpose of simplifying the design of the controller hardware.

Interestingly enough, some digital PID controllers still implement the "interacting" PID equation even though it is no longer a necessary artifact of controller design.

26.10 Pneumatic PID controllers

Many pneumatic PID controllers use the *force-balance* principle. One or more input signals (in the form of pneumatic pressures) exert a force on a beam by acting through diaphragms, bellows, and/or bourdon tubes, which is then counter-acted by the force exerted on the same beam by an output air pressure acting through a diaphragm, bellows, or bourdon tube. The self-balancing mechanical system "tries" to keep the beam motionless through an exact balancing of forces, the beam's position precisely detected by a nozzle/baffle mechanism.

Throughout this section I will make reference to a pneumatic controller mechanism of my own design. This mechanism does not directly correspond to any particular manufacturer or model of pneumatic controller, but shares characteristics common to many. This design is shown here for the purpose of illustrating the development of P, I, and D control actions in as simple a context as possible:



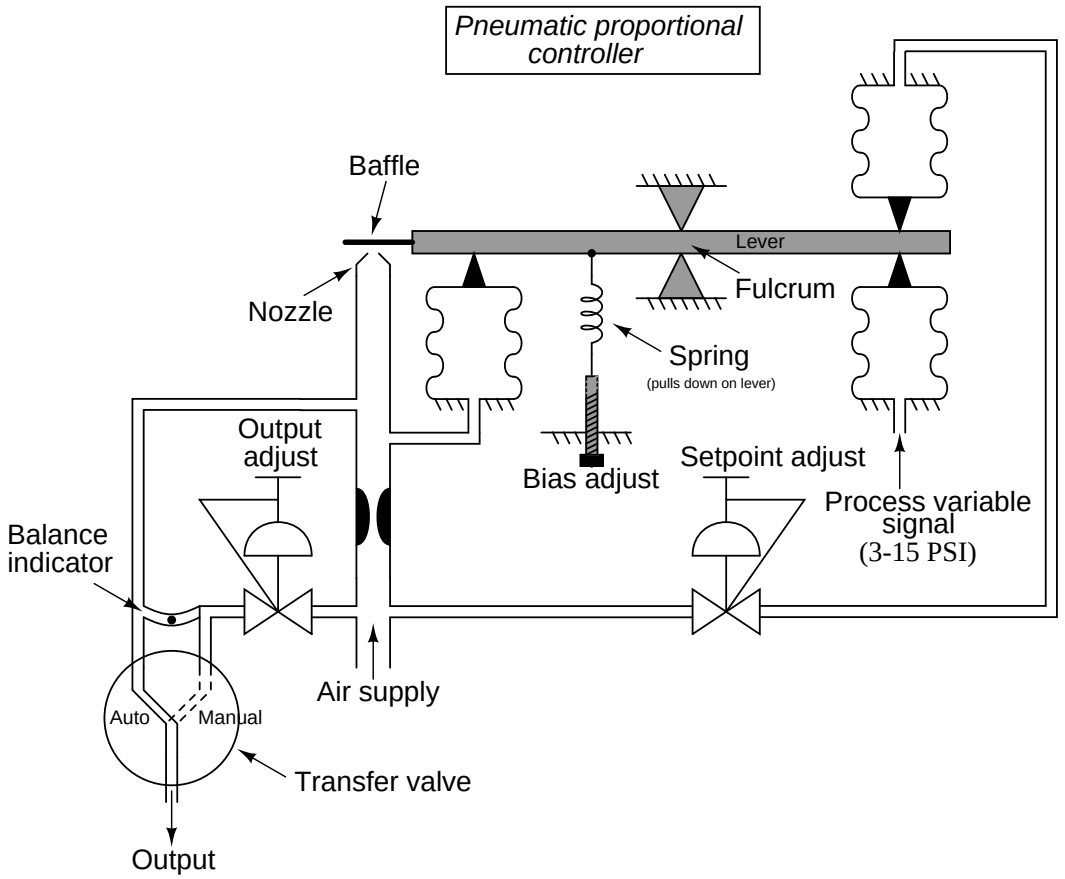
The action of this particular controller is *direct*, since an increase in process variable signal (pressure) results in an increase in output signal (pressure). Increasing process variable (PV) pressure attempts to push the right-hand end of the beam up, causing the baffle to approach the nozzle. This blockage of the nozzle causes the nozzle's pneumatic backpressure to increase, thus increasing the amount of force applied by the output feedback bellows on the left-hand end of the beam and returning the flapper (very nearly) to its original position. If we wished to reverse the controller's

action, all we would need to do is swap the pneumatic signal connections between the input bellows, so that the PV pressure was applied to the upper bellows and the SP pressure to the lower bellows.

Any factor influencing the ratio of input pressure(s) to output pressure may be exploited as a gain (proportional band) adjustment in this mechanism. Changing bellows area (either both the PV and SP bellows equally, or the output bellows by itself) would influence this ratio, as would a change in output bellows position (such that it pressed against the beam at some difference distance from the fulcrum point). Moving the fulcrum left or right is also an option for gain control, and in fact is usually the most convenient to engineer.

26.10.1 Automatic and manual modes

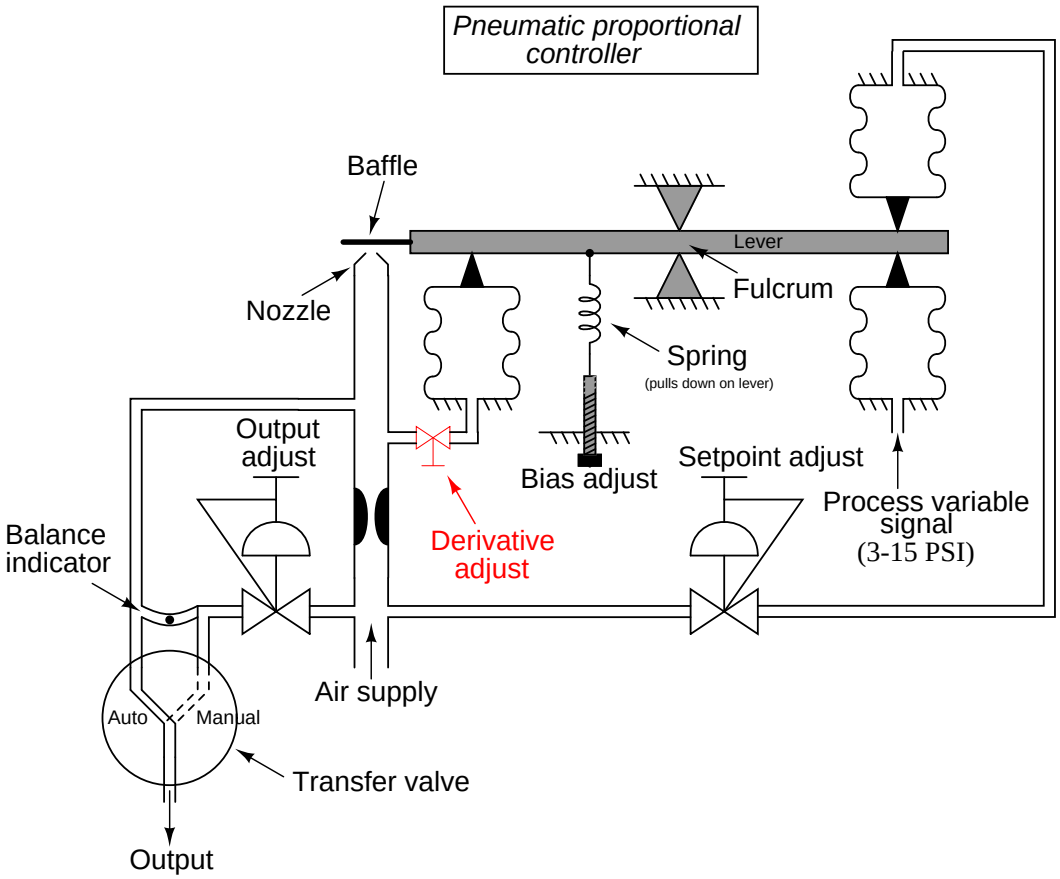
A more practical pneumatic proportional controller mechanism is shown in the next illustration, complete with setpoint and bias adjustments, and a manual control mode:



"Bumpless" transfer between automatic and manual modes is accomplished by the human operator paying attention to the *balance indicator* revealing any air pressure difference between the output bellows and the output adjust pressure regulator. When in automatic mode, a switch to manual mode involves adjusting the regulator until the balance indicator registers zero pressure difference, then switching the transfer valve to the "manual" position. The controller output is then at the direct command of the output adjust pressure regulator, and will not respond to changes in either PV or SP. "Bumplessly" switching back to automatic mode requires that either the output or the setpoint pressure regulators be adjusted until the balance indicator once again registers zero pressure difference, then switching the transfer valve to the "auto" position. The controller output will once again respond to changes in PV and SP.

26.10.2 Derivative and integral actions

Interestingly enough, derivative (rate) and integral (reset) control modes are relatively easy to add to this pneumatic controller mechanism. To add derivative control action, all we need to do is place a restrictor valve between the nozzle tube and the output feedback bellows, causing the bellows to delay filling or emptying its air pressure over time:

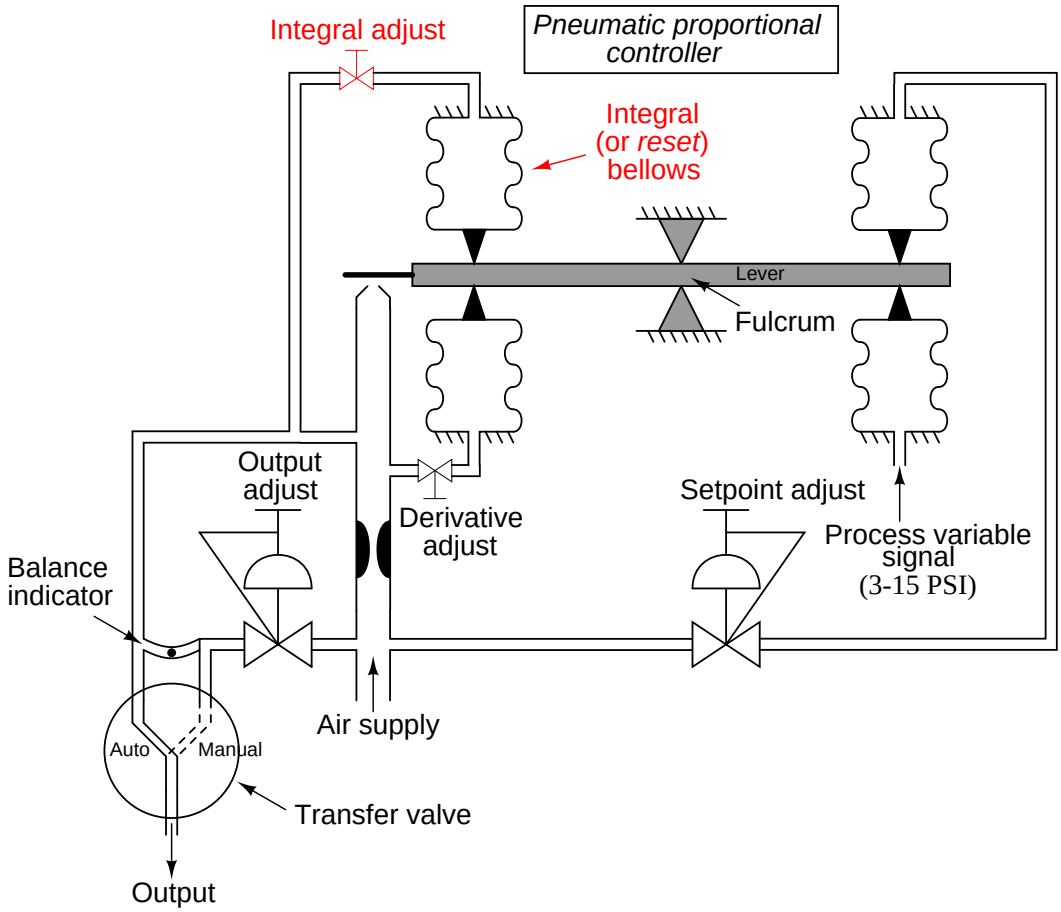


If any sudden change occurs in PV or SP, the output pressure will saturate before the output bellows has the opportunity to equalize in pressure with the output signal tube. Thus, the output pressure "spikes" with any sudden "step change" in input: exactly what we would expect with derivative control action.

If either the PV or the SP ramps over time, the output signal will ramp in direct proportion (proportional action), but there will *also* be an added offset of pressure at the output signal in order to keep air flowing either in or out of the output bellows at a constant rate to generate the force necessary to balance the changing input signal. Thus, derivative action causes the output pressure to shift either up or down (depending on the direction of input change) more than it would with just proportional action alone in response to a ramping input: exactly what we would expect from

a controller with both proportional and derivative control actions.

Integral action requires the addition of a second bellows (a "reset" bellows, positioned opposite the output feedback bellows) and another restrictor valve to the mechanism[8]:

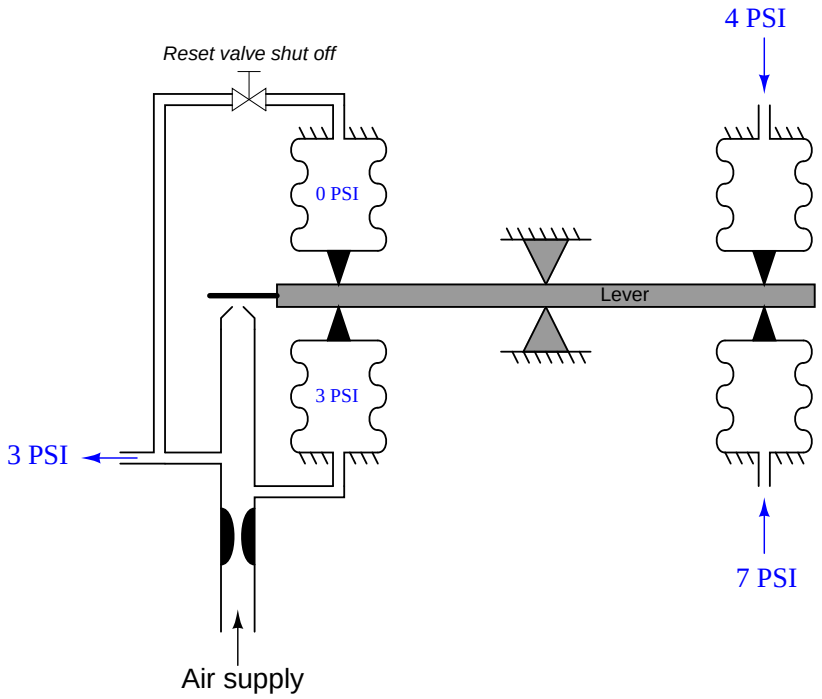


This second bellows takes air pressure from the output line and translates it into force that opposes the original feedback bellows. At first, this may seem counter-productive, for it nullifies the ability of this mechanism to continuously balance the force generated by the PV and SP bellows. Indeed, it would render the force-balance system completely ineffectual if this new "reset" bellows were allowed to inflate and deflate with no time lag. However, with a time lag provided by the restriction of the integral adjustment valve and the volume of the bellows (a sort of pneumatic "RC time constant"), the nullifying force of this bellows becomes delayed over time. As this bellows

[8]Practical integral action also requires the elimination of the bias spring and adjustment, which formerly provided a constant downward force on the left-hand side of the beam to give the output signal the positive offset necessary to avoid saturation at 0 PSI. Not only is a bias adjustment completely unnecessary with the addition of integral action, but it would actually cause problems by making the integral action "think" an error existed between PV and SP when there was none.

slowly fills (or empties) with pressurized air from the nozzle, the change in force on the beam causes the regular output bellows to have to "stay ahead" of the reset bellows action by constantly filling (or emptying) at some rate over time.

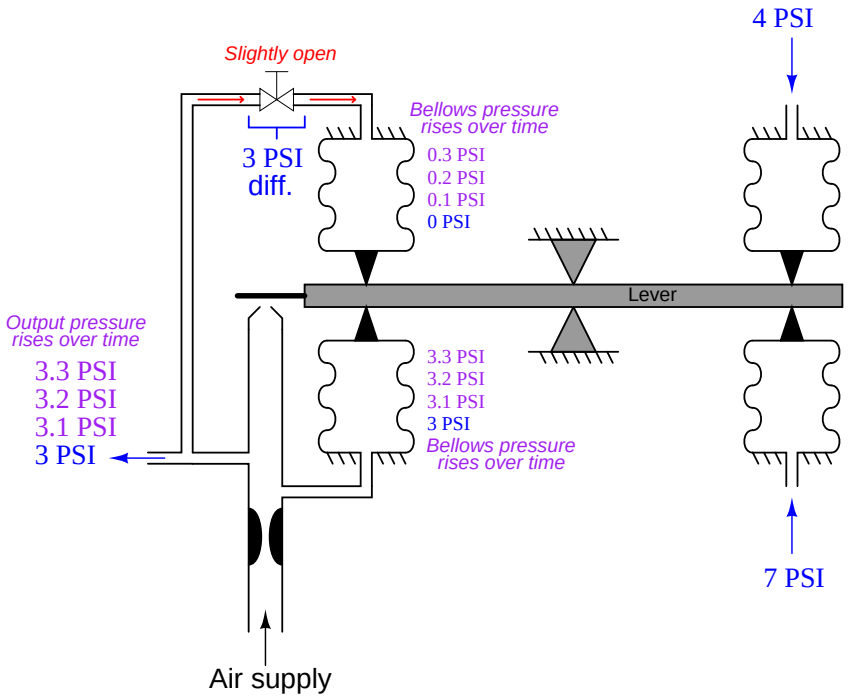
To better understand this integrating action, let us examine a simplified version of the controller. The following mechanism has been stripped of all unnecessary complexity so that we may focus on just the proportional and integral actions. Here, the PV and SP air pressure signals differ by 3 PSI, causing the force-balance mechanism to instantly respond with a 3 PSI output pressure to the feedback bellows (assuming a central fulcrum location, giving a controller gain of 1). The reset (integral) valve has been completely shut off to begin our analysis:



With 0 PSI of air pressure in the reset bellows, it is as though the reset bellows does not exist at all. The mechanism is a simple proportional-only pneumatic controller.

Now, imagine opening up the reset valve just a little bit, so that the output air pressure of 3 PSI begins to slowly fill the reset bellows. As the reset bellows fills with pressurized air, it begins to push down on the left-hand end of the force beam. This forces the baffle closer to the nozzle, causing the output pressure to rise. The regular output bellows has no restrictor valve to impede its filling, and so it *immediately* applies more upward force on the beam with the rising output pressure. With this greater output pressure, the reset bellows has an even greater "final" pressure to achieve, and so its rate of filling continues.

The result of these two bellows' opposing forces (one instantaneous, one time-delayed) is that the lower bellows must always stay 3 PSI *ahead* of the upper bellows in order to maintain a force-balanced condition with the two input bellows whose pressures differ by 3 PSI. This creates a constant 3 PSI differential pressure across the reset restriction valve, resulting in a constant flow of air into the reset bellows at a rate determined by that pressure drop and the opening of the restrictor valve. Eventually this will cause the output pressure to saturate at maximum, but until then the practical importance of this rising pressure action is that the mechanism now exhibits *integral control response* to the constant error between PV and SP:

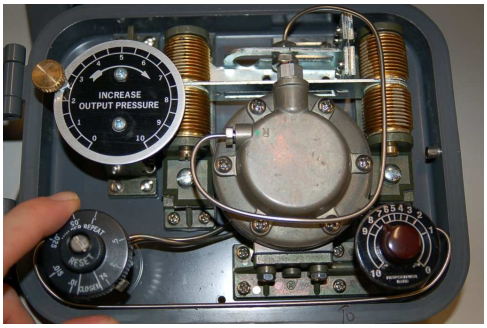


The greater the difference in pressures between PV and SP (i.e. the greater the *error*), the more pressure drop will develop across the reset restriction valve, causing the reset bellows to fill (or empty, depending on the sign of the error) with compressed air at a faster rate[9], causing the output pressure to change at a faster rate. Thus, we see in this mechanism the defining nature of integral control action: that the magnitude of the error determines the *velocity* of the output signal (its rate of change over time, or $\frac{dm}{dt}$). The rate of integration may be finely adjusted by changing the opening of the restrictor valve, or adjusted in large steps by connecting *capacity tanks* to the reset bellows to greatly increase its effective volume.

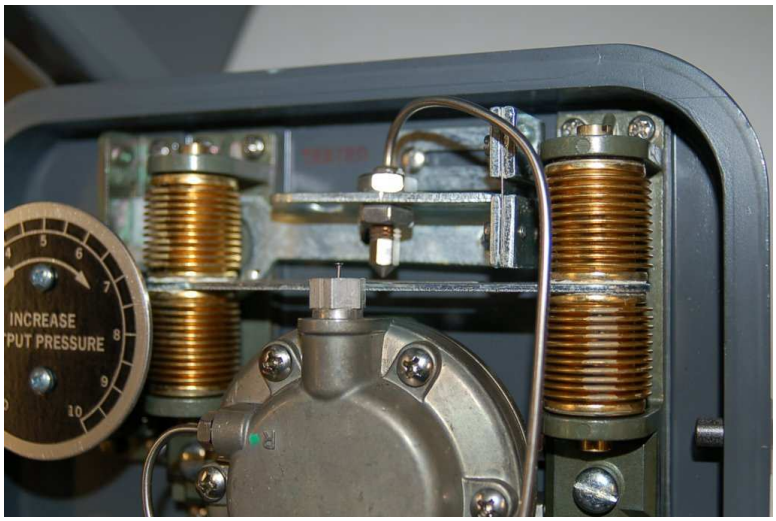
[9]These restrictor valves are designed to encourage laminar air flow, making the relationship between volumetric flow rate and differential pressure drop *linear* rather than quadratic as it is for large control valves. Thus, a doubling of pressure drop across the restrictor valve results in a doubling of flow rate into (or out of) the reset bellows, and a consequent doubling of integration rate. This is precisely what we desire and expect from a controller with integral action.

26.10.3 Fisher MultiTrol

Front (left) and rear (right) photographs of a real pneumatic controller (a Fisher "MultiTrol" unit) appear here:



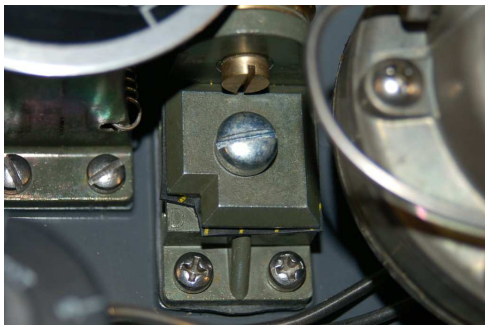
The mechanism is remarkably similar to the one used throughout the explanatory discussion, with the important distinction of being *motion-balance* instead of force balance. Proportional and integral control modes are implemented through the actions of four brass bellows pushing as opposing pairs at either end of a beam:

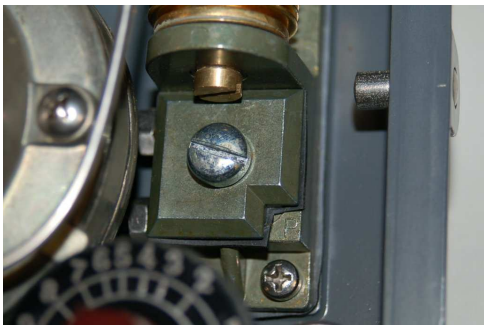


The nozzle may be seen facing down at the middle of the beam, with the center of the beam acting as a baffle. Setpoint control is achieved by moving the position of the nozzle up and down with respect to the beam. A setpoint dial (labeled "Increase Output Pressure") turns a cam which moves the nozzle closer to or further away from the beam. This being a motion-balance system, an offset in nozzle position equates to a biasing of the output signal, causing the controller to seek a new process variable value.

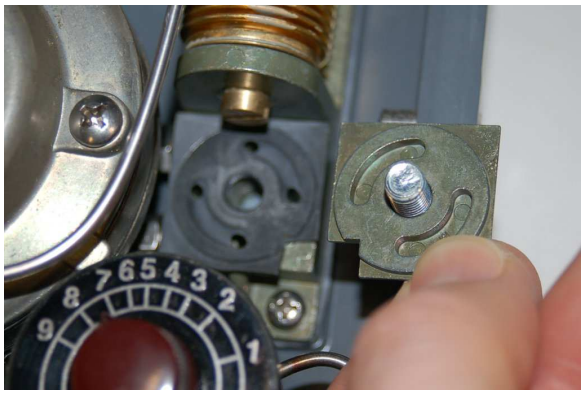
Instead of altering the position of a fulcrum to alter the gain (proportional band) of this controller, gain control is effected through the use of a "pressure divider" valve proportioning the amount of output air pressure sent to the feedback bellows. Integral rate control is implemented exactly the same way as in the hypothetical controller mechanism illustrated in the discussion: by adjusting a valve restricting air flow to and from the reset bellows. Both valves are actuated by rotary knobs with calibrated scales. The reset knob is actually calibrated in units of minutes per repeat, while the proportional band knob is labeled with a scale of arbitrary numbers:

Selection of direct versus reverse action is accomplished in the same way as selection between proportional and snap-action (on-off) control: by movable manifolds re-directing air pressure signals to different bellows in the mechanism. The direct/reverse manifold appears in the left-hand photograph (the letter "D" stands for *direct* action) while the proportional/snap manifold appears in the right-hand photograph (the letter "P" stands for *proportional* control):





Either setting is made by removing the screw holding the manifold plate to the controller body, rotating the plate one-quarter turn, and re-attaching. The following photograph shows one of the manifold plates removed and turned upside-down for inspection of the air passages:

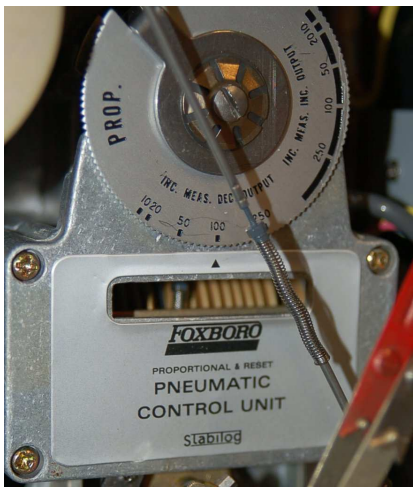


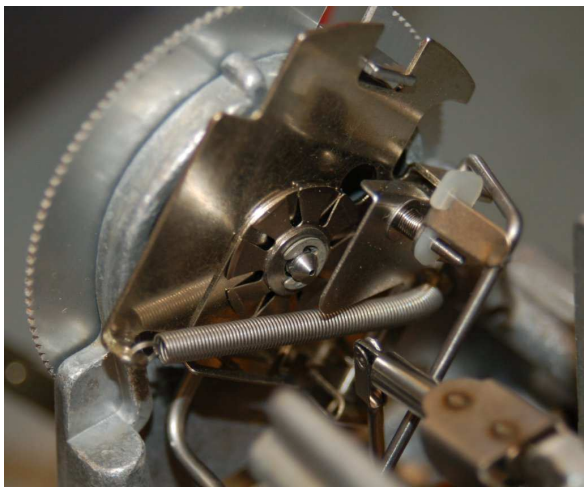
The two quarter-circumference slots seen in the manifold plate connect adjacent air ports together. Rotating the plate 90 degrees connects the four air ports together as two different pairs.

26.10.4 Foxboro model 43AP

The Fisher MultiTrol pneumatic controller is a very simple device, intended for field-mounting near the pneumatic transmitter and control valve to form a control loop for non-precision applications. A more sophisticated field-mounted pneumatic controller is the Foxboro model 43AP, sporting actual PV and SP indicating pointers, plus more precise tuning controls. The following photographs show one of these controllers, with the access door closed (left) and open (right):

At the heart of this controller is a motion-balance "pneumatic control unit" mechanism. A dial for setting proportional band (and direct/reverse action) appears on the front of the mechanism:





Note the simple way in which direct and reverse actions are described on this dial: either *increasing measurement, decreasing output* (reverse action) or *increasing measurement, increasing output* (direct action).

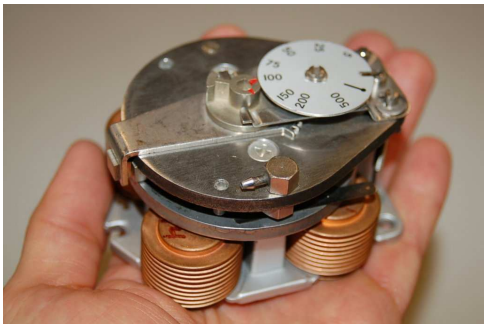
26.10.5 Foxboro model 130

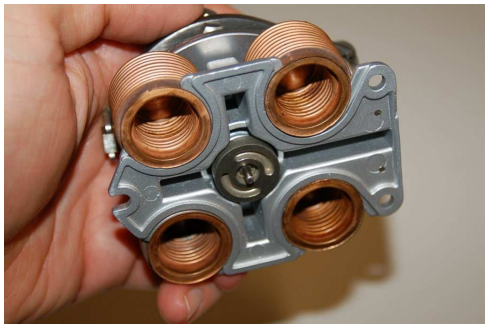
Foxboro also manufactured panel-mounted pneumatic controllers, the model 130 series, for larger-scale applications where multiple controllers needed to be located in one compact space. A bank of four Foxboro model 130 pneumatic controllers appears in the next photograph:

Each controller may be partially removed (slid out) from its slot in the rack, the P, I, and D settings adjustable on the left side panel with a screwdriver:

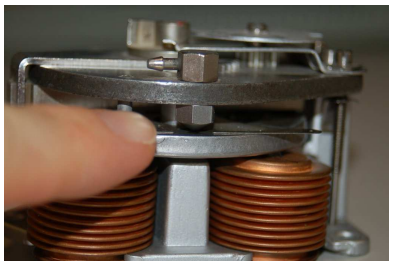
With the side panel removed, the entire mechanism is open to viewing:

The heart of the model 130 controller is a four-bellows force-balance mechanism, identical in principle to the hypothetical force-balance PID controller mechanism used throughout the explanatory discussion. Instead of the four bellows acting against a straight beam, however, these bellows push against a circular disk:

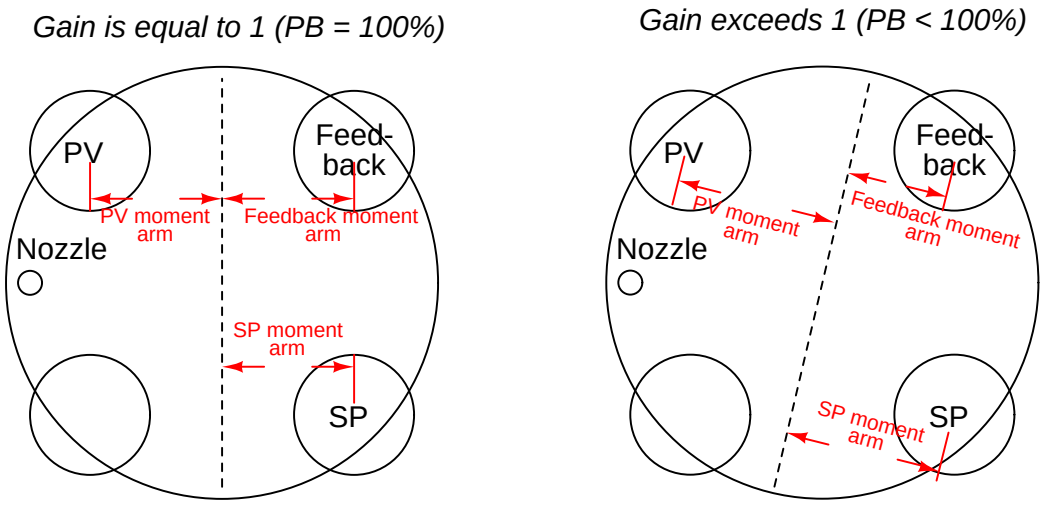




A nozzle (shown in the next photograph) detects if the disk is out of position (unbalanced), sending a back-pressure signal to an amplifying relay which then drives the feedback bellows:



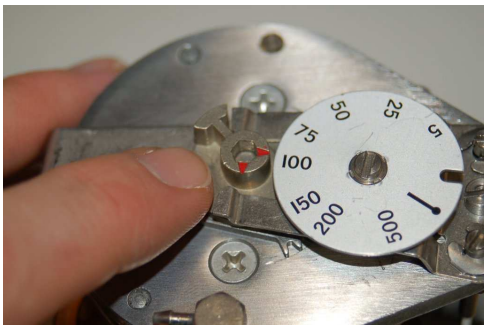
The disk rocks along an axis established by a movable bar. As this bar is rotated at different angles relative to the face of the disk, the fulcrum shifts with respect to the four bellows, providing a simple and effective gain adjustment:

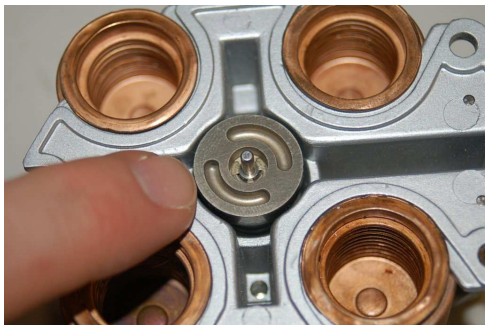


If the moment arms (lever lengths) between the input (PV and SP) bellows and the feedback bellows are equal, both sets of bellows will have equal leverage, and the gain will be one (a proportional band setting of 100%). However, if the fulcrum bar is rotated to give the input bellows more leverage and the feedback bellows less leverage, the feedback bellows will have to "work harder" (exert more force) to counteract any imbalance of force created by the input (PV and SP) bellows, thus creating a greater gain: more output pressure for the same amount of input pressure.

The fourth (lower-left) bellows acting on the disk provides an optional reset (integral) function. Its moment arm (lever length) of course is always equal to that of the feedback bellows, just as the PV and SP bellows' moment arm lengths are always equal, being positioned opposite the fulcrum line.

Selection between direct and reverse action works on the exact same principle as in the Fisher MultiTrol controller – by connecting four air ports in one of two paired configurations. A selector (movable with a hex wrench) turns an air signal port "switch" on the bottom of the four-bellows unit, effectively switching the PV and SP bellows:



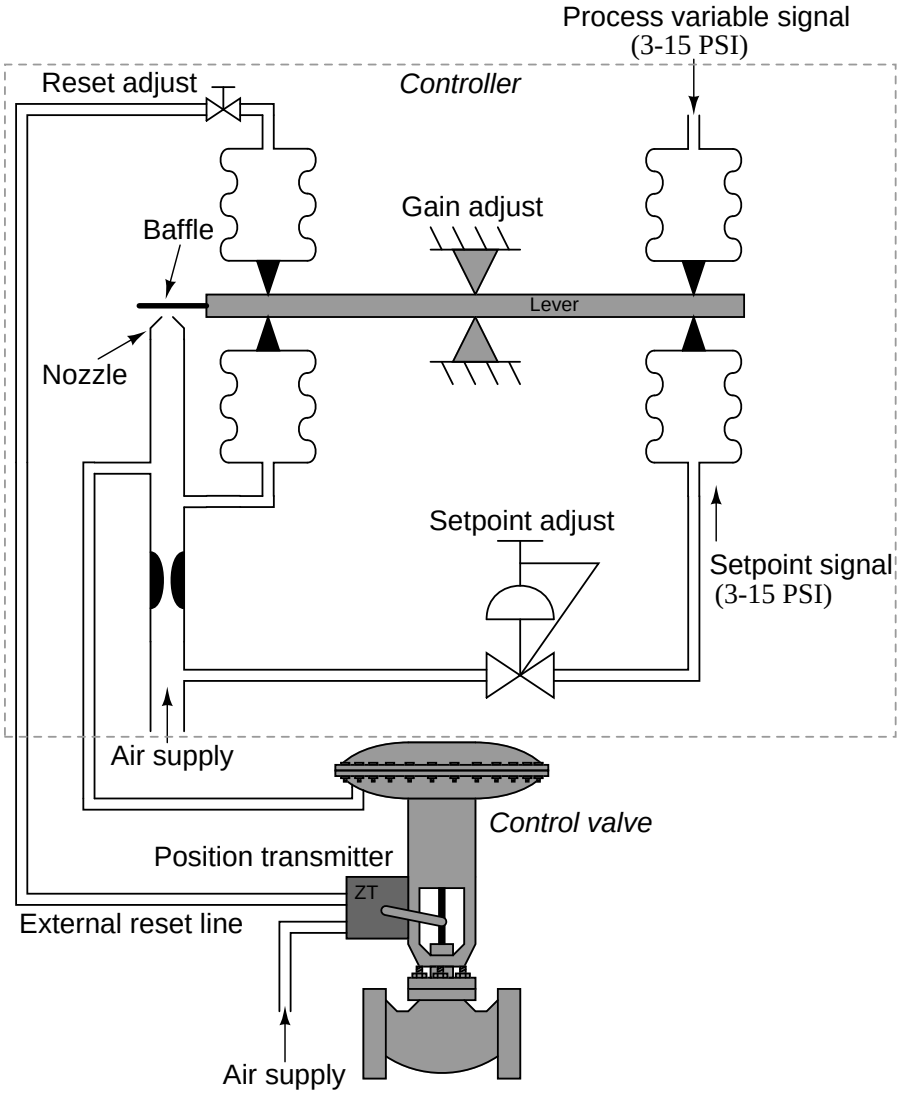


An interesting characteristic of most pneumatic controllers is modularity of function: it is possible to order a pneumatic controller that is proportional-only (P), proportional plus integral (P+I), or full PID. Since each control mode requires additional components to implement, a P-only pneumatic controller costs less than a P+I pneumatic controller, which in turn costs less than a full PID pneumatic controller. This explains the relative scarcity of full PID pneumatic controllers in industry: why pay for additional functionality if less will suffice for the task at hand?

26.10.6 External reset (integral) feedback

Some pneumatic controllers come equipped with an option for *external reset*: a feature useful in control systems to avoid integral windup if and when the process stops responding to changes in controller output. Instead of receiving a pneumatic signal directly from the output line of the controller, the reset bellows receives its signal through another pneumatic line, connected to a location in the control system where the final *effect* of the output signal (m) is seen. If for some reason the final control element cannot achieve the state called for by the controller, the controller will sense this through the external reset signal, and will cease integration to avoid "wind-up."

In the following illustration[10], the external reset signal comes from a pneumatic *position transmitter* (ZT) mounted to the sliding stem of the control valve, sending back a 3-15 PSI signal representing valve stem position:



If something happens to the control valve causing it to freeze position when the controller commands it to move – suppose the stem encounters a mechanical "stop" limiting travel, or a piece of solid material jams the valve trim so it cannot close further – the pneumatic pressure signal sent from the position transmitter to the controller's reset bellows will similarly freeze. After the

[10] In case you are wondering, this controller happens to be *reverse-acting* instead of direct. This is of no consequence to the feature of external reset.

pneumatic lag caused by the reset restrictor valve and bellows passes, the reset bellows force will remain fixed. This halts the controller's integral action, which was formerly based on a "race" between the output feedback bellows and the reset bellows, causing the feedback bellows to "lead" the reset bellows pressure by an amount proportional to the error between PV and SP. This "race" caused the output pressure to wind either up or down depending on the sign of the error. Now that the reset bellows pressure is frozen due to the control valve stem position being frozen, however, the "race" comes to an end and the controller exhibits only proportional action. Thus, the dreaded effect of integral windup – where the integral action of a controller continues to act even though the change in output is of no effect on the process – is averted.

26.11 Analog electronic PID controllers

Although analog electronic process controllers are considered a newer technology than pneumatic process controllers, they are actually "more obsolete" than pneumatic controllers. Panel-mounted (inside a control room environment) analog electronic controllers were a great improvement over panel-mounted pneumatic controllers when they were first introduced to industry, but they were superseded by digital controller technology later on. Field-mounted pneumatic controllers were either replaced by panel-mounted electronic controllers (either analog or digital) or left alone. Applications still exist for field-mounted pneumatic controllers, even now at the beginning of the 21^{st} century, but very few applications exist for analog electronic controllers in any location.

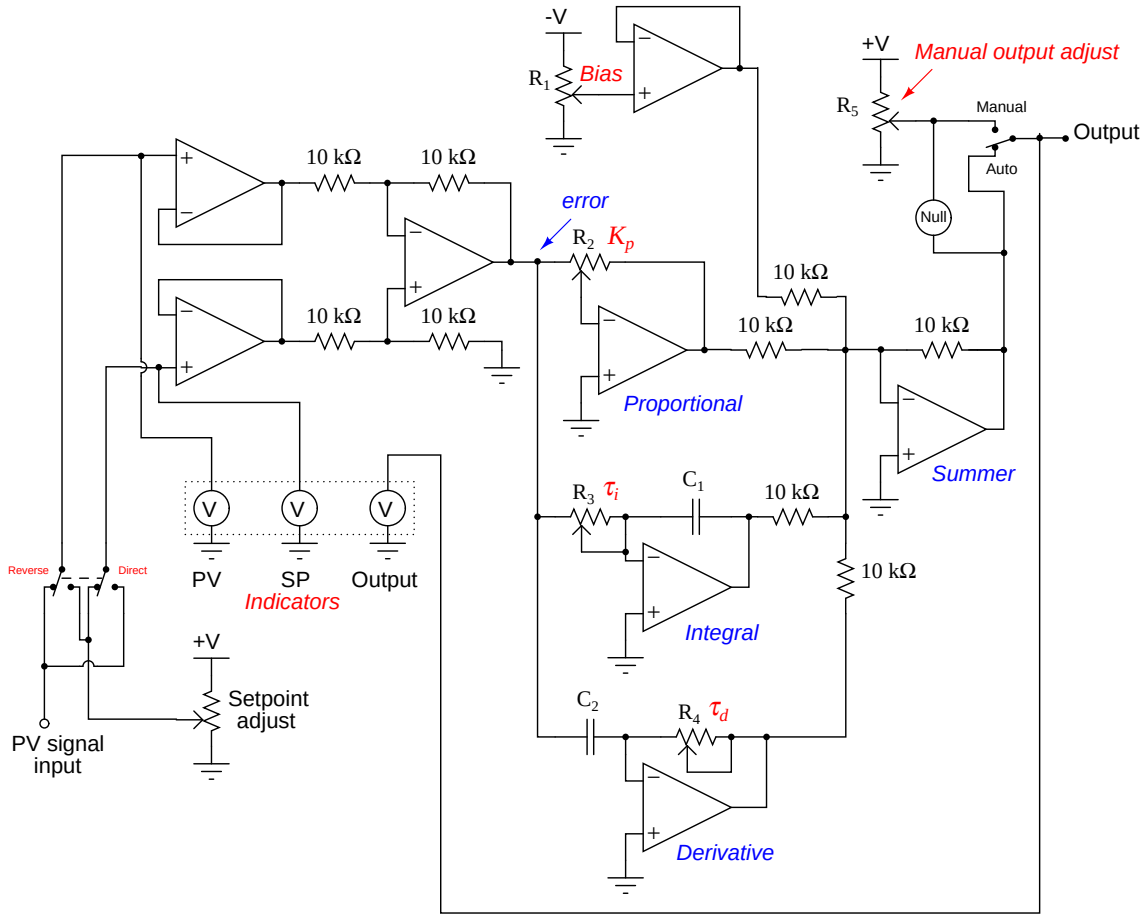
Analog electronic controllers enjoy only two advantages over digital electronic controllers: greater reliability and faster response. Now that digital industrial electronics has reached a very high level of reliability, the first advantage is academic, leaving only the second advantage for practical consideration. The advantage of faster speed may be fruitful in applications such as motion control, but for most industrial processes even the slowest digital controller is fast enough[11]. Furthermore, the numerous advantages offered by digital technology (data recording, networking capability, self-diagnostics, flexible configuration, function blocks for implementing different control strategies) severely weaken the relative importance of reliability and speed.

Most analog electronic PID controllers utilized operational amplifiers in their designs. It is relatively easy to construct circuits performing amplification (gain), integration, differentiation, summation, and other useful control functions with just a few op-amps, resistors, and capacitors.

[11] The real problem with digital controller speed is that the time delay between successive "scans" translates into dead time for the control loop. Dead time is the single greatest impediment to feedback control.

26.11.1 Circuit design

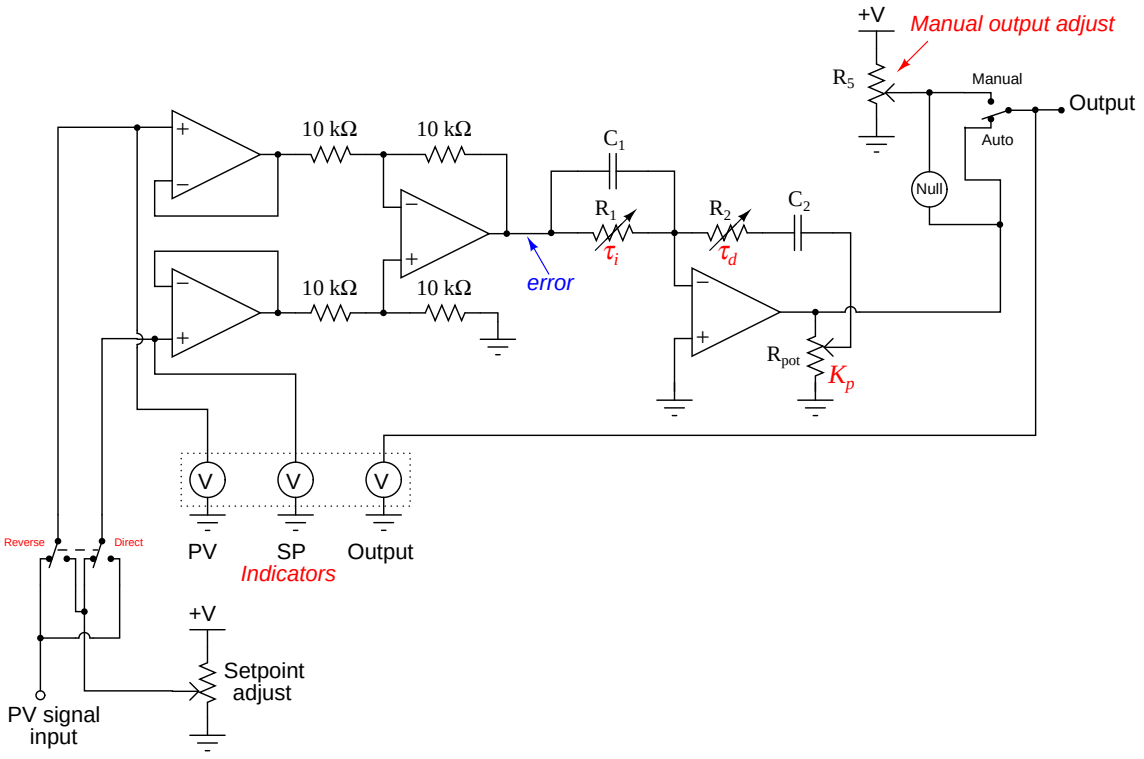
The following schematic diagram shows a full PID controller implemented using eight operational amplifiers, designed to input and output voltage signals representing PV, SP, and Output:



This controller implements the *parallel*, or *independent* PID algorithm, since each tuning adjustment (P, I, and D) act independently of each other:

$$m = K_p e + \frac{1}{\tau_i} \int e\, dt + \tau_d \frac{de}{dt} + b \qquad \textbf{Parallel PID equation}$$

It is possible to construct an analog PID controller with fewer operational amplifiers. An example is shown here:



As you can see, a *single* operational amplifier does all the work of calculating proportional, integral, *and* derivative responses. The first three amplifiers do nothing but buffer the input signals and calculate error (PV − SP, or SP − PV, depending on the direction of action).

This controller design happens to implement the *series* or *interacting* PID equation. Adjusting either the derivative or integral potentiometers also has an effect on the proportional (gain) value, and adjusting the gain of course has an effect on all terms of the PID equation:

$$m = K_p \left(e + \frac{1}{\tau_i} \int e \, dt \right) \left(1 + \tau_d \frac{d}{dt} \right) + b$$

Series or **Interacting PID equation**

It should be apparent to you now why analog controllers tend to implement the series equation instead of the parallel or ideal PID equations: they are simpler and less expensive to build that way.

26.11.2 Single-loop analog controllers

One popular analog electronic controller was the Foxboro model 62H, shown in the following photographs. Like the model 130 pneumatic controller, this electronic controller was designed to fit into a rack next to several other controllers. Tuning parameters were adjustable by moving potentiometer knobs under a side-panel accessible by partially removing the controller from its rack:

The Fisher corporation manufactured a series of analog electronic controllers called the AC^2, which were similar in construction to the Foxboro model 62H, but very narrow in width so that many could be fit into a compact panel space.

Like the pneumatic panel-mounted controllers preceding, and digital panel-mount controllers to follow, the tuning parameters for a panel-mounted analog electronic controller were typically accessed on the controller's side. The controller could be slid partially out of the panel to reveal the P, I, and D adjustment knobs (as well as direct/reverse action switches and other configuration controls).

Indicators on the front of an analog electronic controller served to display the process variable (PV), setpoint (SP), and manipulated variable (MV, or output) for operator information. Many analog electronic controllers did not have separate meter indications for PV and SP, but rather used a single meter movement to display the *error signal*, or difference between PV and SP. On the Foxboro model 62H, a hand-adjustable knob provided both indication and control over SP, while a small edge-reading meter movement displayed the error. A negative meter indication showed that the PV was below setpoint, and a positive meter indication showed that the PV was above setpoint.

The Fisher AC^2 analog electronic controller used the same basic technique, cleverly applied in such a way that the PV was displayed in real engineering units. The setpoint adjustment was a large wheel, mounted so the edge faced the operator. Along the circumference of this wheel was a scale showing the process variable range, from the LRV at one extreme of the wheel's travel to the URV at the other extreme of the wheel's travel. The actual setpoint value was the middle of the wheel from the operator's view of the wheel edge. A single meter movement needle traced an arc along the circumference of the wheel along this same viewable range. If the error was zero (PV = SP), the needle would be positioned in the middle of this viewing range, pointed at the same value along the scale as the setpoint. If the error was positive, the needle would rise up to point to a larger (higher) value on the scale, and if the error was negative the needle would point to a smaller (lower) value on the scale. For any fixed value of PV, this error needle would therefore move in exact step with

the wheel as it was rotated by the operator's hand. Thus, a single adjustment and a single meter movement displayed both SP and PV in very clear and unambiguous form.

Taylor manufactured a line of analog panel-mounted controllers that worked much the same way, with the SP adjustment being a graduated tape reeled to and fro by the SP adjustment knob. The middle of the viewable section of tape (as seen through a plastic window) was the setpoint value, and a single meter movement needle pointed to the PV value as a function of error. If the error happened to be zero (PV = SP), the needle would point to the middle of this viewable section of tape, which was the SP value.

Another popular panel-mounted analog electronic controller was the Moore Syncro, which featured plug-in modules for implementing different control algorithms (different PID equations, nonlinear signal conditioning, etc.). These plug-in function modules were a hardware precursor to the software "function blocks" appearing in later generations of digital controllers: a simple way of organizing controller functionality so that technicians unfamiliar with computer programming could easily configure a controller to do different types of control functions. Later models of the Syncro featured fluorescent bargraph displays of PV and SP for easy viewing in low-light conditions.

Analog single-loop controllers are largely a thing of the past, with the exception of some low-cost or specialty applications. An example of the former is shown here, a simple analog temperature controller small enough to fit in the palm of my hand:

This particular controller happened to be part of a sulfur dioxide analyzer system, controlling the internal temperature of a gas regulator panel to prevent vapors in the sample stream from condensing in low spots of the tubing and regulator system. The accuracy of such a temperature control application was not critical – if temperature was regulated to $\pm$ 5 degrees Fahrenheit it would be more than adequate. This is an application where an analog controller makes perfect sense: it is very compact, simple, extremely reliable, and inexpensive. None of the features associated with digital PID controllers (programmability, networking, precision) would have any merit in this application.

26.11.3 Multi-loop analog control systems

In contrast to single-loop analog controllers, *multi-loop* systems control dozens or even hundreds of process loops at a time. Prior to the advent of reliable digital technology, the only electronic process control systems capable of handling the numerous loops within large industrial installations such as power generating plants, oil refineries, and chemical processing facilities were analog systems, and several manufacturers produced multi-loop analog systems just for these large-scale control applications.

One of the most technologically advanced analog electronic products manufactured for industrial control applications was the Foxboro SPEC 200 system[12]. Although the SPEC 200 system used panel-mounted indicators, recorders, and other interface components resembling panel-mounted control systems, the actual control functions were implemented in a separate equipment rack which Foxboro called a *nest*[13]. Printed circuit boards plugged into each "nest" provided all the control functions (PID controllers, alarm units, integrators, signal selectors, etc.) necessary, with analog signal wires connecting the various functions together with panel-mounted displays and with field instruments to form a working system.

Analog field instrument signals (4-20 mA, or in some cases 10-50 mA) were all converted to a 0-10 VDC range for signal processing within the SPEC 200 nest. Operational amplifiers (mostly the model LM301) formed the "building blocks" of the control functions, with a +/- 15 VDC power supply providing DC power for everything to operate.

[12] Although the SPEC 200 system – like most analog electronic control systems – is considered obsolete, working installations may still be found at the time of this writing (2008). A report published by the Electric Power Research Institute (see References at the end of this chapter) in 2001 documents a SPEC 200 analog control system installed in a nuclear power plant in the United States as recently as 1992, and another as recently as 2001 in a Korean nuclear power plant.

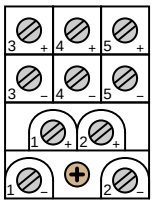
[13] Foxboro provided the option of a self-contained, panel-mounted SPEC 200 controller unit with all electronics contained in a single module, but the split architecture of the display/nest areas was preferred for large installations where many dozens of loops (especially cascade, feedforward, ratio, and other multi-component control strategies) would be serviced by the same system.

An an example of SPEC 200 technology, the following photographs show a model 2AX+A4 proportional-integral (P+I) controller card, inserted into a metal frame (called a "module" by Foxboro). This module was designed to fit into a slot in a SPEC 200 "nest" where it would reside alongside many other similar cards, each card performing its own control function:

Tuning and alarm adjustments may be seen in the right-hand photograph. This particular controller is set to a proportional band value of approximately 170, and an integral time constant of just over 0.01 minutes per repeat. A two-position rotary switch near the bottom of the card selected either reverse ("Dec") or direct ("Inc") control action.

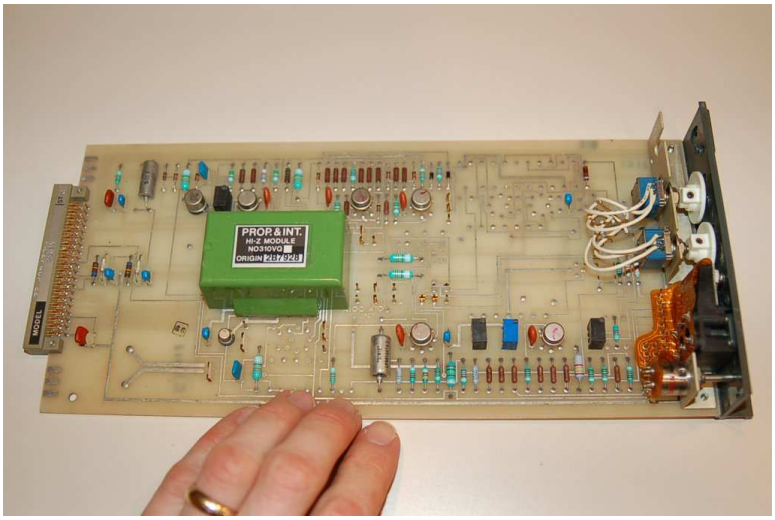
The array of copper pins at the top of the module form the male half of a cable connector, providing connection between the control card and the front-panel instrument accessible to operations personnel. Since the tuning controls appear on the face of this controller card (making it a "card tuned" controller), they were not accessible to operators but rather only to the technical personnel with access to the nest area. Other versions of controller cards ("control station tuned") had blank places where the P and I potentiometer adjustments appear on this model, with tuning adjustments provided on the panel-mounted instrument displays for easier access to operators.

The set of ten screw terminals at the bottom of the module provided connection points for the input and output voltage signals. The following list gives the general descriptions of each terminal pair, with the descriptions for this particular P + I controller written in *italic* type:



- Terminals (1+) and (1-): Input signal #1 (*Process variable input*)
- Terminals (2+) and (2-): Output signal #1 (*Manipulated variable output*)
- Terminals (3+) and (3-): Input #2, Output #4, or Option #1 (*Remote setpoint*)
- Terminals (4+) and (4-): Input #3, Output #3, or Option #2 (*Optional alarm*)
- Terminals (5+) and (5-): Input #4, Output #2, or Option #3 (*Optional 24 VAC*)

A photograph of the printed circuit board (card) removed from the metal module clearly shows the analog electronic components:



Foxboro went to great lengths in their design process to maximize reliability of the SPEC 200 system, already an inherently reliable technology by virtue of its simple, analog nature. As a result, the reliability of SPEC 200 control systems is the stuff of legend[14].

[14] I once encountered an engineer who joked that the number "200" in "SPEC 200" represented the number of years the system was designed to continuously operate. At another facility, I encountered instrument technicians who were a bit afraid of a SPEC 200 system running a section of their plant: the system had *never suffered a failure of any*

26.12 Digital PID controllers

The vast majority of PID controllers in service today are digital in nature. Microprocessors executing PID algorithms provide many advantages over any form of analog PID control (pneumatic or electronic), not the least of which being the ability to network with personal computer workstations and other controllers over wired or wireless (radio) networks.

26.12.1 Stand-alone digital controllers

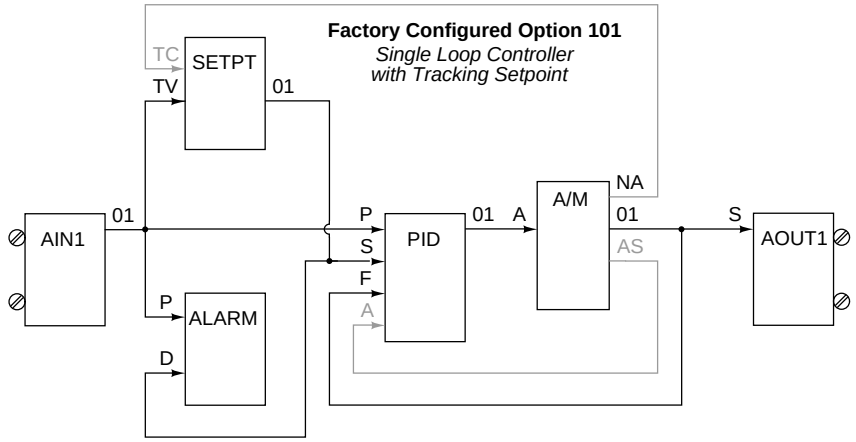
If the internal components of a panel-mounted pneumatic or analog electronic controller (such as the Foxboro models 130 or 62, respectively) were completely removed and replaced by all-digital electronic componentry, the result would be a *stand-alone digital PID controller*. From the outside, such a digital controller looks very similar its technological ancestors, but its capabilities are far greater.

An example of a popular panel-mounted digital controller is the Siemens model 353 (formerly the Moore Products model 353):

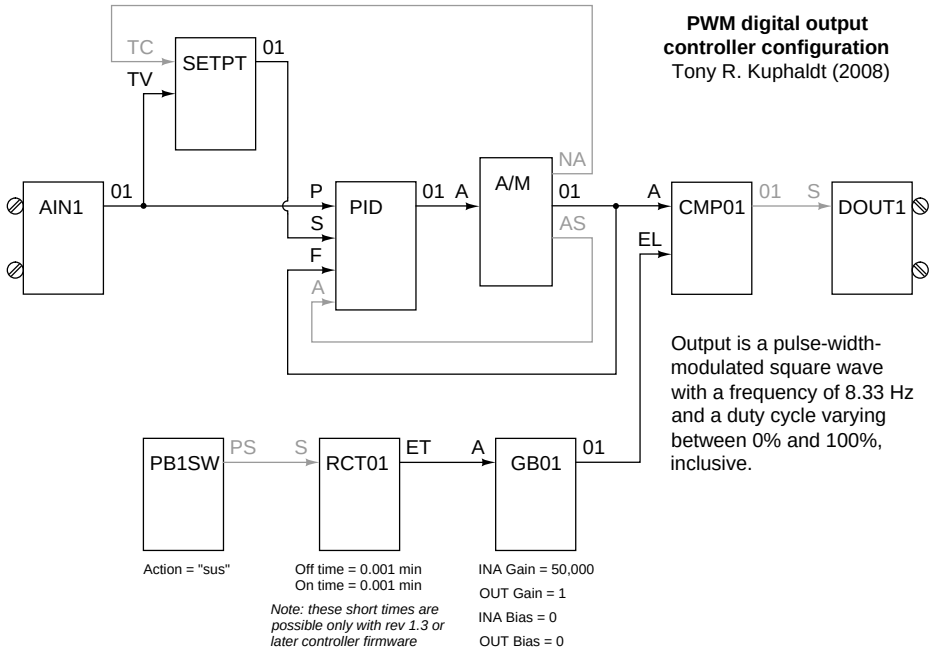
This particular controller, like many high-end digital controllers and larger digital control systems, is programmed in a function block language. Each function block in the controller is a software subroutine performing a specific function on input signals, generating at least one output signal. Each function block has a set of configuration parameters telling it how to behave. For example, the PID function block in a digital controller would have parameters specifying direct or reverse action, gain (K_p), integral time constant (τ_i), derivative time constant (τ_d), output limits, etc.

kind since it was installed decades ago, and as a result no one in the shop had any experience troubleshooting it. As it turns out, the entire facility was eventually shut down and sold, with the SPEC 200 nest running faithfully until the day its power was turned off!

Even the "stock" configuration for simple, single-loop PID control is a collection of function blocks linked together:



The beauty of function block programming is that the same blocks may be easily re-linked to implement custom control strategies. Take for instance the following function block program written for a Siemens model 353 controller to provide a pulse-width-modulation (PWM, or time-proportioned) output signal instead of the customary 4-20 mA DC analog output signal. The application is for an electric oven temperature control system, where the oven's heating element could only be turned on and off fully rather than continuously varied:



In order to specify links between function blocks, each of the used lettered block inputs is mapped to the output channel of another block. In the case of the time-proportioned function block program, for example, the "P" (process variable) input of the PID function block is set to get its signal from the "01" output channel of the AIN1 (analog input 1) function block. The "TV" (tracking value) input of the SETPT (setpoint) function block is also set to the "01" output channel of the AIN1 function block, so that the setpoint value generator has access to the process variable value in order to implement setpoint tracking. Any function block output may drive an unlimited number of function block inputs (fan-out), but each function block input may receive a signal from *only one* function block output. This is a rule followed within all function block languages to prevent multiple block output signals from conflicting (attempting to insert different signal values into the same input).

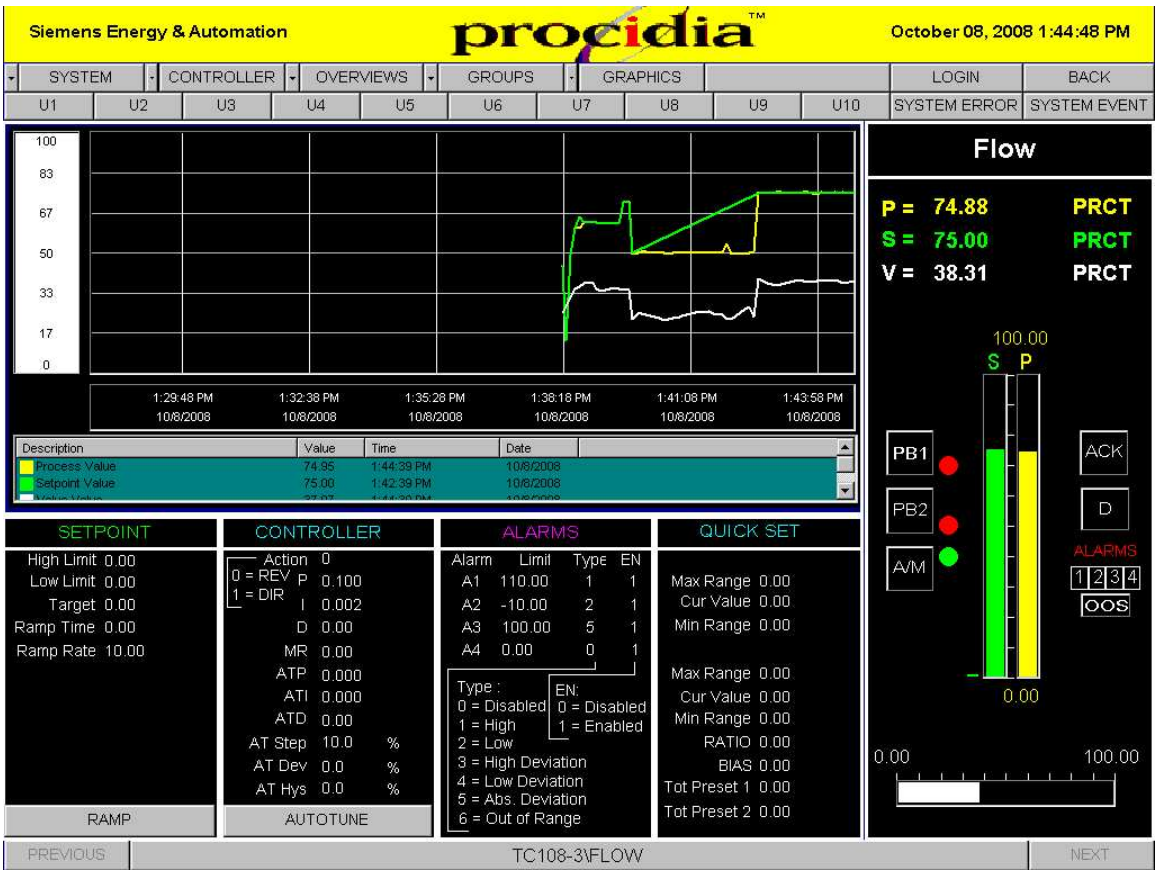
In the Siemens controllers, function block programming may be done by entering configuration data using the front-panel keypad, or by using graphical software running on a personal computer networked with the controller.

For applications not requiring so much capability, and/or requiring a smaller form factor, other panel-mounted digital controllers exist. The Honeywell model UDC3000 is a popular example of a 1/4 DIN (96 mm × 96 mm) size digital controller:

Even smaller panel-mounted controllers are produced by a wide array of manufacturers for applications requiring minimum functionality: 1/8 DIN (96 mm × 48 mm), 1/16 DIN (48 mm × 48 mm), and even 1/32 DIN (48 mm × 24 mm) sizes are available.

One of the advantageous capabilities of modern stand-alone controllers is the ability to exchange data over digital networks. This provides operations and maintenance personnel alike the ability to remotely monitor and even control (adjust setpoints, switch modes, change tuning parameters, etc.) the process controller from a computer workstation. The Siemens model 353 controller (with appropriate options) has the ability to digitally network over Ethernet, a very common and robust digital network standard. The following photographs show three such controllers connected to a network through a common 4-port Ethernet "hub" device:

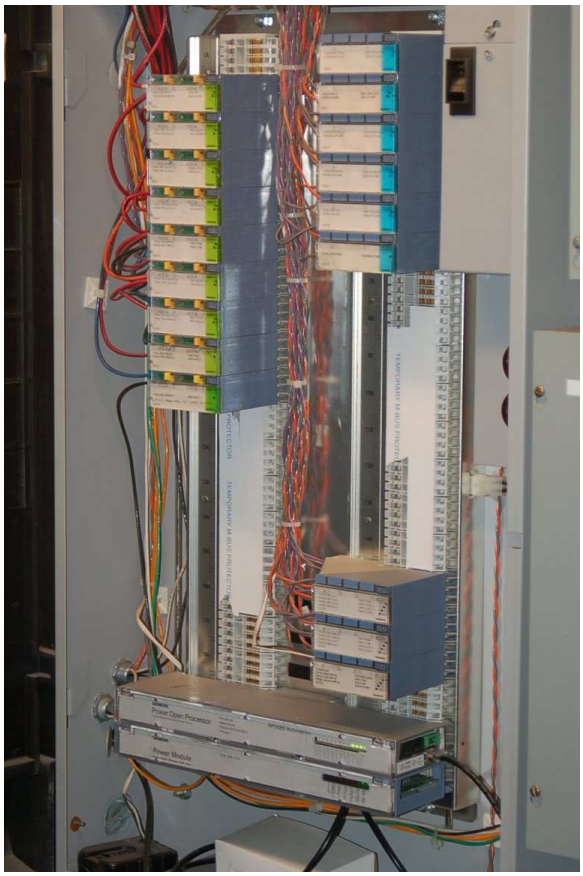
Special software (in this case, Siemens *Procidia*) running on a computer workstation connected to the same Ethernet network acquires data from and sends data to the networked controllers. Screenshots of this software show typical displays allowing complete control over the function of the process controllers:



26.12.2 Direct digital control (DDC)

A microprocessor operating at sufficient clock speed is able to execute more than one PID control algorithm for a process loop, by "time-sharing" its calculating power: devoting slices of time to the evaluation of each PID equation in rapid succession. This not only makes multiple-loop digital control possible for a single microprocessor, but also makes it very attractive given the microprocessor's natural ability to manage data archival, transfer, and networking. A single computer is able to execute PID control for multiple loops, and also make that loop control data accessible between loops (for purposes of cascade, ratio, feedforward, and other control strategies) and accessible on networks for human operators and technicians to easily access.

Such *direct digital* control (DDC) has been applied with great success to the problem of building automation, where temperature and humidity controls for large structures benefit from large-scale data integration. The following photograph shows a Siemens APOGEE building automation system with multiple I/O (input/output) cards providing interface between analog instrument signals and the microprocessor's digital functions:



A close-up view of the processor shows the device handling all mathematical calculations for the PID control:

Other than a few LEDs, there is no visual indication in this panel of what the system is doing at any particular time. Operators, engineers, and technicians alike must use software running on a networked personal computer to access data in this control system.

A smaller-scale example of a DDC system is the Delta model DSC-1280 controller, an example shown in the following photograph:

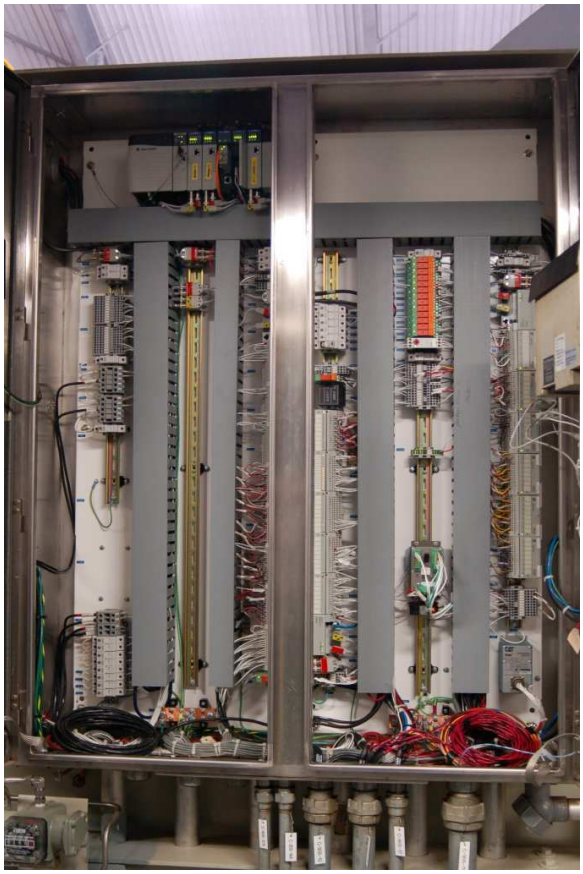
This system does not have plug-in I/O cards like the Siemens APOGEE, but instead is monolithic in design, with all inputs and outputs part of one large "motherboard" PCB. The model DSC-1280 controller has 12 input channels and 8 output channels (hence the model number "1280"). An Ethernet cable (RJ-45 plug) is seen in the upper-left corner of this unit, through which a remotely-located personal computer communicates with the DDC using a high-level protocol called BACnet. In many ways, BACnet is similar to Modbus, residing at layer 7 of the OSI Reference Model (the so-called *Application Layer*), unconcerned with the details of data communication at the Physical or Data Link layers. This means, like Modbus, BACnet commands may be sent and received over a variety of lower-level network standards, with Ethernet being the preferred[15] option at the time of this writing.

A more common application of industrial direct digital control is to use a programmable logic

[15]Thanks to the explosion of network growth accompanying personal computers in the workplace, Ethernet is ubiquitous. The relatively high speed and low cost of Ethernet communications equipment makes it an attractive network standard over which a great many high-level industrial protocols communicate.

controller (PLC) as a PID controller for multiple loops. PLCs were originally invented for on/off (discrete) process control functions, but have subsequently grown in speed and capability to handle analog PID control functions as well. A PLC is often a less expensive option for PID control than a stand-alone controller, which explains the prevalence of this control philosophy in many industrial environments.

This next photograph shows an Allen-Bradley (Rockwell) ControlLogix PLC used to control the operation of a gas turbine engine. The PLC may be seen in the upper-left corner of the enclosure, with the rest of the enclosure devoted to terminal blocks and accessory components:



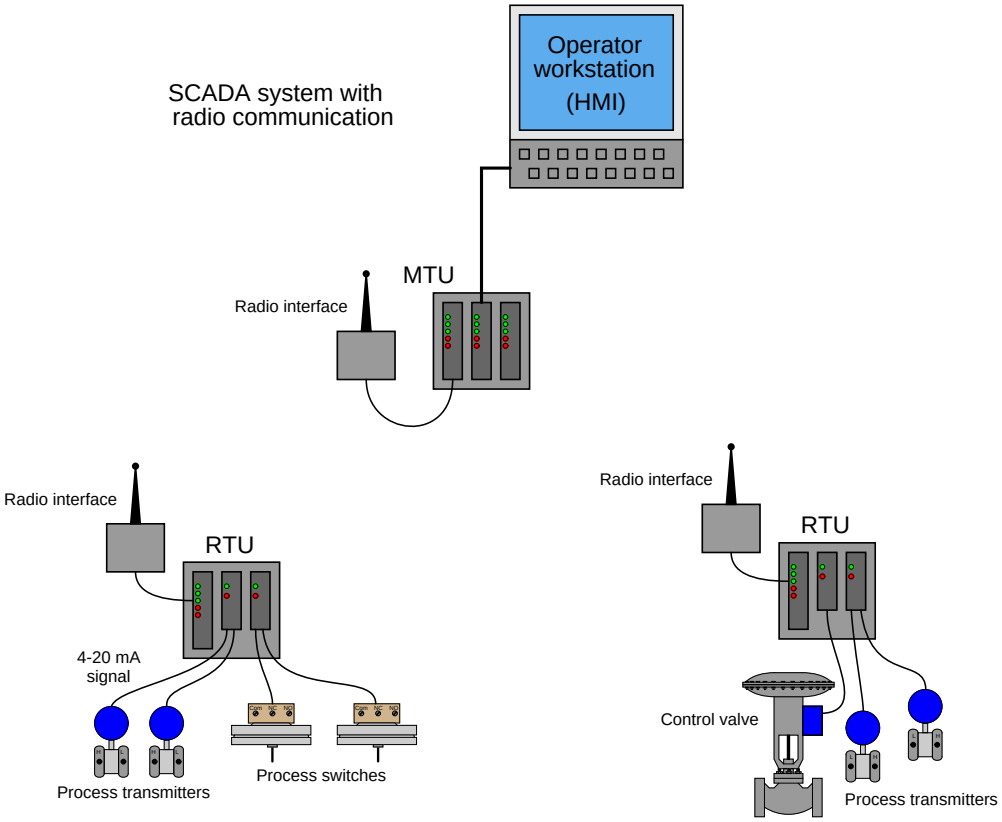
A strong advantage of using PLCs for analog loop control is the ability to easily integrate discrete controls with the analog controls. It is quite easy, for example, to coordinate the sequential start-up and shut-down functions necessary for intermittent operation with the analog PID controls necessary for continuous operation, all within one programmable logic controller. It should be noted, however, that many early PLC implementations of PID algorithms were crude at best, lacking the finesse of stand-alone PID controllers. Even some modern PLC analog functions are mediocre[16] at the time of this writing (2008).

[16] An aspect common to many PLC implementations of PID control is the use of the "parallel" PID algorithm instead of the superior "ISA" or "non-interacting" algorithm. The choice of algorithm may have a profound effect on tuning, and on tuning procedures, especially when tuning parameters must be re-adjusted to accommodate changes in transmitter range.

26.12.3 SCADA and telemetry systems

A similar control system architecture to Direct Digital Control (DDC) – assigning a single microprocessor to the task of managing multiple control functions, with digital communication between the microprocessor units – is used for the management of systems which are by their very nature spread over wide geographical regions. Such systems are generally referred to as *SCADA*, which is an acronym standing for *Supervisory Control And Data Acquisition*.

The typical SCADA system consists of multiple *Remote Terminal Unit* (RTU) devices connected to process transmitters and final control elements, implementing basic control functions such as motor start/stop and PID loop control. These RTU devices communicate digitally to a *Master Terminal Unit* (MTU) device at a central location where human operators may monitor the process and issue commands.

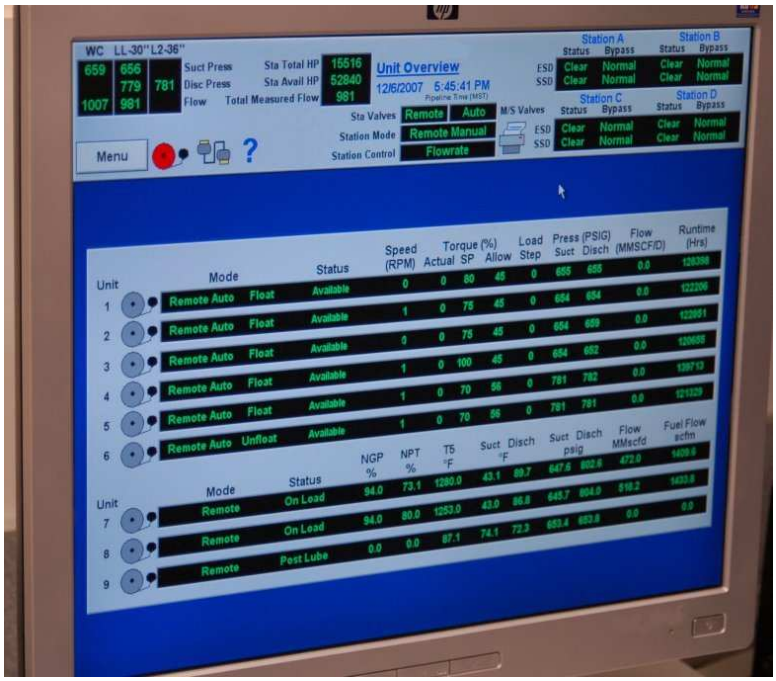


A photograph of an RTU "rack" operating at a large electric power substation is shown here:

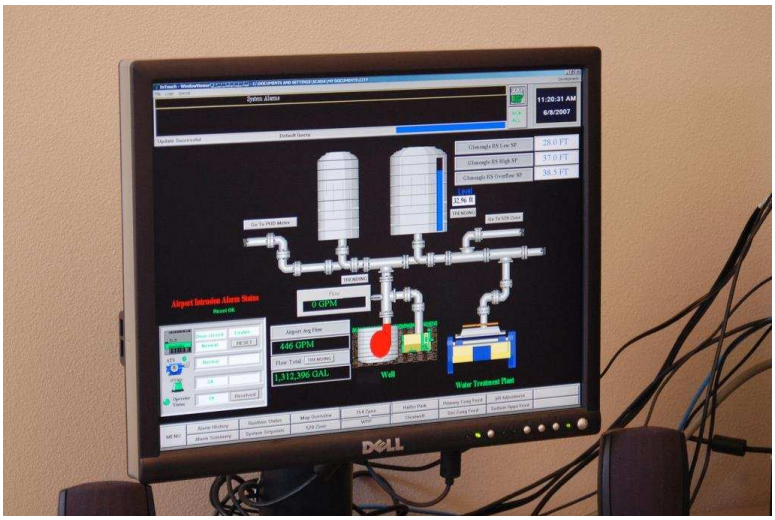
Some RTU hardware, such as the substation monitoring system shown above, is custom-manufactured for the application. Other RTU hardware is more general in purpose, such as the ROC 800 manufactured by Fisher, intended for the monitoring and control of natural gas and oil production wells, but applicable to other applications as well. The Fisher ROC units are designed to operate with a minimum of electrical power, so that a single solar panel and battery will be sufficient for year-round operation in remote environments. Radio communication is a standard feature for ROC units, as there are no runs of communications cable linking remote wells separated by dozens of miles.

Standard programmable logic controllers (PLCs) are ideal candidates for use as RTU devices. Modern PLCs have all the I/O, networking, and control algorithm capability necessary to function as remote terminal units. Commercially available Human-Machine Interface (HMI) software allowing personal computers to display PLC variable values potentially turns every PC into a Master Terminal Unit (MTU) where operators can view process variables, change setpoints, and issue other commands for controlling the process.

A photograph of such HMI software used to monitor a SCADA system for a set of natural gas compressors is shown here:



Another photograph of a similar system used to monitor and control drinking water reservoirs for a city is shown here:



A concept closely related to SCADA is that of *telemetry*, the word literally meaning "distance measuring" (i.e. measuring something over a distance). The acronym SCADA, by containing the word "control," implies two-way communication (measurement and control) between the master location and the remote location. In applications where the flow of information is strictly one-way (simplex) from the remote location to the master location, "telemetry" is a more apt description.

Telemetry systems find wide application in scientific research. Seismographs, river and stream flowmeters, weather stations, and other remotely-located measurement instruments connected (usually by radio links) to some centralized data collection center are all examples of telemetry. Any industrial measurement (-only) application spanning a large distance could likewise be classified as a telemetry system, although you will sometimes find the term "SCADA" applied even when the communication is simplex in nature.

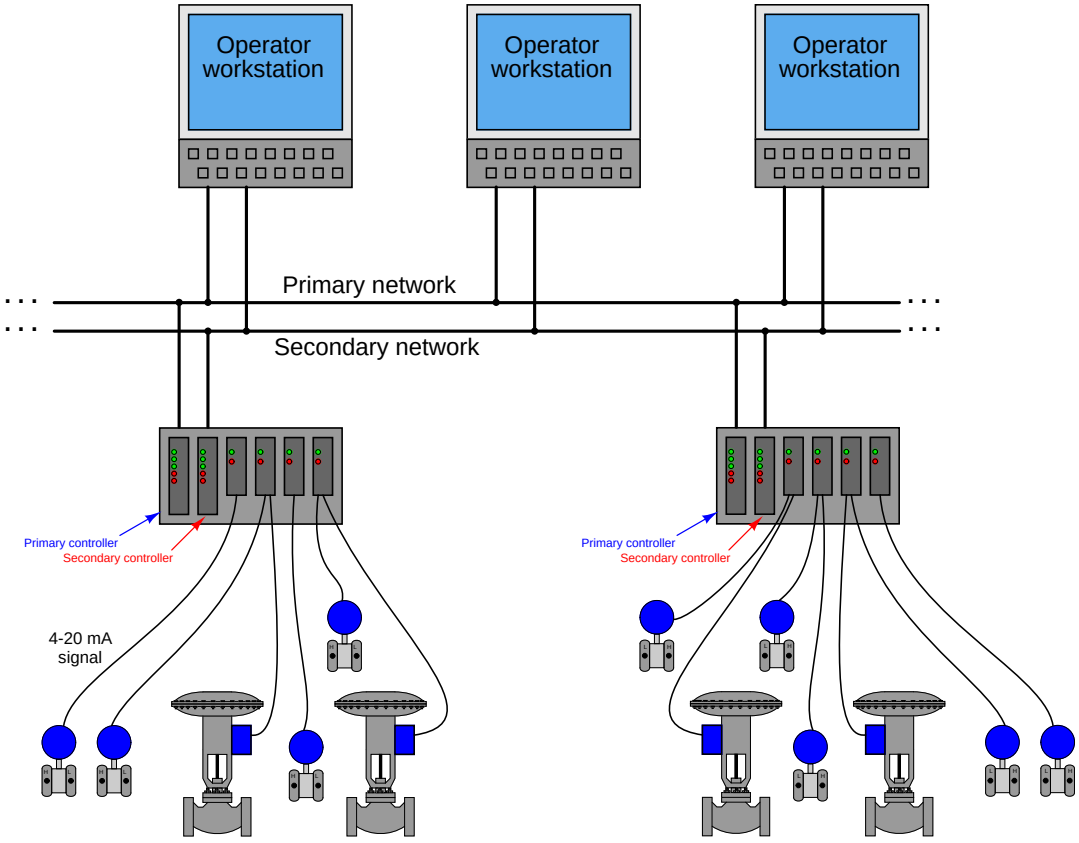
26.12.4 Distributed Control Systems (DCS)

A radically new concept appeared in the world of industrial control in the mid-1970's: the notion of *distributed* digital control. Direct digital control during that era[17] suffered a substantial problem: the potential for catastrophic failure if the single digital computer executing *multiple* PID control functions were to ever halt. Digital control brings many advantages, but it isn't worth the risk if the entire operation will shut down (or catastrophically fail!) following a hardware or software failure within that one computer.

Distributed control directly addressed this concern by having multiple control computers – each one responsible for only a handful of PID loops – distributed throughout the facility and networked together to share information with each other and with operator display consoles. With individual process control "nodes" scattered throughout the campus, each one dedicated to controlling just a few loops, there would be less concentration of liability as there would be with a single-computer DDC system. Such distribution of computing hardware also shortened the analog signal wiring, because now the hundreds or thousands of analog field instrument cables only had to reach as far as the distributed nodes, not all the way to a centralized control room. Only the networking cable had to reach that far, representing a drastic reduction in wiring needs. Furthermore, distributed control introduced the concept of *redundancy* to industrial control systems: where digital signal acquisition and processing hardware units were equipped with "spare" units designed to automatically take over all critical functions in the event of a primary failure.

[17] Modern DDC systems of the type used for building automation (heating, cooling, security, etc.) almost always consist of networked control nodes, each node tasked with monitoring and control of a limited area. The same may be said for modern PLC technology, which not only exhibits advanced networking capability (fieldbus I/O networks, Ethernet, Modbus, wireless communications), but is often also capable of redundancy in both processing and I/O. As technology increases in sophistication, the distinction between a DDC (or a networked PLC system) and a DCS becomes more ambiguous.

The following illustration shows a typical distributed control system (DCS) architecture:



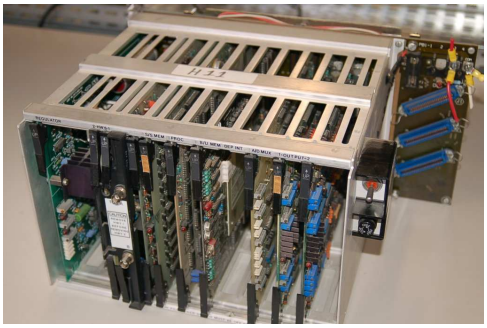
Each "rack" contains a microprocessor to implement all necessary control functions, with individual I/O (input/output) "cards" for converting analog field instrument signals into digital format, and visa-versa. Redundant processors, redundant network cables, and even redundant I/O cards address the possibility of component failure. DCS processors are usually programmed to perform routine self-checks[18] on redundant system components to ensure availability of the spare components in the event of a failure.

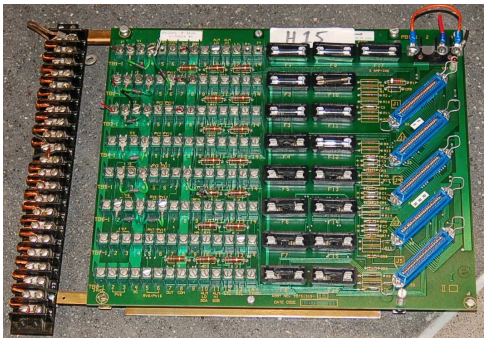
If there ever was a total failure in one of the "control racks" where the redundancy proved insufficient for the fault(s), the only PID loops faulted will be those resident in that rack, not any of the other loops throughout the system. Likewise, if ever the network cables become severed or otherwise faulted, only the information flow between those two points will suffer; the rest of the system will continue to communicate data normally. Thus, one of the "hallmark" features of a DCS is its tolerance to serious faults: even in the event of severe hardware or software faults, the impact to process control is minimized by design.

[18] An example of such a self-check is scheduled switching of the networks: if the system has been operating on network cable "A" for the past four hours, it might switch to cable "B" for the next four hours, then back again after another four hours to continually ensure both cables are functioning properly.

One of the very first distributed control systems in the world was the Honeywell TDC2000 system[19], introduced in 1975. By today's standards, the technology was crude[20], but the concept was revolutionary.

Each rack (called a "box" by Honeywell) consisted of an aluminum frame holding several large printed circuit boards with card-edge connectors. A "basic controller" box appears in the left-hand photograph. The right-hand photograph shows the termination board where the field wiring (4-20 mA) connections were made. A thick cable connected each termination board to its respective controller box:





Controller redundancy in the TDC2000 DCS took the form of a "spare" controller box serving as a backup for up to eight other controller boxes. Thick cables routed all analog signals to this spare controller, so that it would have access to them in the event it needed to take over for a failed controller. The spare controller would become active on the event of *any* fault in any of the (up to eight) other controllers, including failures in the I/O cards. Thus, this redundancy system provided for processor failures as well as I/O failures. All TDC2000 controllers communicated digitally by means of a dual coaxial cable network known as the "Data Hiway." The dual cables provided redundancy in network communications.

[19] To be fair, the Yokogawa Electric Corporation of Japan introduced their CENTUM distributed control system the same year as Honeywell. Unfortunately, while I have personal experience maintaining and using the Honeywell TDC2000 system, I have zero personal experience with the Yokogawa CENTUM system, and neither have I been able to obtain technical documentation for the original incarnation of this DCS (Yokogawa's latest DCS offering goes by the same name). Consequently, I can do little in this chapter but mention its existence, despite the fact that it deserves just as much recognition as the Honeywell TDC2000 system.

[20] Just to give some perspective, the original TDC2000 system used whole-board processors rather than microprocessor chips, and magnetic core memory rather than static or dynamic RAM circuits! Communication between controller nodes and operator stations was handled by thick coaxial cables, implementing master/slave arbitration with a separate device (a "Hiway Traffic Director" or HTD) coordinating all communications between nodes. Like Bob Metcalfe's original version of Ethernet, these coaxial cables were terminated at their end-points by termination resistors, with coaxial "tee" connectors providing branch points for multiple nodes to connect along the network.

A typical TDC2000 operator workstation appears in the next photograph:

Over the years following its 1975 introduction, the Honeywell system grew in sophistication with faster networks (the "Local Control Network" or LCN), more capable controller racks (the "Process Manager" or PM series), and better operator workstations. Many of these improvements were incremental, consisting of add-on components that could work with existing TDC2000 components so that the entire system need not be replaced to accept the new upgrades.

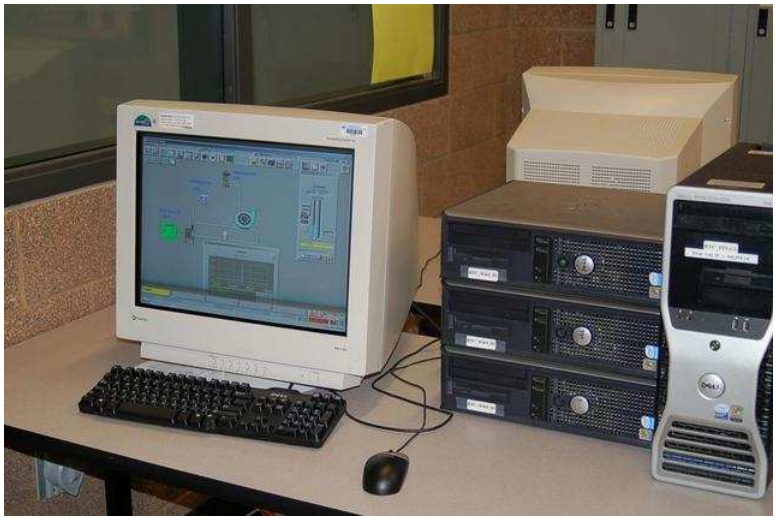
Other control equipment manufacturers responded to the DCS revolution started by Honeywell and Yokogawa by offering their own distributed control systems. The Bailey Network 90 (Net90) DCS, Bailey Infi90 DCS, and the Fisher Provox systems are examples. Foxboro, already an established leader in the control system field with their SPEC 200 analog system, first augmented the SPEC 200 with digital capabilities (the VIDEOSPEC workstation consoles, FOX I/A computer, INTERSPEC and FOXNET data networks), then developed an entirely digital distributed control system, the SPECTRUM.

Some modern distributed control systems offered at the time of this writing (2008) include:

- ABB *800xA*
- Emerson *DeltaV* and *Ovation*
- Foxboro (Invensys) *I/A*
- Honeywell *Experion PKS*
- Yokogawa *CENTUM VP* and *CENTUM CS*

For a visual comparison with the Honeywell TDC2000 DCS, examine the following photograph of an Emerson DeltaV DCS rack, with processor and multiple I/O modules:

Many modern distributed control systems such as the Emerson DeltaV use regular personal computers rather than proprietary hardware as operator workstations. This cost-saving measure leverages existing computer and display technologies without sacrificing control-level reliability (since the control hardware and software is still industrial-grade):



As previously mentioned in the Direct Digital Control (DDC) subsection, programmable logic controllers (PLCs) are becoming more and more popular as PID control platforms due to their ever-expanding speed, functionality, and relatively low cost. It is now possible with modern PLC hardware and networking capabilities to build a truly distributed control system with individual

PLCs as the processing nodes, and with redundancy built into each of those nodes so that any single failure does not interrupt critical control functions. Such a system may be purchased at a fraction of the up-front cost of a fully-fledged DCS.

However, what is currently lacking in the PLC world is the same level of hardware and software integration necessary to build a functional distributed control system that comes as ready-to-use as a system pre-built by a DCS manufacturer. In other words, if an enterprise chooses to build their own distributed control system using programmable logic controllers, they must be prepared to do a *lot* of programming work in order to emulate the same level of functionality and power as a pre-engineered DCS[21]. Any engineer or technician who has experienced the power of a modern DCS – with its self-diagnostic, "smart" instrument management, event auditing, advanced control strategy, pre-engineered redundancy, data collection and analysis, and alarm management capabilities – realizes these features are neither luxuries nor are they trivial to engineer. Woe to anyone who thinks these critical features may be created by incumbent staff at a lesser cost!

26.12.5 Fieldbus control

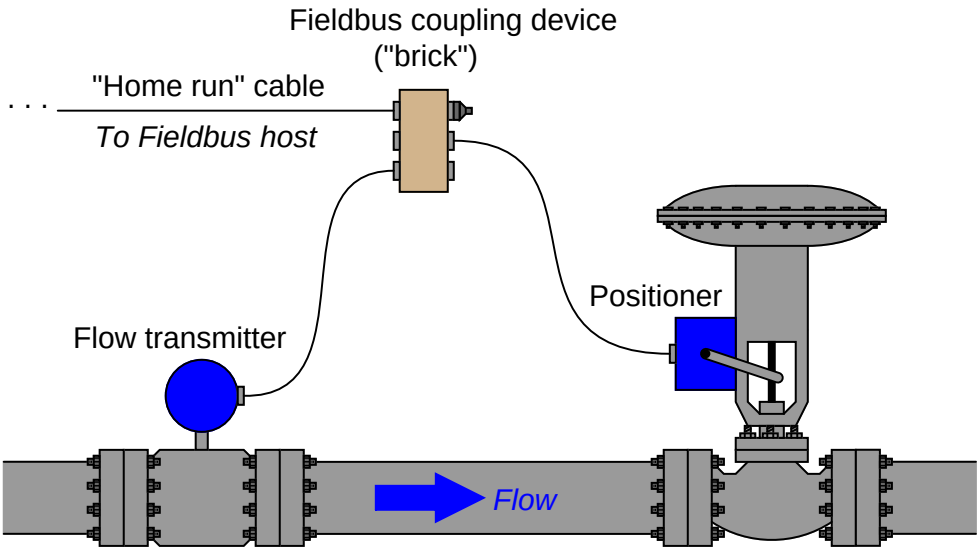
The DCS revolution started in the mid-1970's was fundamentally a moving of control system "intelligence" from a centralized location to distributed locations. Rather than have a single computer (or a panel full of single-loop controllers) located in a central control room implement PID control for a multitude of process loops, many (smaller) computers located closer to the process areas would handle the PID and other control functions, with network cables shuttling data between those distributed locations and the central control room.

Beginning in the late 1980's, the next logical step in this evolution of control architecture saw the relocation of control "intelligence" to the field instruments themselves. In other words, the new idea was to equip individual transmitters and control valve positioners with the necessary computational power to implement PID control all on their own, using digital networks to carry process data between the field instruments and any location desired. This is the fundamental concept of *fieldbus*.

"Fieldbus" as a technical term has multiple definitions. Many manufacturers use the word "fieldbus" to describe any digital network used to transport data to and from field instruments. In this subsection, I use the word "fieldbus" to describe a design philosophy where field instruments possess all the necessary "intelligence" to control the process, with no need for separate centralized (or even distributed) control hardware. *FOUNDATION Fieldbus* is the first standard to embody this fully-distributed control concept, the technical details of this open standard maintained and promoted by the *Fieldbus Foundation*. The aim of this Foundation is to establish an open, technician standard for *any* manufacturer to follow in the design of their fieldbus instruments. This means a FOUNDATION Fieldbus (FF) transmitter manufactured by Smar will work seamlessly with a FF control valve positioner manufactured by Fisher, communicating effortlessly with a FF-aware host system manufactured by ABB, and so on. This may be thought of in terms of being the digital equivalent of the 3-15 PSI pneumatic signal standard or the 4-20 mA analog electronic signal standard: so long as all instruments "talk" according to the same standard, brands and models may be freely interchanged to build any control system desired.

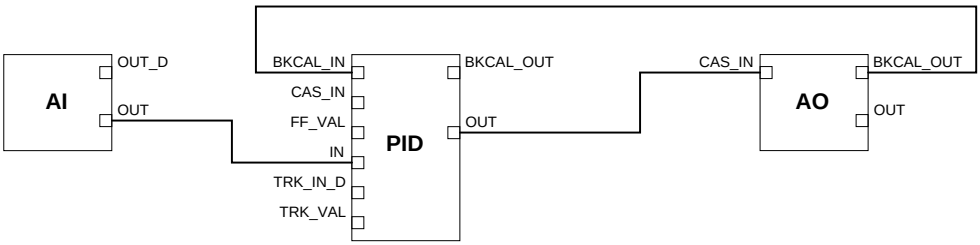
[21] I know of a major industrial manufacturing facility (which shall remain nameless) where a PLC vendor promised the same technical capability as a full DCS at approximately one-tenth the installed cost. Several years and several tens of thousands of man-hours later, the sad realization was this "bargain" did not live up to its promise, and the decision was made to remove the PLCs and go with a complete DCS from another manufacturer. *Caveat emptor!*

To illustrate the general fieldbus concept, consider this flow control system:



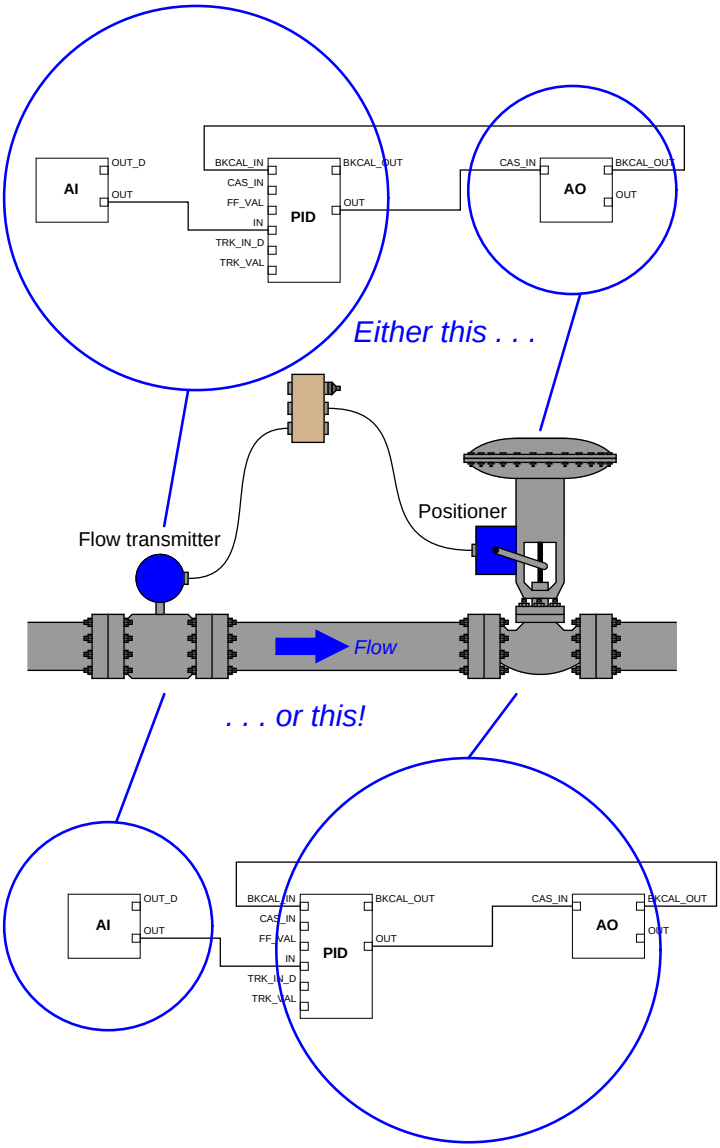
Here, a fieldbus *coupling device* provides a convenient junction point for cables coming from the transmitter, valve positioner, and host system. FOUNDATION Fieldbus devices both receive DC power and communicate digitally over the same twisted-pair cables. In this case, the host system provides DC power for the transmitter and positioner to function, while communication of process data occurs primarily between the transmitter and positioner (with little necessary involvement of the host system[22]).

Just like distributed control systems, FOUNDATION Fieldbus instruments are programmed using a function block language. In this case, we must have an analog input (for the transmitter's measurement), a PID function block, and an analog output (for the valve positioner) to make a complete flow control system:



[22] Although it is customary for the host system to be configured as the *Link Active Scheduler* (LAS) device to schedule and coordinate all fieldbus device communications, this is not absolutely necessary. Any suitable field instrument may also serve as the LAS, which means a host system is not even necessary except to provide DC power to the instruments, and serve as a point of interface for human operators, engineers, and technicians.

The analog input (AI) block must reside in the transmitter, and the analog output (AO) block must reside in the valve positioner, since those blocks necessarily relate to the measured and controlled variables, respectively. However, the PID block may reside in *either* field device:



Practical reasons do exist for choosing one location of the PID function block over the other, most notably the difference in communication loading between the two options[23]. However, there

[23]With the PID function block programmed in the flow transmitter, there will be twice as many scheduled communication events per macrocycle than if the function block is programmed into the valve positioner. This is evident by the number of signal lines connecting circled block(s) to circled block(s) in the above illustration.

is no *conceptual* limitation to the location of the PID function block. In a fieldbus control system where the control "intelligence" is distributed all the way to the field instrument devices themselves, there are no limits to system flexibility.

26.13 Practical PID controller features

In order for any PID controller to be practical, it must be able to do more than just implement the PID equation. This section identifies and explains some of the basic features found on most (but not all!) modern PID controllers:

- Manual versus Automatic mode
- Output tracking
- Setpoint tracking
- Alarming
- PV characterization and damping
- Setpoint limits
- Output limits
- PID tuning security

26.13.1 Manual and automatic modes

When a controller continually calculates output values based on PV and SP values over time, it is said to be operating in *automatic* mode. This mode, of course, is what is necessary to regulate any process. There are times, however, when it is desirable to allow a human operator to manually "override" the automatic action of the PID controller. Applicable instances include process start-up and shut-down events, emergencies, and maintenance procedures. A controller that is being "overridden" by a human being is said to be in *manual mode*.

A very common application of manual mode is during maintenance of the sensing element or transmitter. If an instrument technician needs to disconnect a process transmitter for calibration or replacement, the controller receiving that transmitter's signal cannot be left in automatic mode. If it is, then the controller may[24] take sudden corrective action the moment the transmitter's signal goes dead. If the controller is first placed in manual mode before the technician disconnects the transmitter, however, the controller will ignore any changes in the PV signal, letting its output signal be adjusted at will by the human operator. If there is another indicator of the same process variable as the one formerly reported by the disconnected transmitter, the human operator may elect to read that other indicator and play the part of a PID controller, manually adjusting the final control element to maintain the alternate indicator at setpoint while the technician completes the transmitter's maintenance.

An extension of this "mode" concept applies to controllers configured to receive a setpoint from another device (called a *remote* or *cascaded* setpoint). In addition to an automatic and a manual mode selection, a third selection called *cascade* exists to switch the controller's setpoint from human operator control to remote (or "cascade") control.

[24] The only reason I say "may" instead of "will" is because some modern digital controllers are designed to automatically switch to manual-mode operation in the event of a sensor or transmitter signal loss. Any controller not "smart" enough to shed its operating mode to manual in the event of PV signal loss will react dramatically when that PV signal dies, and this is not a good thing for an operating loop!

26.13.2 Output and setpoint tracking

The provision of manual and automatic operating modes creates a set of potential problems for the PID controller. If, for example, a PID controller is switched from automatic to manual mode by a human operator, and then the output is manually adjusted to some new value, what will the output value do when the controller is switched *back* to automatic mode? In some crude PID controller designs, the result would be an immediate "jump" back to the output value calculated by the PID equation while the controller was in manual. In other words, some controllers never stop evaluating the PID equation – even while in manual mode – and will default to that automatically-calculated output value when the operating mode is switched from manual to automatic.

This can be very frustrating to the human operator, who may wish to use the controller's manual mode as a way to change the controller's bias value. Imagine, for example, that a PD controller (no integral action) is operating in automatic mode at some low output value, which happens to be too low to achieve the desired setpoint. The operator switches the controller to manual mode and then raises the output value, allowing the process variable to approach setpoint. When PV nearly equals SP, the operator switches the controller's mode back to automatic, expecting the PID equation to start working again from this new starting point. In a crude controller, however, the output would jump back to some lower value, right where the PD equation would have placed it for these PV and SP conditions.

A feature designed to overcome this problem – which is so convenient that I consider it an essential feature of any controller with a manual mode – is called *output tracking*. With output tracking, the bias value of the controller shifts every time the controller is placed into manual mode and the output value manually changed. Thus, when the controller is switched from manual mode to automatic mode, the output does *not* immediately jump to some previously-calculated value, but rather "picks up" from the last manually-set value and begins to control from that point as dictated by the PID equation. In other words, output tracking allows a human operator to arbitrarily offset the output of a PID controller by switching to manual mode, adjusting the output value, and then switching back to automatic mode. The output will continue its automatic action from this new starting point instead of the old starting point.

A very important application of output tracking is in the manual correction of integral wind-up (sometimes called *reset windup* or just *windup*). This is what happens to a controller with integral action if for some reason the process variable *cannot* achieve setpoint no matter how far the output signal value is driven by integral action. An example might be on a temperature controller where the source of heat for the process is a steam system. If the steam system shuts down, the temperature controller *cannot* warm the process up to the temperature setpoint value no matter how far open the steam valve is driven by integral action. If the steam system is shut down for too long, the result will be a controller output saturated at maximum value in a futile attempt to warm the process. If and when the steam system starts back up, the controller's saturated output will now send *too much heating steam* to the process, causing the process temperature to overshoot setpoint until integral action drives the output signal back down to some reasonable level. This situation may be averted, however, if the operator switches the temperature controller to manual mode as soon as the steam system shuts down. Even if this preventative step is not taken, the problem of overshoot may be averted upon steam system start-up if the operator uses output tracking by quickly switching the controller into manual mode, adjusting the output down to a reasonable level, and then switching back into automatic mode so that the controller's output value is no longer "wound up" at a high

level[25].

A similar feature to output tracking – also designed for the convenience of a human operator switching a PID controller between automatic and manual modes – is called *setpoint tracking*. The purpose of setpoint tracking is to equalize SP and PV while the controller is in manual mode, so that when the controller gets switched back into automatic mode, it will begin its automatic operation with zero error (PV = SP).

This feature is most useful during system start-ups, where the controller may have difficulty controlling the process in automatic mode under unusual conditions. Operators often prefer to run certain control loops in manual mode from the time of initial start-up until such time that the process is near normal operating conditions. At that point, when the operator is content with the stability of the process, the controller is assigned the responsibility of maintaining the process at setpoint. With setpoint tracking present in the controller, the controller's SP value will be held equal to the PV value (whatever that value happens to be) for the entire time the controller is in manual mode. Once the operator decides it is proper to switch the controller into automatic mode, the SP value freezes at that last manual-mode PV value, and the controller will continue to control the PV at that SP value. Of course, the operator is free to adjust the SP value to any new value while the controller is in automatic mode, but this is at the operator's discretion.

Without setpoint tracking, the operator would *have to* make a setpoint adjustment either before or after switching the controller from manual mode to automatic mode, in order to ensure the controller was properly set up to maintain the process variable at the desired value. With setpoint tracking, the setpoint value will default to the process variable value when the controller was last in manual mode, which (it is assumed) will be close enough to the desired value to suffice for continued operation.

Unlike output tracking, for which there is virtually no reason not to have the feature present in a PID controller, there may very well be applications where we do not wish to have setpoint tracking. For some processes[26], the setpoint value *should* remain fixed at all times, and as such it would be undesirable to have the setpoint value drift around with the process variable value every time the controller was placed into manual mode.

[25] I once had the misfortune of working on an analog PID controller for a chlorine-based wastewater disinfection system that lacked output tracking. The chlorine sensor on this system would occasionally fail due to sample system plugging by algae in the wastewater. When this happened, the PV signal would fail low (indicating abnormally low levels of chlorine gas dissolved in the wastewater) even though the actual dissolved chlorine gas concentration was adequate. The controller, thinking the PV was well below SP, would ramp the chlorine gas control valve further and further open over time, as integral action attempted to reduce the error between PV and SP. The error never went away, of course, because the chlorine sensor was plugged with algae and simply could not detect the actual chlorine gas concentration in the wastewater. By the time I arrived to address the "low chlorine" alarm, the controller output was already wound up to 100%. After cleaning the sensor, and seeing the PV value jump up to some outrageously high level, the controller would take a long time to "wind down" its output because its integral action was very slow. I could not use manual mode to "unwind" the output signal, because this controller lacked the feature of output tracking. My "work-around" solution to this problem was to re-tune the integral term of the controller to some really fast time constant, watch the output "wind down" in fast-motion until it reached a reasonable value, then adjust the integral time constant back to its previous value for continued automatic operation.

[26] Boiler steam drum water level control, for example, is a process where the setpoint really should be left at a 50% value at all times, even if there maybe legitimate reasons for occasionally switching the controller into manual mode.

26.13.3 Alarm capabilities

A common feature on many instrument systems is the ability to alert personnel to the onset of abnormal process conditions. The general term for this function is *alarm*. Process alarms may be triggered by process switches directly sensing abnormal conditions (e.g. high-temperature switches, low-level alarms, low-flow alarms, etc.), in which case they are called *hard alarms*. A *soft alarm*, by contrast, is an alarm triggered by some continuous measurement (i.e. a signal from a process transmitter rather than a process switch) exceeding a pre-programmed alarm limit value.

Since PID controllers are designed to input continuous process measurements, it makes sense that a controller could be equipped with programmable alarm limit values as well, to provide "soft" alarm capability without adding additional instruments to the loop[27]. Not only is PV alarming easy to implement in most PID controllers, but *deviation* alarming is easy to implement as well. A "deviation alarm" is a soft alarm triggered by excessive deviation (error) between PV and SP. Such an event indicates control problems, since a properly-operating feedback loop should be able to maintain reasonable agreement between PV and SP at all times.

Alarm capabilities find their highest level of refinement in modern distributed control systems (DCS), where the networked digital controllers of a DCS provide convenient access and advanced management of hard and soft alarms alike. Not only can alarms be accessed from virtually any location in a facility in a DCS, but they are usually time-stamped and archived for later analysis, which is an *extremely* important feature for the analysis of emergency events, and the continual improvement of process safety.

26.13.4 Output and setpoint limiting

In some process applications, it may not be desirable to allow the controller to automatically manipulate the final control element (control valve, variable-speed motor, heater) over its full 0% - 100% range. In such applications, a useful controller feature is an *output limit*. For example, a PID flow controller may be configured to have a minimum output limit of 5%, so that it is not able to close the control valve any further than the 5% open position in order to maintain "minimum flow" through a pump. The valve may still be fully closed (0% stem position) in manual mode, but just not in automatic mode[28].

Similarly, setpoint values may be internally limited in some PID controllers, such that an operator cannot adjust the setpoint above some limiting value or below some other limiting value. In the event that the process variable *must* be driven outside these limits, the controller may be placed in manual mode and the process "manually" guided to the desired state by an operator.

[27] It is very important to note that soft alarms are not a replacement for hard alarms. There is much wisdom in maintaining both hard and soft alarms for a process, so there will be redundant, non-interactive levels of alarming. Hard and soft alarms should complement each other in any critical process.

[28] Some PID controllers limit manual-mode output values as well, so be sure to check the manufacturer's documentation for output limiting on your particular PID controller!

26.13.5 Security

There is justifiable reason to prevent certain personnel from having access to certain parameters and configurations on PID controllers. Certainly, operations personnel need access to setpoint adjustments and automatic/manual mode controls, but it may be unwise to grant those same operators unlimited access to PID tuning constants and output limits. Similarly, instrument technicians may require access to a PID controller's tuning parameters, but perhaps should be restricted from editing configuration programs maintained by the engineering staff.

Most digital PID controllers have some form of security access control, allowing for different levels of permission in altering PID controller parameters and configurations. Security may be crude (a hidden switch located on a printed circuit board, which only the maintenance personnel should know about), sophisticated (login names and passwords, like a multi-user computer system), or anything in between, depending on the level of development invested in the feature by the controller's manufacturer.

An interesting solution to the problem of security in the days of analog control systems was the architecture of Foxboro's SPEC 200 analog electronic control system. The controller displays, setpoint adjustments, and auto/manual mode controls were located on the control room panel where anyone could access them. All other adjustments (PID settings, alarm settings, limit settings) could be located in the *nest* area where all the analog circuit control cards resided. Since the "nest" racks could be physically located in a room separate from the control room, personnel access to the nest room served as access security to these system parameters.

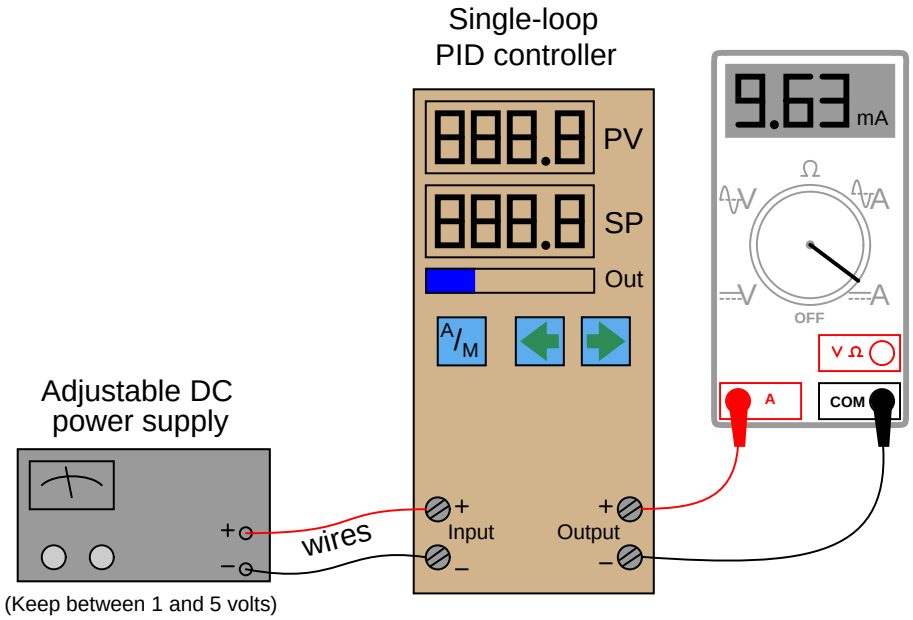
At first, the concept of controller parameter security may seen distrustful and perhaps even insulting to those denied access, especially when the denied persons possess the necessary knowledge to understand the functions and consequences of those parameters. It is not uncommon for soft alarm values to be "locked out" from operator access despite the fact that operators understand very well the purpose and functions of these alarms. At some facilities, PID tuning is the exclusive domain of process engineers, with instrument technicians and operators alike barred from altering PID tuning constants even though some operators and many technicians may well understand PID controller tuning.

When considering security access, there is more to regard than just knowledge or ability. At a fundamental level, security is a task of limiting access commensurate with *responsibility*. In other words, security restrictions exist to exclude those not charged with particular responsibilities. Knowledge and ability are necessary conditions of responsibility (i.e. one cannot reasonably be held responsible for something beyond their knowledge or control), but they are not *sufficient* conditions of responsibility (i.e. knowing how to, and being able to perform a task does not confer responsibility for that task getting completed). An operator may very well understand how and why a soft alarm on a controller works, but the responsibility for altering the alarm value may reside with someone else whose job description it is to ensure the alarm values correspond to plant-wide policies.

26.14 Note to students

PID control can be a frustrating subject for many students, even those with previous knowledge of calculus. At times it can seem like an impossibly abstract concept to master.

Thankfully, there is a relatively simple way to make PID control more "real," and that is hands-on experience with a real PID controller. I advise you acquire an electronic single-loop PID controller[29] and set it up with an adjustable DC power supply and milliammeter as such:



Most electronic controllers input a 1 to 5 VDC signal for the process variable (often with a 250 ohm resistor connected across the input terminals to generate a 1-5 VDC drop from a 4-20 mA current signal, which you will not need here). By adjusting the DC power supply between 1 and 5 volts DC, you will simulate a transmitter signal to the controller's input between 0% and 100%.

The milliammeter reads current output by the controller, 4 mA representing a 0% output signal and 20 mA representing a 100% output signal. With the power supply and milliammeter both connected to the appropriate terminals on the controller, you are all set to simulate input conditions and watch the controller's output response.

This arrangement does not simulate a process, and so there will be no feedback for you to observe. The purpose of this setup is to simply learn how the controller is supposed to respond to different PV and SP conditions, so that you may gain an intuitive "feel" for the PID algorithm to supplement your theoretical understanding of it. Experimentation with a real process (or even a simulated process) comes later (see section 27.6, beginning on page 1574).

[29] Many instrument manufacturers sell simple, single-loop controllers for reasonable prices, comparable to the price of a college textbook. You need to get one that accepts 1-5 VDC input signals and generates 4-20 mA output signals, and has a "manual" mode of operation in addition to automatic – these features are *very important!* Avoid controllers that can only accept thermocouple inputs, and/or only have time-proportioning (PWM) outputs.

Once you have all components connected, you should check to see that everything works:

- Set power supply to 1, 3, and then 5 volts DC. The controller's PV display should read 0%, 50%, and 100%, respectively. The PV display should follow closely to the power supply voltage signal over time. If the display seems to "lag" behind the power supply adjustment, then it means the controller has damping configured for the input signal. You should keep the damping set to the minimum possible value, so the controller is as responsive as it can be.
- Put the controller in manual mode and set the output to 0%, 50%, and then 100%. The milliammeter should register 4 mA, 12 mA, and 20 mA, respectively.

After checking these basic functions, you may proceed to do the following experiments. For each experiment, I recommend setting the PV input signal to 3 volts DC (50%), and manually setting the output to 50% (12 mA on the milliammeter). When you are ready to test the P,I,D responses of the controller, place the controller into automatic mode and then observe the results.

26.14.1 Proportional-only control action

1. Set the controller PV input to 50% (3 volts) and the output value to 50% in manual mode.
2. Configure the controller for reverse action (this is typically the default setting).
3. Configure the PID settings for proportional action only. This may be done by setting the gain equal to 1 (P.B. = 100%), the integral setting to zero repeats per minute (maximum minutes per repeat), and the derivative setting to zero minutes. Some controllers have the ability to switch to a "proportional-only" algorithm – if your controller has that ability, this is the best way to get set up for this exercise.
4. Switch the controller mode to "automatic."
5. Adjust the PV signal to 75% (4 volts) and observe the output. How far does the output signal move from its starting value of 50%? How does the magnitude of this step relate to the magnitude of the PV step? Does the output signal drift or does it remain the same when you stop changing the PV signal?
6. Adjust the PV signal to 25% (2 volts) and observe the output. How far does the output signal move from its starting value of 50%? How does the magnitude of this step relate to the magnitude of the PV step? Does the output signal drift or does it remain the same when you stop changing the PV signal?
7. Change the controller's gain setting to some different value and repeat the previous two steps. How does the output step magnitude relate to the input step-changes in each case? Do you see the relationship between controller gain and how the output responds to changes in the input?
8. Smoothly vary the input signal back and forth between 0% and 100% (1 and 5 volts). How does the output respond when you do this? Try changing the gain setting again and re-checking.
9. Switch the controller's action from *reverse* to *direct*, then repeat the previous step. How does the output respond now?

26.14.2 Integral-only control action

1. Set the controller PV input to 50% (3 volts) and the output value to 50% in manual mode.
2. Configure the controller for reverse action (this is typically the default setting).
3. Configure the PID settings for integral action only. If the controller has an "I-only" mode, this is the best way to get set up for this exercise. If there is no way to completely turn off proportional action, then I recommend setting the gain value to the minimum non-zero value allowed, and setting the integral constant to an aggressive value (many repeats per minute, or fractions of a minute per repeat). If your controller does have an integral-only option, I recommend setting the integral time constant at 1 minute. Set derivative action at zero minutes.
4. Switch the controller mode to "automatic."
5. Adjust the PV signal to 75% (4 volts) and observe the output. Which way does the output signal move? Does the output signal drift or does it remain the same when you stop changing the PV signal? How does this action compare with the proportional-only test?
6. Adjust the PV signal to 25% (2 volts) and observe the output. Which way does the output signal move? Does the output signal drift or does it remain the same when you stop changing the PV signal? How does this action compare with the proportional-only test?
7. Change the controller's integral setting to some different value and repeat the previous two steps. How does the rate of output ramping relate to the input step-changes in each case? Do you see the relationship between the integral time constant and how the output responds to changes in the input?
8. Smoothly vary the input signal back and forth between 0% and 100% (1 and 5 volts). How does the output respond when you do this? Try changing the integral setting again and re-checking.
9. Where must you adjust the input signal to get the output to stop moving? When the output finally does settle, is its value consistent (i.e. does it always settle at the same value, or can it settle at different values)?
10. Switch the controller's action from *reverse* to *direct*, then repeat the previous two steps. How does the output respond now?

26.14.3 Proportional plus integral control action

1. Set the controller PV input to 50% (3 volts) and the output value to 50% in manual mode.
2. Configure the controller for reverse action (this is typically the default setting).
3. Configure the PID settings with a proportional (gain) value of 1 (P.B. = 100%) and an integral value of 1 repeat per minute (or 1 minute per repeat). Set derivative action at zero minutes.
4. Switch the controller mode to "automatic."
5. Adjust the PV signal to 75% (4 volts) and observe the output. Which way does the output signal move? Does the output signal drift or does it remain the same when you stop changing the PV signal? How does this action compare with the proportional-only test and with the integral-only test?
6. Adjust the PV signal to 25% (2 volts) and observe the output. Which way does the output signal move? Does the output signal drift or does it remain the same when you stop changing the PV signal? How does this action compare with the proportional-only test and with the integral-only test?
7. Change the controller's gain setting to some different value and repeat the previous two steps. Can you tell which aspect of the output signal's response is due to proportional action and which aspect is due to integral action?
8. Change the controller's integral setting to some different value and repeat those same two steps. Can you tell which aspect of the output signal's response is due to proportional action and which aspect is due to integral action?
9. Smoothly vary the input signal back and forth between 0% and 100% (1 and 5 volts). How does the output respond when you do this? Try changing the gain and/or integral settings again and re-checking.
10. Switch the controller's action from *reverse* to *direct*, then repeat the previous two steps. How does the output respond now?

26.14.4 Proportional plus derivative control action

1. Set the controller PV input to 50% (3 volts) and the output value to 50% in manual mode.
2. Configure the controller for reverse action (this is typically the default setting).
3. Configure the PID settings with a proportional (gain) value of 1 (P.B. = 100%) and a derivative value of 1 minute. Set integral action at zero repeats per minute (maximum number of minutes per repeat).
4. Switch the controller mode to "automatic."
5. Adjust the PV signal to 75% (4 volts) and observe the output. Which way does the output signal move? How does the output signal value compare while you are adjusting the input voltage versus after you reach 4 volts and take your hand off the adjustment knob? How does this action compare with the proportional-only test?
6. Adjust the PV signal to 25% (2 volts) and observe the output. Which way does the output signal move? How does the output signal value compare while you are adjusting the input voltage versus after you reach 4 volts and take your hand off the adjustment knob? How does this action compare with the proportional-only test?
7. Change the controller's gain setting to some different value and repeat the previous two steps. Can you tell which aspect of the output signal's response is due to proportional action and which aspect is due to derivative action?
8. Change the controller's derivative setting to some different value and repeat those same two steps. Can you tell which aspect of the output signal's response is due to proportional action and which aspect is due to derivative action?
9. Smoothly vary the input signal back and forth between 0% and 100% (1 and 5 volts). How does the output respond when you do this? Try changing the derivative setting again and re-checking.
10. Switch the controller's action from *reverse* to *direct*, then repeat the previous two steps. How does the output respond now?

26.14.5 Full PID control action

1. Set the controller PV input to 50% (3 volts) and the output value to 50% in manual mode.
2. Configure the controller for reverse action (this is typically the default setting).
3. Configure the PID settings with a proportional (gain) value of 1 (P.B. = 100%), an integral value of 1 repeat per minute (or 1 minute per repeat), and a derivative action of 1 minute.
4. Switch the controller mode to "automatic."
5. Adjust the PV signal to 75% (4 volts) and observe the output. Which way does the output signal move? Does the output signal drift or does it remain the same when you stop changing the PV signal? How does magnitude of the output signal compare while you are changing the input voltage, versus when the input signal is steady?
6. Adjust the PV signal to 25% (2 volts) and observe the output. Which way does the output signal move? Does the output signal drift or does it remain the same when you stop changing the PV signal? How does magnitude of the output signal compare while you are changing the input voltage, versus when the input signal is steady?
7. Change the controller's gain setting to some different value and repeat the previous two steps. Can you tell which aspect of the output signal's response is due to proportional action, which aspect is due to integral action, and which aspect is due to derivative action?
8. Change the controller's integral setting to some different value and repeat the same two steps. Can you tell which aspect of the output signal's response is due to proportional action, which aspect is due to integral action, and which aspect is due to derivative action?
9. Change the controller's derivative setting to some different value and repeat the same two steps. Can you tell which aspect of the output signal's response is due to proportional action, which aspect is due to integral action, and which aspect is due to derivative action?
10. Smoothly vary the input signal back and forth between 0% and 100% (1 and 5 volts). How does the output respond when you do this? Try changing the gain, integral, and/or derivative settings again and re-checking.
11. Switch the controller's action from *reverse* to *direct*, then repeat the previous two steps. How does the output respond now?

References

"FOUNDATION Fieldbus", document L454 EN, Samson AG, Frankfurt, Germany, 2000.

"Identification and Description of Instrumentation, Control, Safety, and Information Systems and Components Implemented in Nuclear Power Plants", EPRI, Palo Alto, CA: 2001. 1001503.

Lavigne, John R., *Instrumentation Applications for the Pulp and Paper Industry*, The Foxboro Company, Foxboro, MA, 1979.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Control Volume II*, Third Edition, CRC Press, Boca Raton, FL, 1999.

Mollenkamp, Robert A., *Introduction to Automatic Process Control*, Instrument Society of America, Research Triangle Park, NC, 1984.

"Moore 353 Process Automation Controller User's Manual", document UM353-1, Revision 11, Siemens Energy and Automation, 2003.

Shinskey, Francis G., *Energy Conservation through Control*, Academic Press, New York, NY, 1978.

Shinskey, Francis G., *Process-Control Systems – Application / Design / Adjustment*, Second Edition, McGraw-Hill Book Company, New York, NY, 1979.

"SPEC 200 Systems", technical information document TI 200-100, Foxboro, 1980.

"SPEC 200 System Configuration", technical information document TI 200-105, Foxboro, January 1975.

"SPEC 200 System Wiring", technical information document TI 200-260, Foxboro, 1972.

Ziegler, J. G., and Nichols, N. B., *Optimum Settings for Automatic Controllers*, Transactions of the American Society of Mechanical Engineers (ASME), Volume 64, pages 759-768, Rochester, NY, November 1942.

Chapter 27

Process dynamics and PID controller tuning

To *tune* a feedback control system means to adjust parameters in the controller to achieve robust control over the process. "Robust" in this context is usually defined as stability of the process variable despite changes in load, fast response to changes in setpoint, minimal oscillation following either type of change, and minimal offset (error between setpoint and process variable) over time.

"Robust control" is far easier to define than it is to achieve. With PID (Proportional-Integral-Derivative) control being the most common feedback control algorithm used in industry, it is important for the instrument technician (and engineer!) to understand how to tune these controllers effectively and with a minimum investment of time.

Different types of processes, having different dynamic (time-dependent) behaviors, require different levels of proportional, integral, and derivative control action to achieve stability and robust response. It is therefore imperative for anyone seeking to tune a PID controller to understand the dynamic nature of the process being controlled. For this reason, the chapter begins with an exploration of common process characteristics before introducing techniques useful in choosing practical P, I, and D tuning parameter values.

27.1 Process characterization

Perhaps the most important rule of controller tuning is to *know the process before attempting to adjust the controller's tuning*. Unless you adequately understand the nature of the process you intend to control, you will have little hope in actually controlling it well. This section of the book is dedicated to an investigation of different process characteristics and how to identify each.

Quantitative PID tuning methods (see section 27.3 beginning on page 1549) attempt to map the characteristics of a process so good PID parameters may be chosen for the controller. The goal of this section is for you to understand various process types by observation and qualitative analysis so you may comprehend why different tuning parameters are necessary for each type, rather than mindlessly following a step-by-step PID tuning procedure.

The three major classifications of process response are *self-regulating*, *integrating*, and *runaway*. Each of these process types is defined by its response to a step-change in the manipulated variable (e.g. control valve position or state of some other final control element). A "self-regulating" process responds to a step-change in the final control element's status by settling to a new, stable value. An "integrating" process responds by ramping either up or down at a rate proportional to the magnitude of the final control element's step-change. Finally, a "runaway" process responds by ramping either up or down at a rate that increases over time, headed toward complete instability without some form of corrective action from the controller.

Self-regulating, integrating, and runaway processes have very different control needs. PID tuning parameters that may work well to control a self-regulating process, for example, will *not* work well to control an integrating or runaway process, no matter how similar any of the other characteristics of the processes may be[1]. By first identifying the characteristics of a process, we may draw some general conclusions about the P, I, and D setting values necessary to control it well.

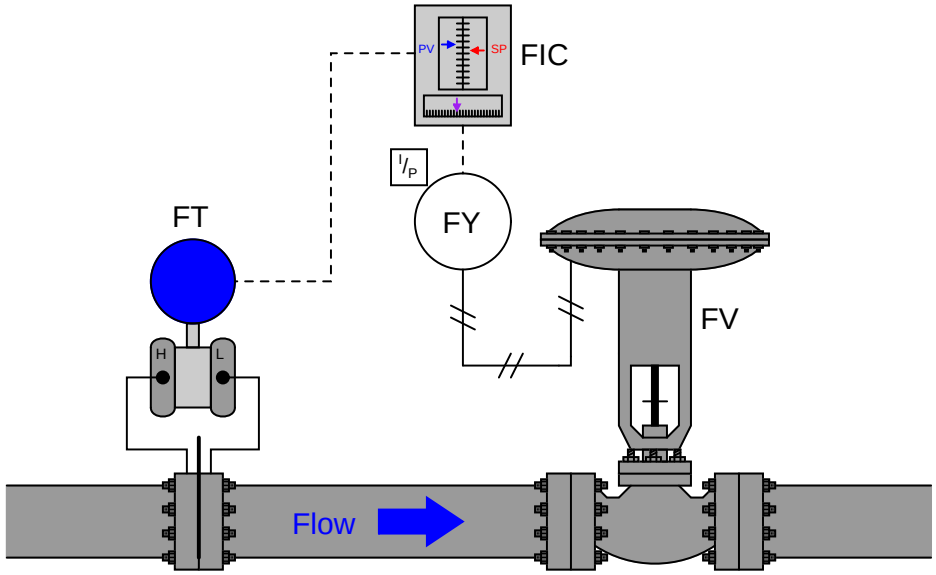
Perhaps the best method for testing a process to determine its natural characteristics is to place the controller in *manual mode* and introduce a step-change to the controller output signal. It is critically important that the loop controller be in manual mode whenever process characteristics are being explored. If the controller is left in the automatic mode, the response seen from the process to a setpoint or load change will be partly due to the natural characteristics of the process itself *and* partly due to the corrective action of the controller. The controller's corrective action thus interferes with our goal of exploring process characteristics. By placing the controller in "manual" mode, we turn off its corrective action, effectively removing its influence by breaking the feedback loop between process and controller, controller and process. In manual mode, the response we see from the process to an output (manipulated variable) or load change is *purely* a function of the natural process dynamics, which is precisely what we wish to discern.

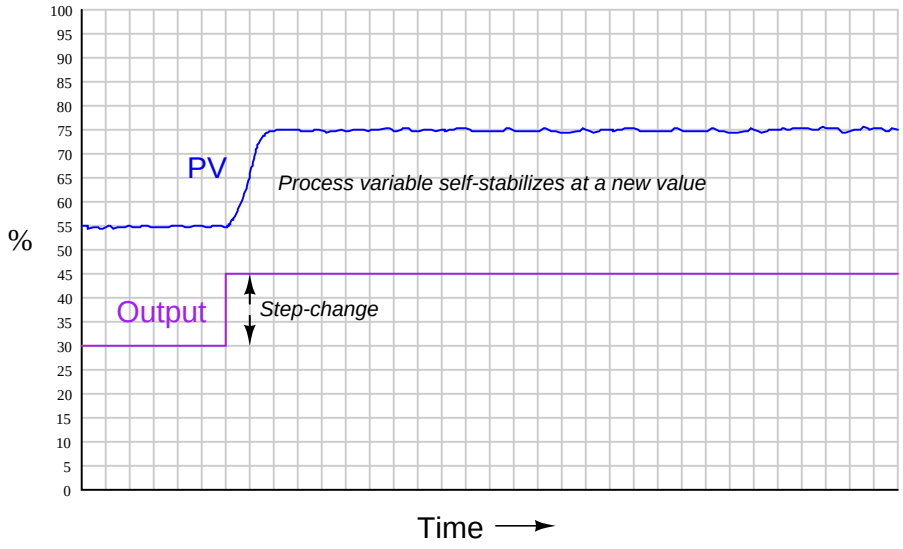
A test of process characteristics with the loop controller in manual mode is often referred to as an *open-loop* test, because the feedback loop has been "opened" and is no longer a complete loop. Open-loop tests are the fundamental diagnostic technique applied in the following subsections.

[1] To illustrate, self-regulating processes require significant integral action from a controller in order to avoid large offsets between PV and SP, with minimal proportional action and no derivative action. Integrating processes, in contrast, may be successfully controlled primarily on proportional action, with minimal integral action to eliminate offset. Runaway processes absolutely require derivative action for dynamic stability, but derivative action alone is not enough: some integral action will be necessary to eliminate offset. Even if knowledge of a process's dominant characteristic does not give enough information for us to quantify P, I, or D values, it will tell us which tuning constant will be most important for achieving stability.

27.1.1 Self-regulating processes

A good example of a self-regulating process is liquid flow control. If a control valve is opened in a step-change fashion, liquid flow through the pipe generally stabilizes at a new rate rather quickly. The following illustration shows a typical liquid flow-control installation, with a process trend showing the flow response to a step-change in valve position (with the controller in manual mode):





The defining characteristic of a self-regulating process is its inherent ability to settle at a new

process variable value without any corrective action on the part of the controller. In other words, there is a unique process variable value for each possible output (valve) value.

A corollary to the above statement is that a unique output value *will be required* to achieve a new process variable value. For example, to achieve a greater flow rate, the control valve must be opened further and held at that further-open position for as long as the greater flow rate is desired. This presents a fundamental problem for a proportional-only controller. Recall the formula for a proportional-only controller, defining the output value (m) by the error (e) between process variable and setpoint multiplied by the gain (K_p) and added to the bias (b):

$$m = K_p e + b$$

Where,
m = Controller output
e = Error (difference between PV and SP)
K_p = Proportional gain
b = Bias

Suppose we find the controller in a situation where there is zero error (PV = SP), and the flow rate is holding steady at some value. If we then increase the setpoint value (calling for a greater flow rate), the error will increase, driving the valve further open. As the control valve opens further, flow rate naturally increases to match. This increase in process variable drives the error back toward zero, which in turn causes the controller to decrease its output value back toward where it was before the setpoint change. However, the error can never go all the way back to zero because if it did, the valve would return to its former position, and that would cause the flow rate to self-regulate back to its original value before the setpoint change was made. What happens instead is that the control valve begins to close as flow rate increases, and eventually the process finds some equilibrium point where the flow rate is steady at some value less than the setpoint, creating just enough error to drive the valve open just enough to maintain that new flow rate. Unfortunately, due to the need for an error to exist, this new flow rate will fall shy of our setpoint. We call this error *proportional-only offset*, or *droop*, and it is an inevitable consequence of a proportional-only controller attempting to control a self-regulating process.

For any fixed bias value, there will be only one setpoint value that is perfectly achievable for a proportional-only controller in a self-regulating process. Any other setpoint value will result in some degree of offset in a self-regulating process. If dynamic stability is more important than absolute accuracy (zero offset) in a self-regulating process, a proportional-only controller may suffice. A great many self-regulating processes in industry have been and still are controlled by proportional-only controllers, despite some inevitable degree of offset between PV and SP.

The amount of offset experienced by a proportional-only controller in a self-regulating process may be minimized by increasing the controller's gain. If it were possible to increase the gain of a proportional-only controller to infinity, it would be able to achieve any setpoint desired with zero offset! However, there is a practical limit to the extent we may increase the gain value, and that limit is *oscillation*. If a controller is configured with too much gain, the process variable will begin to oscillate over time, never stabilizing at any value at all, which of course is highly undesirable for any automatic control system. Even if the gain is not great enough to cause sustained oscillations, excessive values of gain will still cause problems by causing the process variable to oscillate with decreasing amplitude for a period of time following a sudden change in either setpoint or load.

Determining the optimum gain value for a proportional-only controller in a self-regulating process is, therefore, a matter of compromise between excessive offset and excessive oscillation.

Recall that the purpose of integral (or "reset") control action was the elimination of offset. Integral action works by ramping the output of the controller at a rate determined by the magnitude of the offset: the greater the difference between PV and SP for an integral controller, the faster that controller's output will ramp over time. In fact, the output will stabilize at some value *only* if the error is diminished to zero (PV = SP). In this way, integral action works tirelessly to eliminate offset.

It stands to reason then that a self-regulating process *absolutely requires* some amount of integral action in the controller in order to achieve zero offset for every setpoint or load condition. The more aggressive (faster) a controller's integral action, the sooner offset will be eliminated. Just how much integral action a self-regulating process can tolerate depends on the magnitudes of any time lags in the system. The faster a process's natural response is to a manual step-change in controller output, the better it will respond to aggressive integral controller action once the controller is placed in automatic mode. Aggressive integral control action in a slow process, however, will result in oscillation due to integral wind-up[2].

It is not uncommon to find self-regulating processes being controlled by *integral-only* controllers. An "integral-only" process controller is an instrument lacking proportional or derivative control modes. Liquid flow control is a nearly ideal candidate process for integral-only control action, due to its self-regulating and fast-responding nature.

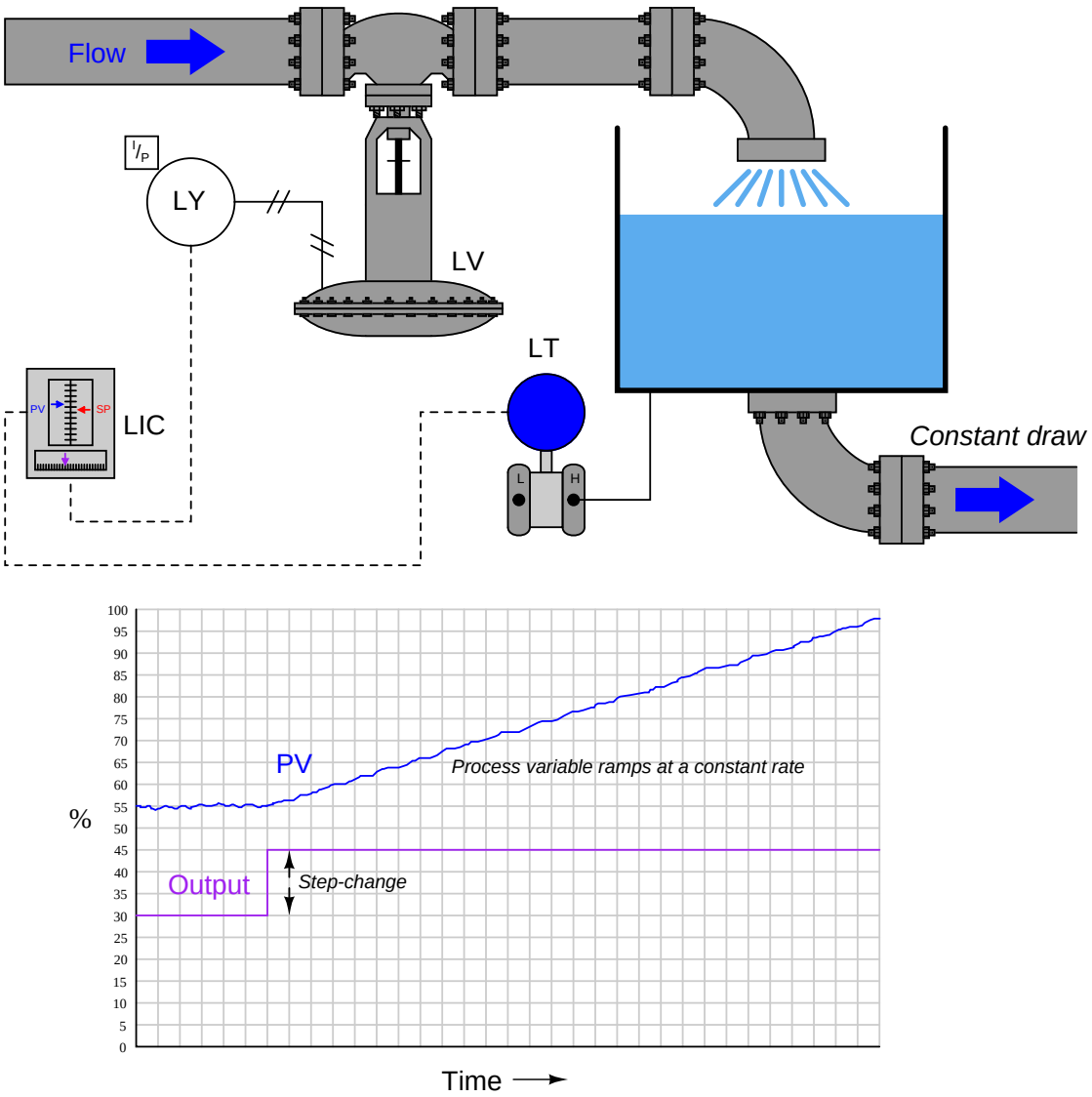
Summary:

- Self-regulating processes are characterized by their natural ability to stabilize at a new process variable value following changes in the control element value or load(s).
- Self-regulating processes *absolutely require* integral controller action to eliminate offset between process variable and setpoint.
- Faster integral controller action results in quicker elimination of offset.
- The amount of integral controller action tolerable in a self-regulating process depends on the degree of time lag in the system. Too much integral action will result in oscillation, just like too much proportional control action.

[2] Recall that wind-up is what happens when integral action "demands" more from a process than the process can deliver. If integral action is too aggressive for a process (i.e. fast integral controller action in a process with slow time lags), the output will ramp too quickly, causing the process variable to overshoot setpoint which then causes integral action to wind the other direction. Just like proportional action, too much integral action will cause a self-regulating process to oscillate.

27.1.2 Integrating processes

A good example of an integrating process is liquid level control, where either the flow rate of liquid into or out of a vessel is constant and the other flow rate is controlled. If a control valve is opened in a step-change fashion, liquid level in the vessel ramps at a rate proportional to the difference in flow rates in and out of the vessel. The following illustration shows a typical liquid level-control installation, with a process trend showing the level response to a step-change in valve position (with the controller in manual mode):



It is critically important to realize that this ramping action of the process variable over time is a

characteristic of the process itself, not the controller. When liquid flow rates in and out of a vessel are mis-matched, the liquid level within that vessel will change at a rate proportional to the difference in flow rates. The trend shown here reveals a fundamental characteristic of the process, not the controller (this should be obvious once it is realized that the step-change in output is something that would only ever happen with the controller in *manual* mode).

Mathematically, we may express the integrating nature of this process using calculus notation. First, we may express the *rate of change* of volume in the tank over time ($\frac{dV}{dt}$) in terms of the flow rates in and out of the vessel:

$$\frac{dV}{dt} = Q_{in} - Q_{out}$$

For example, if the flow rate of liquid going into the vessel was 450 gallons per minute, and the constant flow rate drawn out of the vessel was 380 gallons per minute, the volume of liquid contained within the vessel would increase over time at a rate equal to 70 gallons per minute: the difference between the in-flow and the out-flow rates.

Another way to express this mathematical relationship between flow rates and liquid volume in the vessel is to use the calculus function of *integration*:

$$V = \int_0^T (Q_{in} - Q_{out})\, dt$$

The amount of liquid volume accumulated in the vessel (V) between time 0 and time T is equal to the sum ($\int$) of the products (multiplication) of difference in flow rates in and out of the vessel ($Q_{in} - Q_{out}$) and infinitesimal intervals of time (dt).

In the given scenario of a liquid level control system where the out-going flow is held constant, this means the level will be stable only at one in-coming flow rate (where $Q_{in} = Q_{out}$). At any other controlled flow rate, the level will either be increasing over time or decreasing over time.

This process characteristic perfectly matches the characteristic of a proportional-only controller, where there is one unique output value when the error is zero (PV = SP). Imagine this process controlled by a proportional-only controller in automatic mode, where the bias value (b) of the controller was set to the exact value needed by the control valve to make in-coming flow exactly equal to out-going flow. This means that when the process variable is exactly equal to setpoint (PV = SP), the liquid level will hold constant. If now an operator were to increase the setpoint value (with the controller in automatic mode), it would cause the valve to open further, adding liquid at a faster rate to the vessel. The naturally integrating nature of the process will result in an increasing liquid level. As level increases, the amount of error in the controller decreases, causing the valve to return to its original (bias) position. When the level exactly reaches the new setpoint, the controller output will have returned to its original (bias) value, which will make the control valve go to its original position and hold the level constant once again. Thus, a proportional-only controller will achieve the new setpoint with absolutely no offset ("droop").

The more aggressive the controller's proportional action, the sooner the integrating process will reach new setpoints. Just how much proportional action (gain) an integrating process can tolerate depends on the magnitudes of any time lags in the system as well as the magnitude of noise in the process variable signal. Any process system with time lags will oscillate if the controller has sufficient gain. Noise is a problem because proportional action directly reproduces process variable noise on the output signal: too much gain, and just a little bit of PV noise translates into a control valve whose stem position constantly jumps around.

Unlike a self-regulating process, a purely integrating process has no need for integral action inside the controller to eliminate offset. It is as though the integrating action inherent to the process naturally eliminates offset without any external assistance. More than that, the presence of any integral action in the controller will actually force the process variable to overshoot setpoint in a purely integrating process! Imagine a controller with integral action responding to a step-change in setpoint for this process. As soon as an error develops, the integral action will begin "winding up" the output value, forcing the valve to open more than proportional action alone would demand. By the time the liquid level reaches the new setpoint, the valve will have reached a position greater than where it originally was before the setpoint change[3], which means the liquid level will *not* stop rising when it reaches setpoint, but in fact will overshoot setpoint. Only after the liquid level has spent sufficient time above setpoint will the integral action of the controller "wind" back down to its previous level, allowing the liquid level to finally achieve the new setpoint.

Of course, this is an idealized scenario. There will be factors requiring the use of some integral action when controlling an integrating process. Consider, for example, if the out-going flow rate were to change. Now, a new valve position will be required to achieve stable (unchanging) level in the vessel. A proportional-only controller is able to generate a new valve position *only* if an error develops between PV and SP. Without at least some degree of integral action configured in the controller, that error will persist indefinitely. Or consider if the liquid supply pressure upstream of the control valve were to change, resulting in a different rate of incoming flow for the same valve stem position as before. Once again, the controller would have to generate a different output value to compensate for this process change and stabilize liquid level, and the only way a proportional-only controller could do that is to let the process variable drift a bit from setpoint (the definition of an error or offset).

The example of an integrating process used here is but one of many possible processes where we are dealing with either a *mass balance* or an *energy balance* problem. "Mass balance" is the accounting of all mass into and out of a process. Since the Law of Mass Conservation states the impossibility of mass creation or destruction, all mass into and out of a process must be accounted for. If the mass flow rate into a process does not equal the mass flow rate out of a process, the process must be either gaining or losing an internal store of mass. The same may be said for energy: all energy flowing into and out of a process must be accounted for, since the Law of Energy Conservation states the impossibility of energy creation or destruction. If the energy flow rate (input power) into a process does not equal the energy flow rate (output power) out of a process, the process must be either gaining or losing an internal store of energy.

[3] In a proportional-only controller, the output is a function of error (PV − SP) and bias. When PV = SP, bias alone determines the output value (valve position). However, in a controller with integral action, the zero-offset output value is determined by *how long* and *how far* the PV has previously strayed from SP. In other words, there is no fixed bias value anymore. Thus, the output of a controller with integral action will *not* return to its previous value once the new SP is reached. In a purely integrating process, this means the PV will *not* reach stability at the new setpoint, but will continue to rise until all the "winding up" of integral action is un-done.

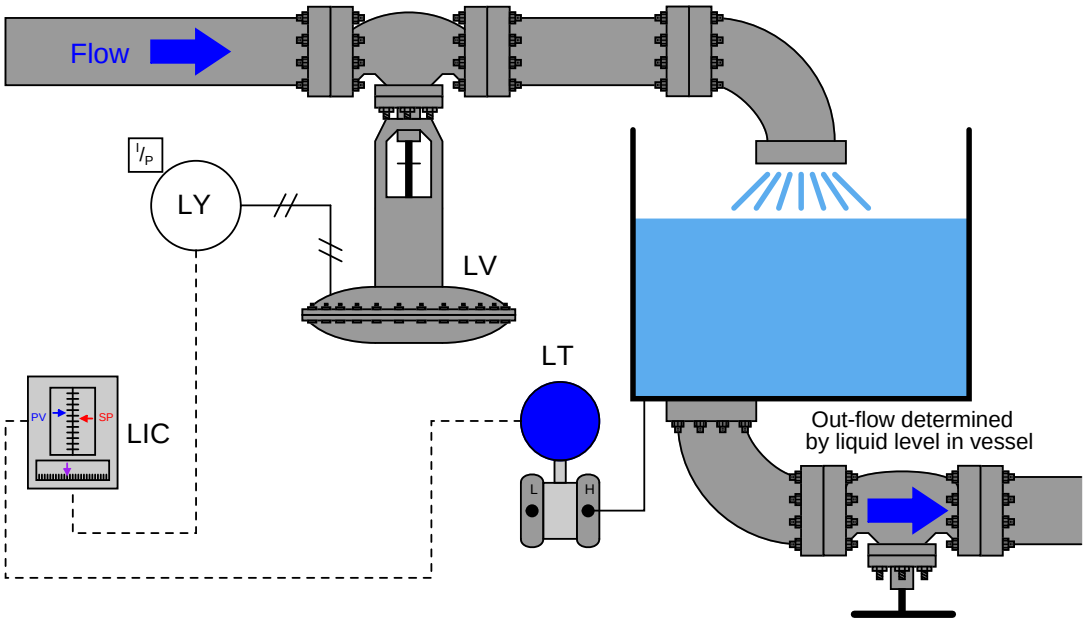
Other examples of integrating processes include the following:

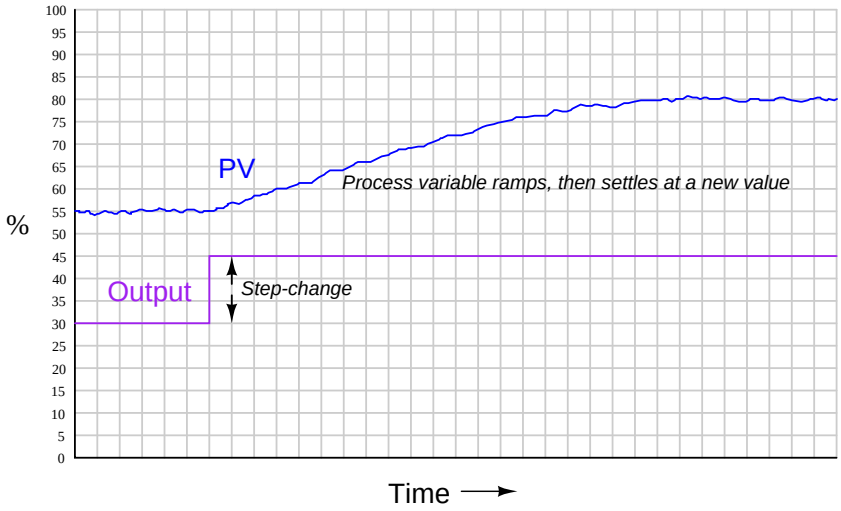
- Gas pressure control – *mass balance* – when the flow of gas either into or out of a vessel is manipulated, and the other flows in or out of the vessel are constant
- Storage bin level control – *mass balance* – when the conveyor feed rate into the bin is manipulated, and the draw from the bin is constant
- Temperature control – *energy balance* – when the flow of heat into or out of a process is manipulated, and all other heat flows are constant
- Speed control – *energy balance* – when the force (linear) or torque (angular) applied to a mass is manipulated, and all other loads are constant in force or torque

In a self-regulating process, the control element (valve) exerts control over *both* the in-flow and the out-flow of either mass or energy. In the previous subsection, where liquid flow control was the process example, the mass balance consisted of liquid flow into the valve and liquid flow out of the valve. Since the piping was essentially a "series" path for an incompressible fluid, where input flow must equal output flow at any given time, mass in and mass out were *guaranteed* to be in a state of balance, with one valve controlling both. This is why a change in valve position resulted in an almost immediate change and re-stabilization of flow rate: the valve exerts immediate control over both the incoming and the outgoing flow rates, with both in perfect balance. Therefore, nothing "integrates" over time in a liquid flow control process because there can never be an imbalance between in-flow and out-flow.

In an integrating process, the control element (valve) exerts control over *either* the in-flow *or* the out-flow of mass or energy, but never both. Thus, changing valve position in an integrating process causes an imbalance of mass flow and/or energy flow, resulting in the process variable ramping over time as either mass or energy accumulates in (or depletes from) the process.

Our "simple" example of an integrating (level-control) process becomes a bit more complicated if the outgoing flow is dependent on level, as is the case with a gravity-drained vessel:





If we subject the control valve to a manual step-change increase, the flow rate of liquid into the vessel immediately increases. This causes an imbalance of incoming and outgoing flow, resulting in the liquid level rising over time. As level rises, however, increasing hydrostatic pressure across the manual valve at the vessel outlet causes the outgoing flow rate to increase. This causes the mass imbalance rate to be less than it was before, resulting in a decreased integration rate (rate of level

rise). Thus, the liquid level still rises, but at a slower and slower rate as time goes on. Eventually, the liquid level will become high enough that the pressure across the manual valve forces a flow rate out of the vessel equal to the flow rate into the vessel. At this point, with matched flow rates, the liquid level stabilizes with no corrective action from the controller (remember, the step-change in output was made in manual mode!). Note the final result of letting the outgoing flow be a function of liquid level: *what used to be an integrating process has now become a self-regulating process,* albeit one with a substantial lag time.

Many processes ideally categorized as integrating actually behave in this manner. Although the manipulated variable may control the flow rate into or out of a process, the other flow rates often change with the process variable. Returning to our list of integrating process examples, we see how a PV-variable load in each case can make the process self-regulate:

- Gas pressure control – *mass balance* – if uncontrolled flow rates into the vessel are allowed to naturally decrease with increasing vessel pressure, and uncontrolled flow rates out of the vessel are allowed to naturally increase with increasing vessel pressure, the vessel's pressure will tend to self-regulate instead of integrate
- Storage bin level control – *mass balance* – if the draw from the bin increases with bin level (greater weight pushing material out at a faster rate), the bin's level will tend to self-regulate instead of integrate
- Temperature control – *energy balance* – if the process loses heat at a faster rate as temperature increases, the process temperature will tend to self-regulate instead of integrate
- Speed control – *energy balance* – if drag forces on the object increase with speed (as they usually do for any fast-moving object), the speed will tend to self-regulate instead of integrate

This one detail completely alters the fundamental characteristic of a process from integrating to self-regulating, and therefore changes the necessary controller parameters. Now, at least some integral controller action is *required* to attain new setpoint values, where none was required before (and in fact where any integral action at all would result in overshoot!).

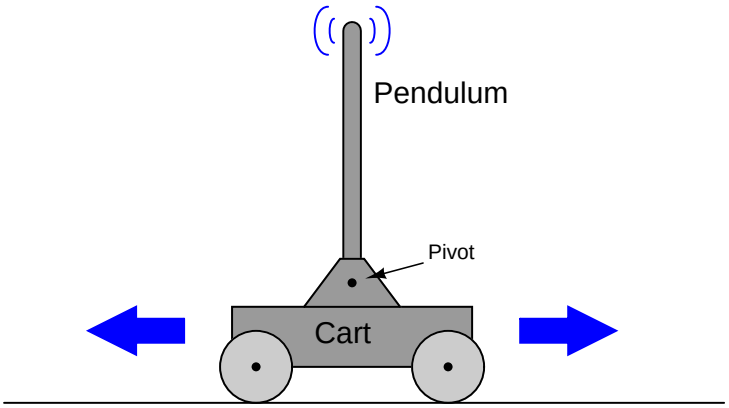
Summary:

- Integrating processes are characterized by a ramping of the process variable in response to a step-change in the control element value or load(s).
- This integration occurs as a result of either *mass flow imbalance* or *energy flow imbalance* in and out of the process
- Integrating processes are ideally controllable with proportional controller action alone.
- Integral controller action guarantees setpoint overshoot in a purely integrating process.
- Some integral controller action will be required in integrating processes to compensate for load changes.
- The amount of proportional controller action tolerable in an integrating process depends on the degree of time lag and process noise in the system. Too much proportional action will result in oscillation (time lags) and/or erratic control element motion (noise).

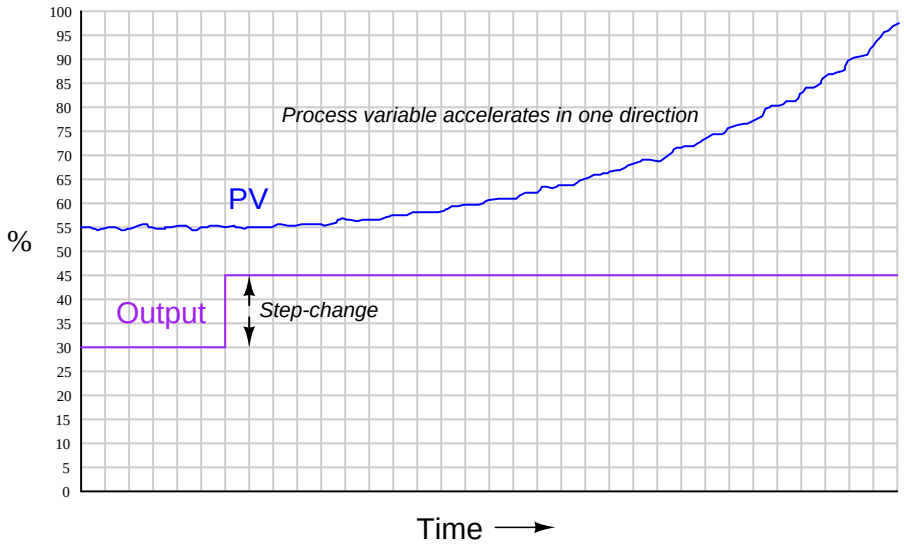
- An integrating process will become self-regulating if loads are allowed to vary with the process variable, thus requiring integral controller action to eliminate offset from setpoint

27.1.3 Runaway processes

A classic "textbook" example of a runaway process is an inverted pendulum: a vertical stick balanced on its end my moving the bottom side-to-side. Inverted pendula are typically constructed in a laboratory environment by fixing a stick to a cart by a pivot, then equipping the cart with wheels and a reversible motor to give it lateral control ability. A sensor (usually a potentiometer) detects the stick's angle from vertical, reporting that angle to the controller as the process variable. The cart's motor is the final control element:



The defining characteristic of a runaway process is its tendency to accelerate away from a condition of stability with no corrective action applied. To chart this behavior on a process trend:



A synonym for "runaway" is *negative self-regulation* or *negative lag*, because the process variable curve over time for a runaway process resembles the mathematical inverse of a self-regulating curve

with a lag time: it races away from the horizontal, while a self-regulating process variable draws closer and closer to the horizontal over time.

The "Segway^{TM}" personal transport device is a practical example of an inverted pendulum, with wheel motion controlled by a computer attempting to maintain the body of the vehicle in a vertical position. As the human rider leans forward, it causes the controller to spin the wheels with just the right amount of acceleration to maintain balance. There are many examples of runaway processes in motion-control applications, especially automated controls for vertical-flight vehicles such as helicopters and vectored-thrust aircraft such as the Harrier jet.

Fortunately, runaway processes are less common in the process industries. I say "fortunately" because these processes are notoriously difficult to control and usually pose more danger than inherently self-regulating processes. Many runaway processes are also nonlinear, making their behavior less intuitive to human operators. Exothermic chemical reaction processes are likely to exhibit "runaway" behavior, at least within certain ranges of operation.

An interesting example of a (potentially) runaway process is a nuclear (fission) reactor under certain conditions. Nuclear fission is a process by which the nuclei of specific types of atoms (most notably uranium-235 and plutonium-239) undergo spontaneous disintegration upon the absorption of an extra neutron, with the release of significant thermal energy and additional neutrons. A quantity of fissile material is subjected to a source of neutron particle radiation, which initiates the fission process, releasing massive quantities of heat which may then be used to boil water into steam and drive steam turbine engines to generate electricity. The "chain reaction" of neutrons splitting fissile atoms, which then eject more neutrons to split more fissile atoms, is inherently exponential in nature. The rate at which neutron activity within a fission reactor grows or decays over time is determined by the *multiplication factor*[4], and this factor is easily controlled by the insertion of neutron-absorbing *control rods* into the reactor core.

If the multiplication factor of a fission reactor were solely controlled by the positions of these control rods, it would be a classic "runaway" process, with the reactor's power level tending to increase toward infinity or decrease toward zero if the rods were at any position other than one yielding a multiplication factor of unity (1). This would make nuclear reactors extremely difficult (if not impossible) to safely control. Fortunately, there are ways to engineer the reactor core so that neutron activity *naturally* self-stabilizes without active control rod action. The liquid coolant used to transfer heat out of the reactor core and into a boiler to produce steam plays a double-role: it also offsets the multiplication factor inversely proportional to temperature. As the reactor core heats up, the coolant density changes, and the neutrons emitted by fission become less likely[5] to be captured

[4] When a nucleus of uranium or plutonium undergoes fission ("splits"), it releases more neutrons capable of splitting additional uranium or plutonium nuclei. The ratio of new nuclei "split" versus old nuclei "split" is the multiplication factor. If this factor has a value of one (1), the chain reaction will sustain at a constant power level, with each new generation of atoms "split" equal to the number of atoms "split" in the previous generation. If this multiplication factor exceeds unity, the rate of fission will increase over time. If the factor is less than one, the rate of fission will decrease over time. Like an inverted pendulum, the chain reaction has a tendency to "fall" toward infinite activity or toward no activity, depending on the value of its multiplication factor.

[5] The mechanism by which this occurs varies with the reactor design, and is too detailed to warrant a full explanation here. In pressurized, light-water reactors, which are the dominant design in the United States of America, this action occurs due to the water's ability to *moderate* (slow down) the velocity of neutrons. Slow neutrons have a greater probability of being "captured" by fissile nuclei than fast neutrons, and so the water's moderating ability will have a direct effect on the reactor core's multiplication factor. As a light-water reactor core increases temperature, the water becomes less dense and therefore less effective at moderating (slowing down) fast neutrons emitted by "splitting" nuclei. These fast(er) neutrons then "miss" the nuclei of atoms they would have otherwise split, effectively reducing the reactor's multiplication factor without any need for regulatory control rod motion. The reactor's power level

by other, fissile nuclei.

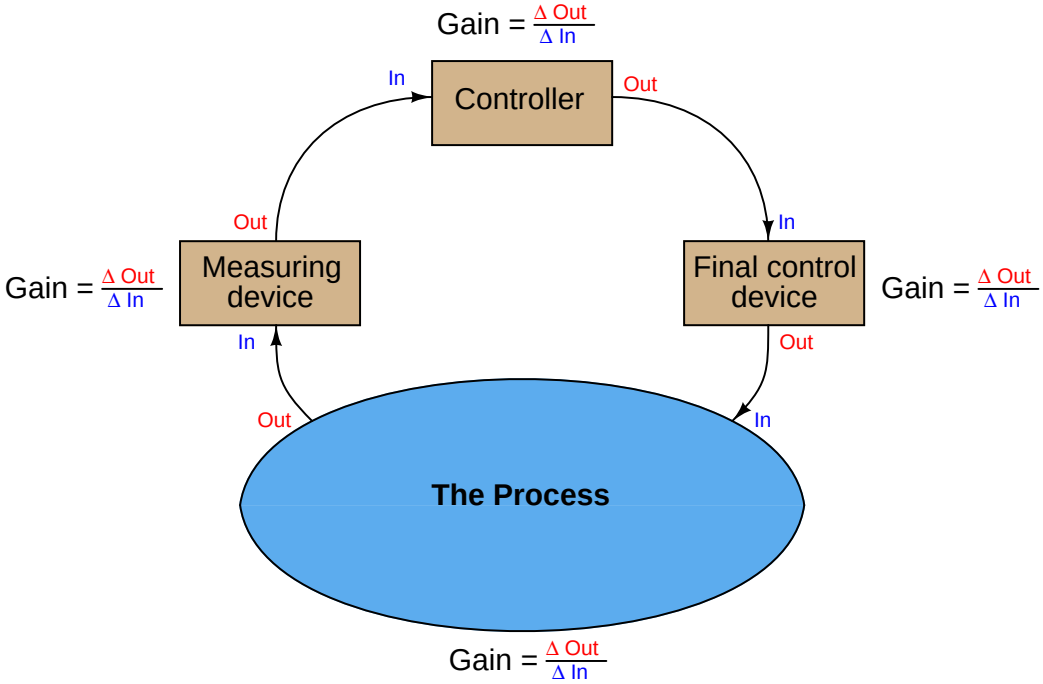
Some nuclear fission reactor designs are capable of "runaway" behavior, though. The ill-fated reactor at Chernobyl (Ukraine, Russia) was of a design where its power output could "run away" under certain operating conditions, and that is exactly what happened on April 26, 1986. The design of the Chernobyl reactor core was such that its cooling water did not provide a natural self-regulation characteristic, especially at low power levels where the reactor was being tested on the day of the accident. A combination of poor management decisions, unusual operating conditions, and bad design characteristics led to the reactor's destruction with massive amounts of radiation released into the surrounding environment. It stands at the time of this writing as the world's worst nuclear incident[6].

therefore stabilizes as it heats up, rather than "running away" to dangerously high levels, and may thus be classified as a *self-regulating* process.

[6] Discounting, of course, the intentional discharge of nuclear weapons, whose sole design purpose is to self-destruct in a "runaway" chain reaction.

27.1.4 Steady-state process gain

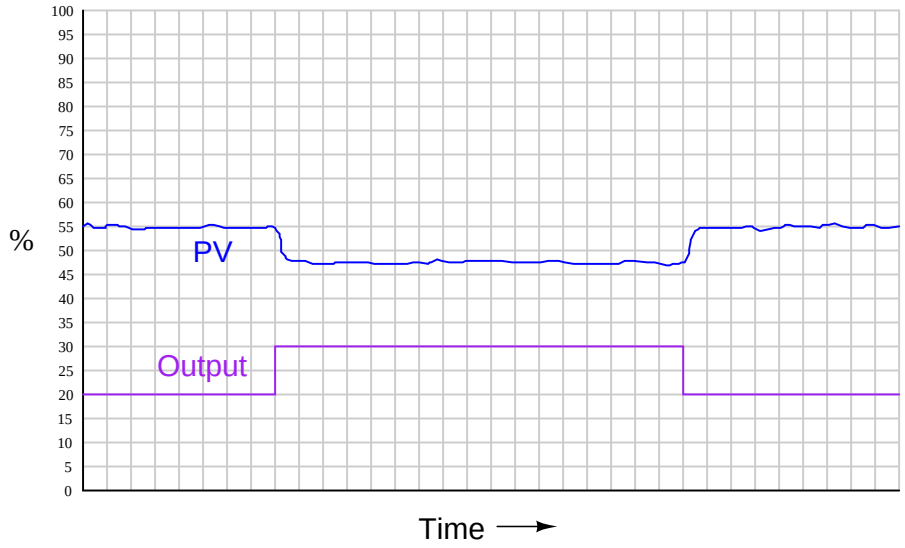
When we speak of a controller's *gain*, we refer to the aggressiveness of its proportional control action: the ratio of output change to input change. However, we may go a step further and characterize each component within the feedback loop as having its own gain (a ratio of output change to input change):



The gains intrinsic to the measuring device (transmitter), final control device (e.g. control valve), and the process itself are all important in helping to determine the necessary controller gain to achieve robust control. The greater the combined gain of transmitter, process, and valve, the less gain is needed from the controller. The less combined gain of transmitter, process, and valve, the more gain will be needed from the controller. This should make some intuitive sense: the more "responsive" a process appears to be, the less aggressive the controller needs to be in order to achieve stable control (and visa-versa).

These combined gains may be empirically determined by means of a simple test performed with the controller in "manual" mode. By placing the controller in manual mode (and thus disabling its automatic correction of process changes) and adjusting the output signal by some fixed amount, the resulting change in process variable may be measured and compared. If the process is self-regulating, a definite ratio of PV change to controller output change maybe determined.

For instance, examine this process trend graph showing a manual "step-change" and process variable response:



Here, the output step-change is 10% of scale, while the resulting process variable step-change is about 7.5%. Thus, the "gain" of the process[7] (together with transmitter and final control element) is approximately 0.75, or 75% (Gain = $\frac{7.5\%}{10\%}$). Incidentally, it is irrelevant that the PV steps *down* in response to the controller output stepping *up*. All this means is the process is reverse-responding, which necessitates *direct* action on the part of the controller in order to achieve negative feedback. When we calculate gains, we usually ignore directions (mathematical signs) and speak in terms of absolute values.

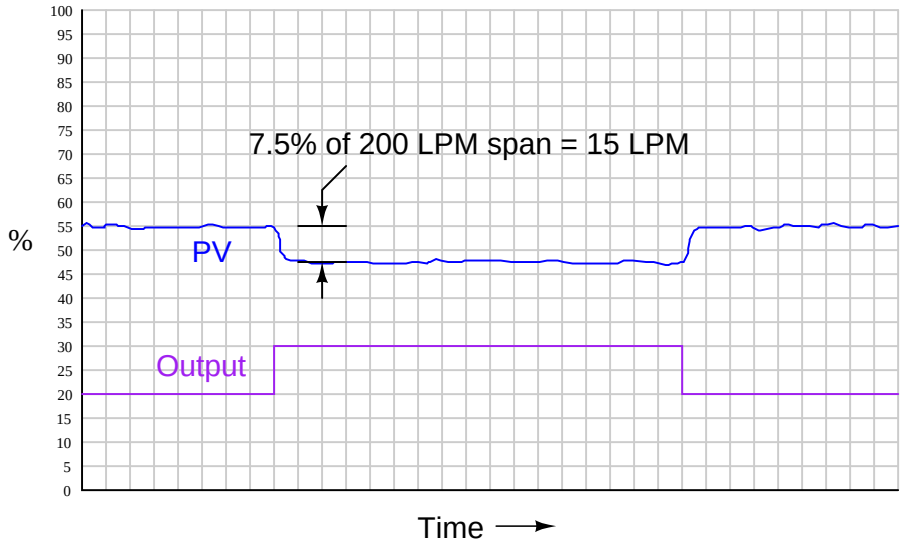
We commonly refer to this gain as the *steady-state gain* of the process, because the determination of gain is made after the PV settles to its self-regulating value.

Since, from the controller's perspective, the individual gains of transmitter, final control element, and physical process meld into one over-all gain value, process may be made to appear more or less responsive (more or less gain) by altering the gain of the transmitter and/or the gain of the final control element.

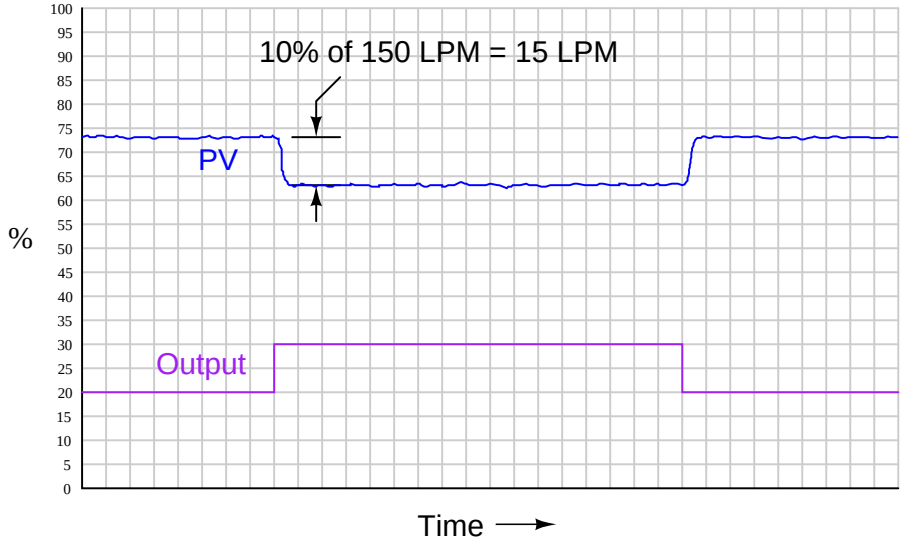
Consider, for example, if we were to reduce the span of the transmitter in this process. Suppose this was a flow control process, with the flow transmitter having a calibrated range of 0 to 200 liters per minute (LPM). If a technician were to re-range the transmitter to a new range of 0 to 150 LPM, what effect would this have on the apparent process gain?

[7] The general definition of gain is the ratio of output change over input change ($\frac{\Delta \text{Out}}{\Delta \text{In}}$). Here, you may have noticed we calculate process gain by dividing the process variable change (7.5%) by the controller output change (10%). If this seems "inverted" to you because we placed the *output* change value in the denominator of the fraction instead of the numerator, you need to keep in mind the perspective of our gain measurement. We are not calculating the gain of the controller, but rather the gain of the *process*. Since the output of the controller is the "input" to the process, it is entirely appropriate to refer to the 10% manual step-change as the change of *input* when calculating process gain.

To definitively answer this question, we must re-visit the process trend graph for the old calibrated range:



We see here that the 7.5% PV step-change equates to a change of 15 LPM given the flow transmitter's span of 200 LPM. However, if a technician re-ranges the flow transmitter to have just three-quarters that amount of span (150 LPM), the exact same amount of output step-change will *appear* to have a more dramatic effect on flow, even though the physical response of the process has the same as it was before:



From the controller's perspective – which only knows percent of signal range – the process gain

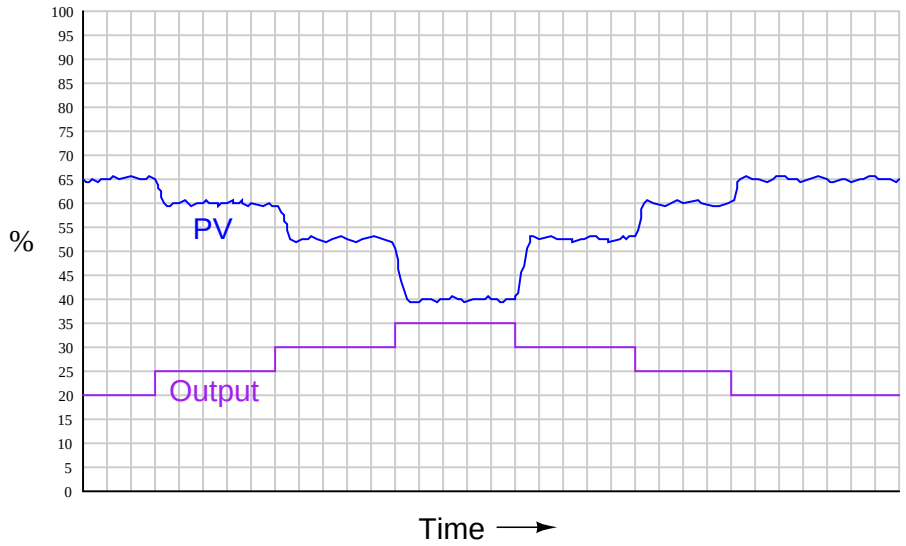
appears to have increased from 0.75 to 1, with nothing more than a re-ranging of the transmitter. Since the process is now "more responsive" to controller output signals than it was before, there may be a tendency for the loop to oscillate in automatic mode even if it did not oscillate previously with the old transmitter range. A simple fix for this problem is to decrease the controller's gain by the same factor that the process gain increased: we need to make the controller's gain $\frac{3}{4}$ what it was before, since the process gain is now $\frac{4}{3}$ what it was before.

The exact same effect occurs if the final control element – usually a control valve – is re-sized. If the same change in controller output signal percentage results in a different amount of influence on the process thanks to the final control element becoming more or less influential, the process gain will appear to change (at least from the controller's perspective), and re-tuning may become necessary to preserve robust control.

If and when re-tuning is needed to compensate for a change in loop instrumentation, all control modes should be proportionately adjusted. This is automatically done if the controller uses the *Ideal* or *ISA* PID equation, or if the controller uses the *Series* or *Interacting* PID equation[8]. All that needs to be done to an Ideal-equation controller in order to compensate for a change in process gain is to change that controller's proportional (P) constant setting. Since this constant directly affects all terms of the equation, the other control modes (I and D) will be adjusted along with the proportional term. If the controller happens to be executing the *Parallel* PID equation, you will have to manually alter all three constants (P, I, and D) in order to compensate for a change in process gain.

[8] For more information on different PID equations, refer to Section 26.9 beginning on page 1443.

A very important process characteristic for us to be aware of is how *consistent* process gain is over the measurement range. It is entirely possible (and in fact very likely) that a process may be more responsive (have higher gain) in some areas of control than in others. Take for instance this hypothetical trend showing process response to a series of manual-mode step-changes:



Note how the PV changes about 5% for the first 5% step-change in output, corresponding to a process gain of 1. Then, the PV changes about 7.5% for the next 5% output step-change, for a process gain of 1.5. The final increasing 5% step-change yields a PV change of about 12.5%, a process gain of 2.5. Clearly, the process being controlled here is not equally responsive throughout the measurement range. This is a concern to us in tuning the PID controller because any set of tuning constants that work well to control the process around a certain setpoint may not work as well if the setpoint is changed to a different value, simply because the process may be more or less responsive at that different process variable value.

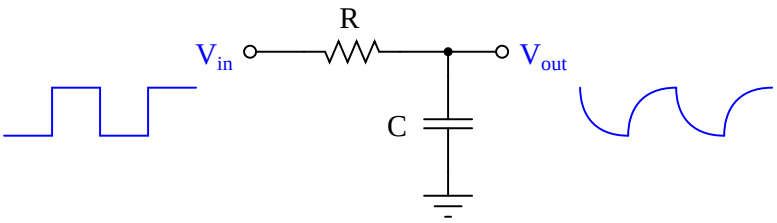
Inconsistent process gain is a problem inherent to many different process types, which means it is something you will need to be aware of when investigating a process prior to tuning the controller. The best way to reveal inconsistent process gain is to perform a series of step-changes to the controller output while in manual mode, "exploring" the process response throughout the safe range of operation.

Compensating for inconsistent process gain is much more difficult than merely detecting its presence. If the gain of the process follows an unchanging progression from one end of the range to the other (e.g. low gain at low output values and high gain at high output values, or visa-versa), a control valve with a different characteristic may be applied to counter-act the process gain.

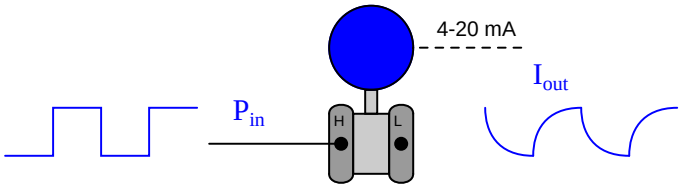
If the process gain follows some pattern more closely related to PV value rather than controller output value, the best solution is a type of controller known as an *adaptive gain controller*. In an adaptive gain controller, the proportional setting is made to vary in a particular way as the process changes, rather than be a fixed constant set by a human technician or engineer.

27.1.5 Lag time

If a square-wave signal is applied to an RC passive integrator circuit, the output signal will appear to have a "sawtooth" shape, the crisp rising and falling edges of the square wave replaced by damped curves:

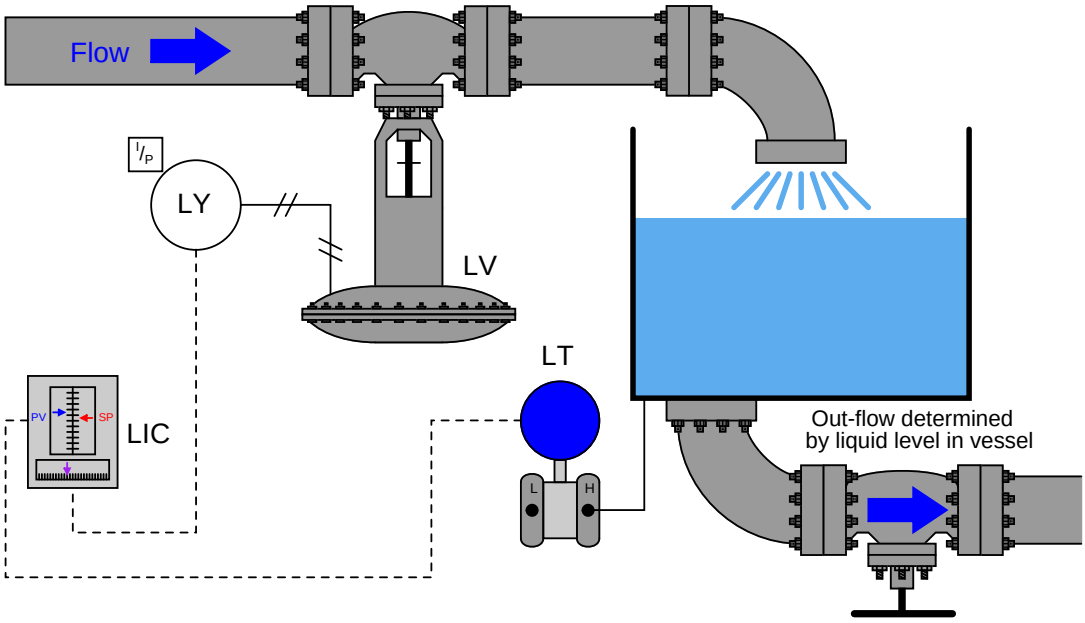


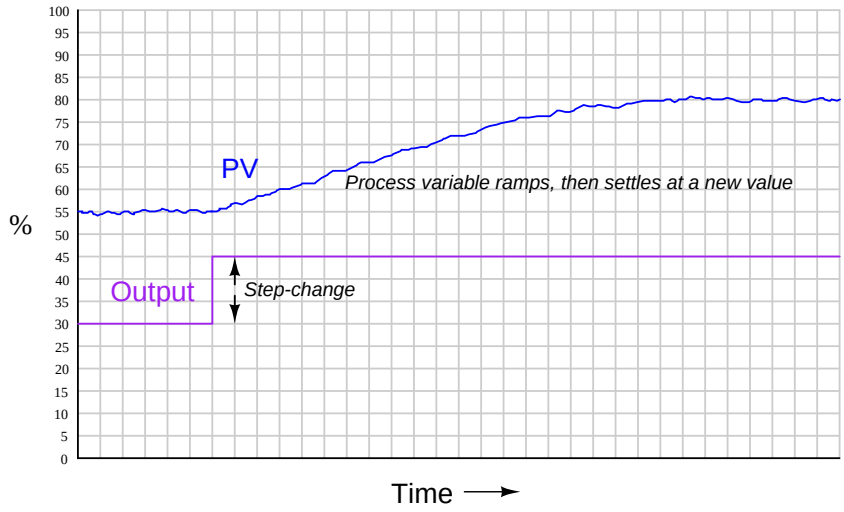
Most mechanical and chemical processes exhibit a similar tendency: an "inertial" opposition to rapid changes. Even instruments themselves naturally[9] damp sudden stimuli. We could have just as easily subjected a pressure transmitter to a series of pressure pulses resembling square waves, and witnessed the output signal exhibit the same damped response:



[9] It is also possible to *configure* many instruments to deliberately damp their response to input conditions. See Section 17.3 on page 733 for more detail.

The gravity-drained level-control process highlighted in an earlier subsection exhibits a very similar response to a sudden change in control valve position:





For any particular flow rate into the vessel, there will be a final (self-regulating) point where the liquid level "wants" to settle[10]. However, the liquid level does not *immediately* achieve that new

[10] Assuming a constant discharge valve position. If someone alters the hand valve's position, the relationship between incoming flow rate and final liquid level changes.

level if the control valve jumps to some new position, owing to the "capacity" of the vessel and the dynamics of gravity flow.

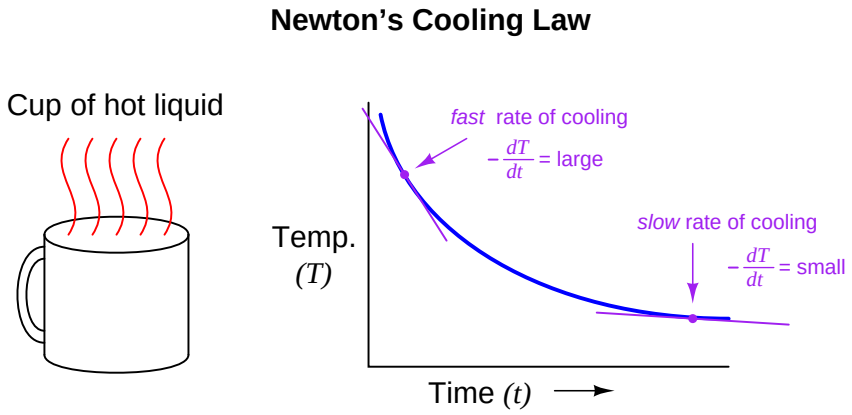
Any physical behavior exhibiting the same "settling" behavior over time may be said to illustrate a *first-order lag*. A classic "textbook" example of a first-order lag is a cup of hot liquid, gradually equalizing with room temperature. The liquid's temperature drops rapidly at first, but then slows its approach to ambient temperature as time progresses. This natural tendency is described by *Newton's Law of Cooling*, mathematically represented in the form of a *differential equation* (an equation containing a variable along with one or more of its derivatives). In this case, the equation is a *first-order* differential equation, because it contains the variable for temperature (T) and the first derivative of temperature ($\frac{dT}{dt}$) with respect to time:

$$\frac{dT}{dt} = -k(T - T_{ambient})$$

Where,

T = Temperature of liquid in cup
$T_{ambient}$ = Temperature of the surrounding environment
k = Constant representing the thermal conductivity of the cup
t = Time

All this equation tells us is that the rate of cooling ($\frac{dT}{dt}$) is directly proportional ($-k$) to the difference in temperature between the liquid and the surrounding air ($T - T_{ambient}$). The hotter the temperature, the faster the object cools (the faster rate of temperature fall):



A general solution to this equation is as follows:

$$T = \left(\frac{T_{initial} - T_{ambient}}{e^{\frac{t}{\tau}}} \right) + T_{ambient}$$

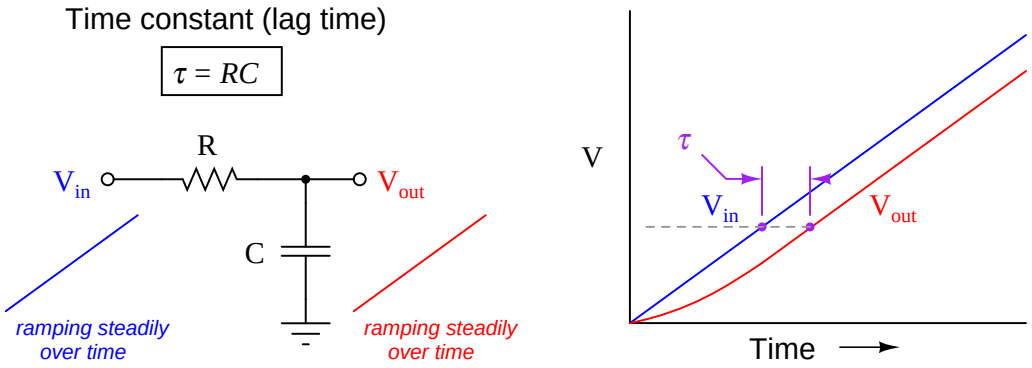
Where,

T = Temperature of liquid in cup
$T_{initial}$ = Temperature of liquid at time ($t = 0$)
$T_{ambient}$ = Temperature of the surrounding environment
e = Euler's constant
k = Constant representing the thermal conductivity of the cup
t = Time
τ = "Time constant" of the system

This mathematical analysis introduces a descriptive quantity of the system: something called a *time constant*. The "time constant" of a first-order system is the amount of time necessary for the system to come to within 36.8% (e^{-1}) of its final value (i.e. the time required for the system to go 63.2% of the way from the starting point to its ultimate settling point: $1 - e^{-1}$). After two time-constants' worth of time, the system will have come to within 13.5% (e^{-2}) of its final value (gone 86.5% of the way: $1 - e^{-2}$); after three time-constants' worth of time, to within 5% (e^{-3}) of the final value, (gone 95% of the way: $1 - e^{-3}$). After five time-constants' worth of time, the system will be within 1% (e^{-5}) of its final value, which is often close enough to consider it "settled" for most practical purposes.

Students of electronics will immediately recognize this concept, since it is widely used in the analysis and application of capacitive and inductive circuits. However, you should recognize the fact that the concept of a "time constant" for capacitive and inductive electrical circuits is but one case of a more general phenomenon. Literally *any* physical system described by the same first-order differential equation may be said to have a "time constant." Thus, it is perfectly valid for us to speak of a hot cup of coffee as having a time constant (τ), and to say that the coffee's temperature will be within 1% of room temperature after five of those time constants have elapsed.

In the world of process control, it is more customary to refer to this as a *lag time* than a *time constant*, but these are really interchangeable terms. The term "lag time" makes sense if we consider a first-order system *driven* to achieve a constant rate of change. For instance, if we subjected our RC circuit to a ramping voltage so that the output ramped as well instead of passively settling at some final value, we would find that the amount of time separating equal input and output voltage values was equal to this time constant (in an RC circuit, $\tau = RC$):

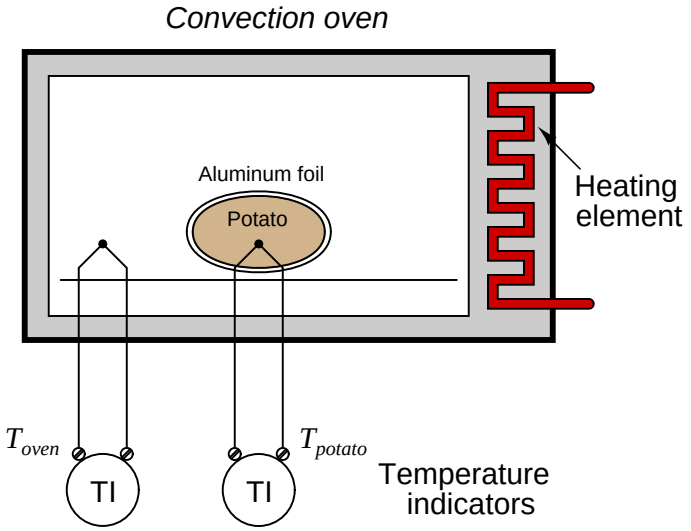


In other words, when a first-order process is ramping at any constant rate, the difference in time between when the process variable reaches a certain value and when it *would have* reached that same value were it not for the existence of lag in the system, is the lag time. This is the length in time that the ramping output *lags behind* the ramping input, regardless of the ramp rate.

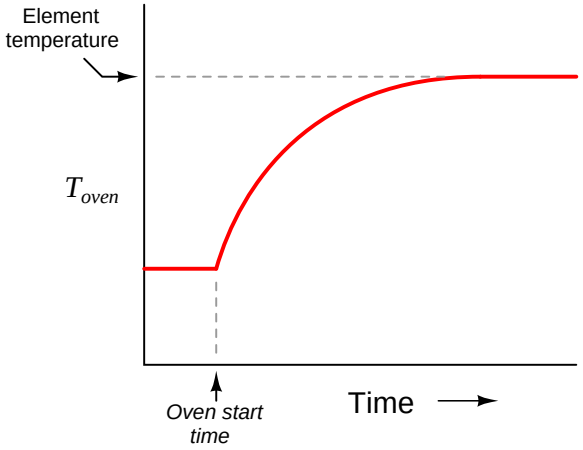
When an engineer or a technician describes a process being "fast" or "slow," they generally refer to the magnitude of the lag time. This makes lag time very important to our selection of PID controller tuning values. Integral and derivative control actions in particular are sensitive to the amount of lag time in a process, since both those actions are time-based. "Slow" processes (i.e. process types having large lag times) cannot tolerate aggressive integral action, where the controller "impatiently" winds the output up or down at a rate that is too rapid for the process to respond to. Derivative action, however, is generally useful on processes having large lag times.

27.1.6 Multiple lags (orders)

Simple, self-regulating processes tend to be first-order: that is, they have only one mechanism of lag. More complicated processes often consist of multiple sub-processes, each one with its own lag time. Take for example a convection oven, heating a potato. Being instrumentation specialists in addition to cooks, we decide to monitor both the oven temperature and the potato temperature using thermocouples and remote temperature indicators:



The oven itself is a first-order process. Given enough time and sufficiently thorough air circulation, the oven's air temperature will eventually self-stabilize at the heating element's temperature. If we graph its temperature over time with the heater power fixed in "manual" mode (no thermostat to control it), we will see a classic first-order function:

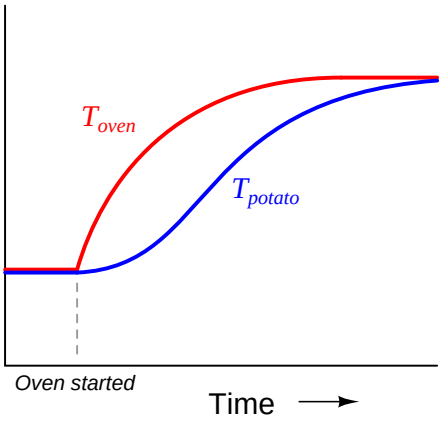


The potato forms another first-order process, absorbing heat from the air within the oven (heat

transfer by convection), gradually warming up until its temperature (eventually) reaches that of the oven[11]. From the perspective of the heating element to the oven air temperature, we have a first-order process. From the perspective of the heating element to the potato, however, we have a *second*-order process.

Intuition might lead you to believe that a second-order process is just like a first-order process – except slower – but that intuition would be wrong. Cascading two first-order lags creates a fundamentally different time dynamic. In other words, two first-order lags do not simply result in a *longer* first-order lag, but rather a second-order lag with its own unique characteristics.

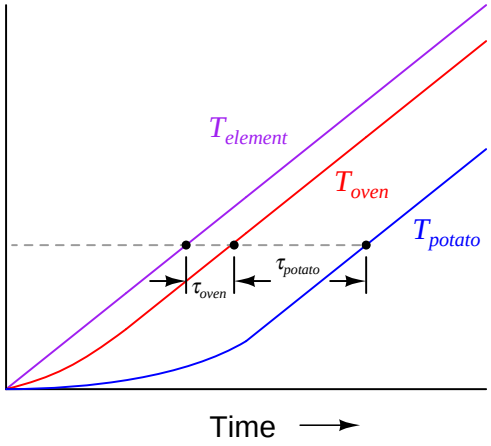
If we superimpose a graph of the potato temperature with a graph of the oven temperature (once again assuming constant power output from the heating element, with no thermostatic control), we will see that the *shape* of this second-order lag is different. The curve now has an "S" shape, rather than a consistent downward concavity:



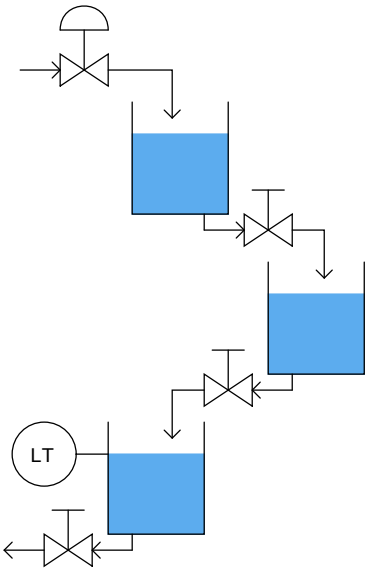
This, in fact, is the tell-tale signature of multiple lags in a process: an "S"-shaped curve rather than the characteristically abrupt initial rise of a first-order curve.

[11] Given the presence of water in the potato which turns to steam at 212 ^{o}F, things are just a bit more complicated than this, but let's assume a completely dry potato for now!

If we were able to ramp the heater power at a constant rate and graph the heater element, air, and potato temperatures, we would clearly see the separate lag times of the oven and the potato as offsets in time at any given temperature:



As another example, let us consider the control of level in three cascaded, gravity-drained vessels:



From the perspective of the level transmitter on the last vessel, the control valve is driving a *third-order* process, with three distinct lags cascaded in series. This would be a challenging process to control, and not just because of the possibility of the intermediate vessels overflowing (since their levels are not being measured)!

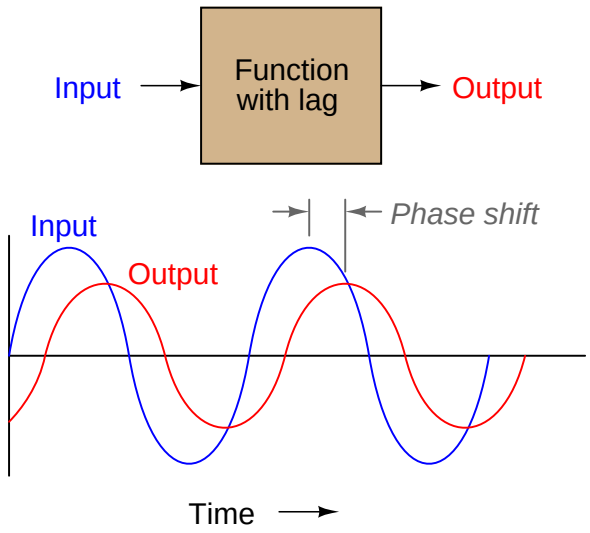
When we consider the dynamic response of a process, we are usually concerned primarily with the physical process itself. However, the instruments attached to that process also influence lag

orders and lag times. As discussed in the previous subsection, almost every physical function exhibits some form of lag. Even the instruments we use to measure process variables have their own (usually very short) lag times. Control valves may have substantial lag times, measured in the tens of seconds for some large valves. Thus, a "slow" control valve exerting control over a first-order process effectively creates a second-order loop response. Thermowells used with temperature sensors such as thermocouples and RTDs can also introduce lag times into a loop (especially if the sensing element is not fully contacting the bottom of the well!).

This means it is nearly impossible to have a control loop with a purely first-order response. Many real loops come close to being first-order, but only because the lag time of the physical process swamps (dominates) the relatively tiny lag times of the instruments. For inherently fast processes such as liquid flow and liquid pressure control, however, the process response is so fast that even short time lags in valve positioners, transmitters, and other loop instruments significantly alter the loop's dynamic character.

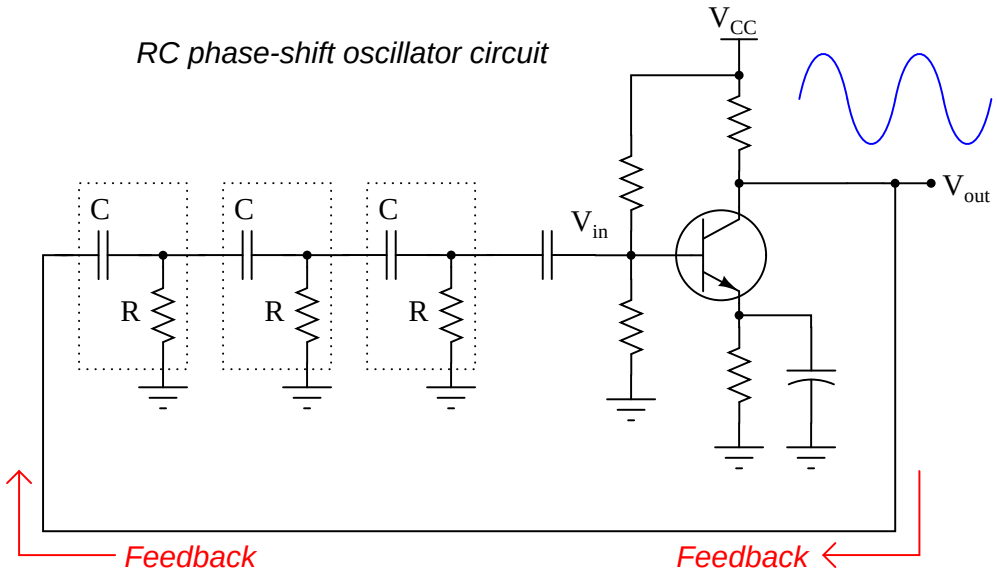
Multiple-order lags are relevant to the issue of PID loop tuning because they make the process harder to control with proportional and integral actions. The more lags there are in a system, the more delayed and "detached" the process variable becomes from the influence of the controller's output signal.

A mathematically convenient way to model the lags in a system is in terms of *phase shift* when driven by a sinusoidal-shaped stimulus. A function exhibiting lag will cause the outgoing waveform to lag behind the input waveform by a certain number of degrees at one frequency. The exact amount of phase shift usually depends on frequency – the higher the frequency, the more phase shift (to a maximum of -90^o for a first-order lag):



The phase shifts of multiple, cascaded lag functions (or processes, or physical effects) add up. This means each lag in a system contributes an additional negative *phase shift* to the loop. This may be a bad thing for negative feedback, which by definition is a 180^o phase shift. If sufficient lags exist in a system, the total loop phase shift may approach 360^o, in which case the feedback becomes *positive* (regenerative): a necessary condition for oscillation.

It is worthy to note that multiple-order lags are constructively applied in electronics when the express goal is to create oscillations. If a series of RC "lag" networks are used to feed the output of an inverting amplifier circuit back to its input with sufficient signal strength intact[12], and those networks introduce another 180 degrees of phase shift, the total loop phase shift will be 360^o (i.e. positive feedback) and the circuit will self-oscillate. This is called an *RC phase-shift oscillator* circuit:



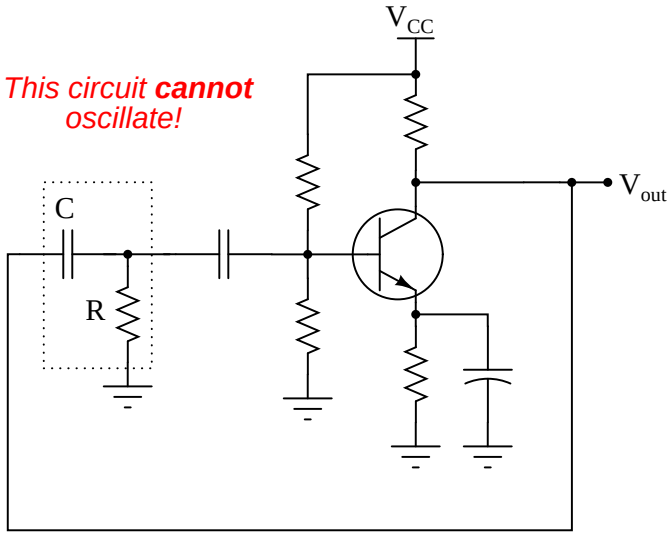
The amplifier works just like a proportional-only process controller, with action set for negative feedback. The resistor-capacitor networks act like the lags inherent to the process being controlled. Given enough controller (amplifier) gain, the cascaded lags in the process (RC networks) create the perfect conditions for self-oscillation. The amplifier creates the first 180^o of phase shift (being inverting in nature), while the RC networks collectively create the other 180^o of phase shift to give a total phase shift of 360^o (positive, or *regenerative* feedback).

In theory, the most phase shift a single RC network can create is -90^o, but even that is not practical[13]. This is why more than two RC phase-shifting networks are required for successful operation of an RC phase-shift oscillator circuit.

[12] The conditions necessary for self-sustaining oscillations to occur is a total phase shift of 360^o *and* a total loop gain of 1. Merely having positive feedback *or* having a total gain of 1 or more will not guarantee self-sustaining oscillations; both conditions must simultaneously exist. As a measure of how close any feedback system is to this critical confluence of conditions, we may quantity a system's *phase margin* (how many degrees of phase shift the system is away from 360^o while at a loop gain of 1) and/or a system's *gain margin* (how many decibels of gain the system is away from 0 dB while at a phase shift of 360^o). The less phase or gain margin a feedback system has, the closer it is to a condition of instability.

[13] At maximum phase shift, the gain of any first-order RC network is zero. Both phase shift and attenuation in an RC lag network are frequency-dependent: as frequency increases, phase shift grows larger (from 0^o to a maximum of -90^o) and the output signal grows weaker. At its theoretical maximum phase shift of exactly -90^o, the output signal would be reduced to nothing!

As an illustration of this point, the following circuit is not capable[14] of self-oscillation. Its lone RC phase-shifting network cannot create the -180^o phase shift necessary for the overall loop to have positive feedback and oscillate:



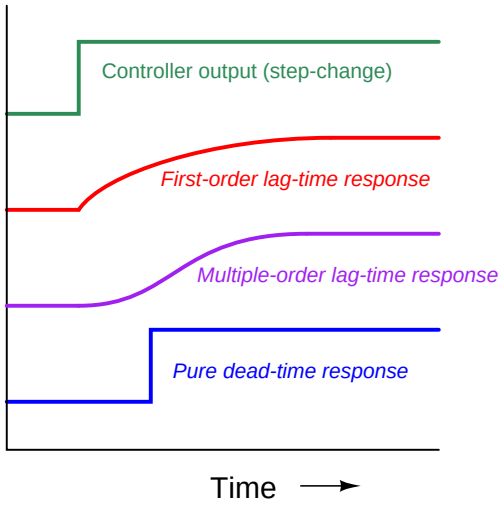
The RC phase-shift oscillator circuit design thus holds a very important lesson for us in PID loop tuning. It clearly illustrates how multiple orders of lag are a more significant obstacle to robust control than a single lag time of any magnitude. A purely first-order process will tolerate enormous amounts of controller gain without ever breaking into oscillations, because it lacks the phase shift necessary to self-oscillate. This means – barring any other condition limiting our use of high gain, such as process noise – we may use very aggressive proportional-only action (e.g. gain values of 20 or more) to achieve robust control on a first-order process[15]. Multiple-order processes are less forgiving of high controller gains, because they *are* capable of generating enough phase shift to self-oscillate.

[14] In its pure, theoretical form at least. In practice, even a single-lag circuit may oscillate given enough gain due to the unavoidable presence of parasitic capacitances and inductances in the wiring and components causing multiple orders of lag (and even some dead time). By the same token, even a "pure" first-order process will oscillate given enough controller gain due to unavoidable lags and dead times in the field instrumentation (especially the control valve). The point I am trying to make here is that there is more to the question of stability (or instability) than loop gain.

[15] Truth be told, the same principle holds for purely integrating processes as well. A purely integrating process *always* exhibits a phase shift of -90^o at any frequency, because that is the nature of integration in calculus. A purely first-order lag process will exhibit a phase shift anywhere from 0^o to -90^o depending on frequency, but never more lagging than -90^o, which is not enough to turn negative feedback into positive feedback. In either case, so long as we don't have process noise to deal with, we can increase the controller's gain all the way to *eleven*. If that last sentence (a joke) does not make sense to you, be sure to watch the 1984 movie *This is Spinal Tap* as soon as possible. Seriously, I have used controller gains as high as *50* on low-noise, first-order processes such as furnace temperature control. With such high gain in the controller, response to setpoint and load changes is quite swift, and integral action is almost unnecessary because the offset is naturally so small.

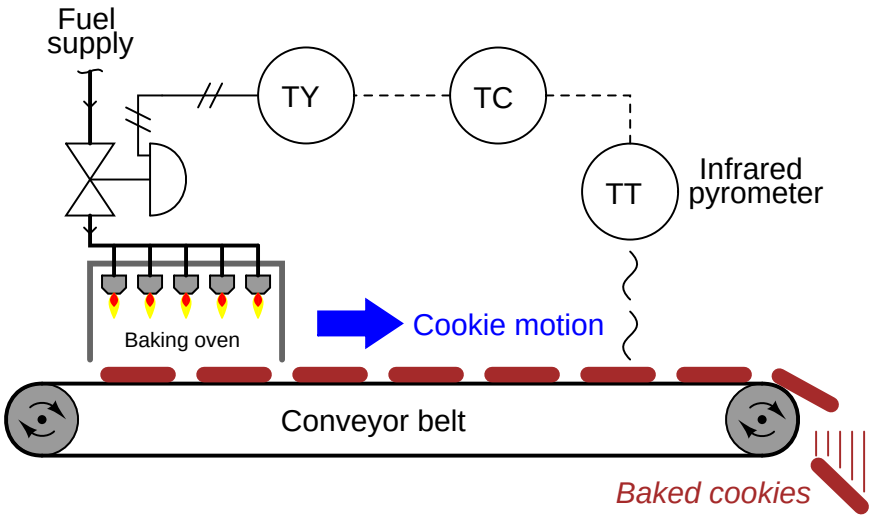
27.1.7 Dead time

Lag time refers to a damped response from a process, from a change in manipulated variable (e.g. control valve position) to a measured change in process variable: the initial effect of a change in controller output is immediately seen, but the final effect takes time to develop. *Dead time*, by contrast, refers to a period of time during which a change in manipulated variable produces *no effect whatsoever* in the process variable: the process appears "dead" for some amount of time before showing a response. The following graph contrasts first-order and multiple-order lag times against pure dead time, as revealed in response to a manual step-change in the controller's output (an "open-loop" test of the process characteristics):



Although the first-order response does takes some time to settle at a stable value, there is no time delay between when the output steps up and the first-order response *begins* to rise. The same may be said for the multiple-order response, albeit with a slower rate of initial rise. The dead-time response, however, is actually delayed some time after the output makes its step-change. There is a period of time where the dead-time response does *absolutely nothing* following the output step-change.

Dead time is also referred to as *transport delay*, because the mechanism of dead time is often a time delay caused by the transportation of material at finite speed across some distance. The following cookie-baking process has dead time by virtue of the time delay inherent to the cookies' journey from the oven to the temperature sensor:

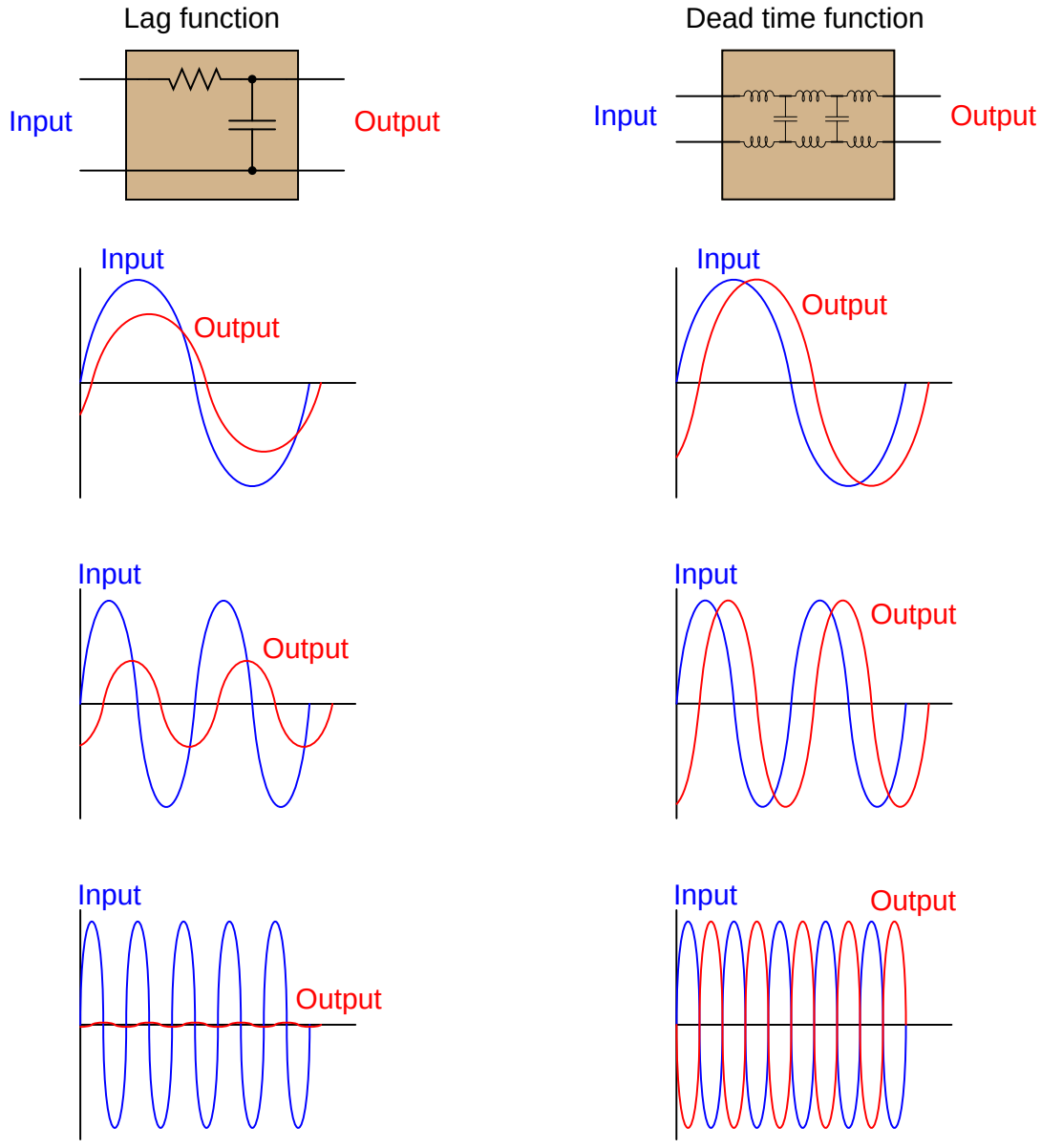


Dead time is a far worse problem for feedback control systems than lag time. The reason why is best understood from the perspective of phase shift: the delay (measured in degrees of angular displacement) between input and output for a system driven by a sinusoidal stimulus. Excessive phase shift in a feedback system makes possible self-sustaining oscillations, turning what is supposed to be negative feedback into positive feedback. Systems with lag produce phase shift that is frequency-dependent (the greater the frequency, the more the output "lags" behind the input), but this phase shift has a natural limit. For a first-order lag function, the phase shift has an absolute maximum value of -90^o; second-order lag functions have a theoretical maximum phase shift of -180^o; and so on. Dead time functions also produce phase shift that increases with frequency, but there is no ultimate limit to the amount of phase shift. This means a single dead-time element in a feedback control loop is capable of producing *any* amount of phase shift given the right frequency[16]. What is more, the gain of a dead time function usually does not diminish with frequency, unlike the gain of a lag function.

Recall that a feedback system will self-oscillate if two conditions are met: a total phase shift of 360^o (or -360^o: the same thing) and a total loop gain of at least one. Any feedback system meeting these criteria will oscillate, be it an electronic amplifier circuit or a process control loop. In the interest of achieving robust process control, we need to prevent these conditions from ever occurring simultaneously.

[16] A sophisticated way of saying this is that a dead-time function has no *phase margin*, only *gain margin*. All that is needed in a feedback system with dead time is sufficient gain to make the system oscillate.

A visual comparison between the phase shifts and gains exhibited by lag versus dead time functions may be seen here, the respective functions modeled by the electrical entities of a simple RC network (lag) and a transmission line (dead time):

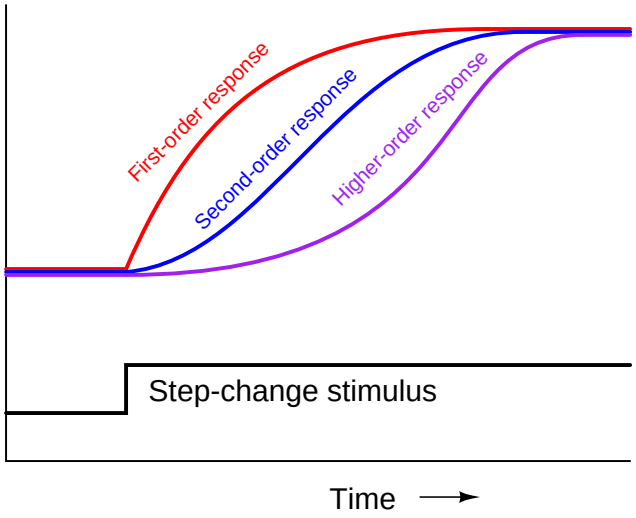


As frequency increases, the lag function's phase shift asymptotically approaches -90^o while its attenuation becomes more severe. Ultimately, when the phase shift reaches its maximum of -90^o, the output signal amplitude is reduced to nothing. By contrast, the dead time function's phase

shift grows linearly with frequency (to -180^o and beyond!) while its attenuation remains unchanged. Clearly, dead time better fulfills the dual criteria of sufficient phase shift and sufficient loop gain needed for feedback oscillation than lag time, which is why dead time is more detrimental to robust process control than lag time.

Pure dead-time processes are rare. Usually, an industrial process will exhibit at least some degree of lag time in addition to dead time. As strange as it may sound, this is a fortunate for the purpose of feedback control. The presence of lag(s) in a process guarantees a degradation of loop gain with frequency increase, which may help avoid oscillation in such a loop. The greater the ratio between dead time and lag time in a loop, the more unstable it will tend to be.

The appearance of dead time may be created in a process by the cascaded effect of multiple lags. As mentioned in an earlier subsection, multiple lags create a process response to step-changes that is "S"-shaped, responding gradually at first instead of immediately following the step-change. Given enough lags acting in series, the beginning of this "S" curve may be so flat that it appears "dead:"



While dead time may be impossible to eliminate in some processes, it should be minimized wherever possible due to its detrimental impact on feedback control. Once an open-loop (manual-mode step-change) test on a process confirms the existence of dead time, the source of dead time should be identified and eliminated if at all possible.

One technique applied to the control of dead-time-dominant processes is a special variation of the PID algorithm called *sample-and-hold*. In this variation of PID, the controller effectively alternates between "automatic" and "manual" modes according to a pre-programmed cycle. For a short period of time, it switches to "automatic" mode in order to "sample" the error (PV − SP) and calculate a new output value, but then switches right back into "manual" mode ("hold") so as to give time for the effects of those corrections to propagate through the process dead time before taking another sample and calculating another output value. This sample-and-hold cycle of course slows the controller's response to changes such as setpoint adjustments and load variations, but it does allow for more aggressive PID tuning constants than would otherwise work in a continuously sampling controller.

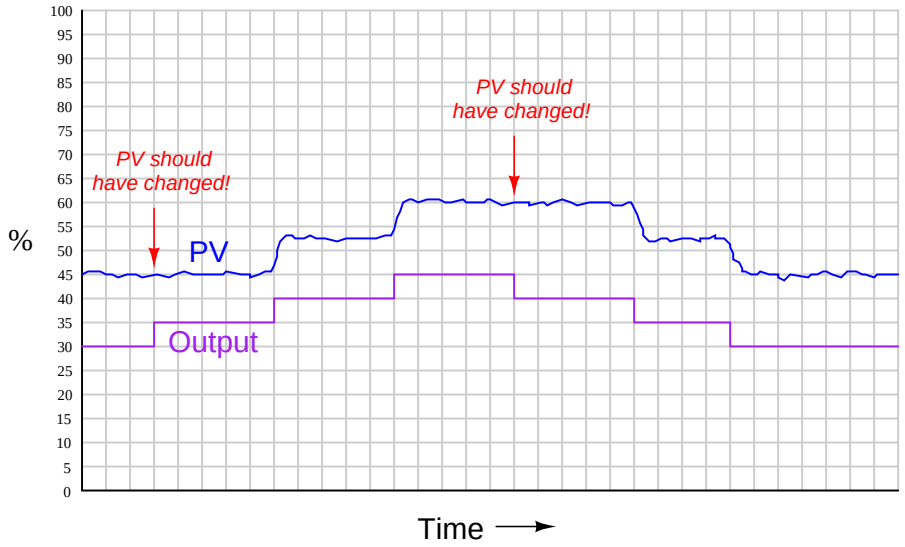
27.1.8 Hysteresis

A detrimental effect to feedback control is a characteristic known as *hysteresis*: a lack of responsiveness to a change in direction. Although hysteresis typically resides with instruments rather than the physical process they connect to, it is most easily detected by a simple open-loop ("step-change") test with the controller in manual mode just like all the important process characteristics (self-regulating versus integrating, steady-state gain, lag time, dead time, etc.).

The most common source of hysteresis is found in pneumatically-actuated control valves possessing excess stem friction. The "jerky" motion of such a valve to smooth increases or decreases in signal is sometimes referred to as *stiction*. Similarly, a pneumatically-actuated control valve with excess friction will be unresponsive to small reversals in signal direction. To illustrate, this means the control valve's stem position will not be the same at a *rising* signal value of 50% (typically 12 mA, or 9 PSI) as it will be at a *falling* signal value of 50%.

Control valve stiction may be quite severe in valves with poor maintenance histories, and/or lacking positioners to correct for deviations between controller signal value and actual stem position. I have personally encountered control valves with hysteresis values in excess of 10%[17], and have heard of even more severe cases.

Detecting hysteresis in a control loop is as simple as performing "up-and-down" tests of the controller output signal in manual mode. The following trend shows how hysteresis might appear in a self-regulating process:



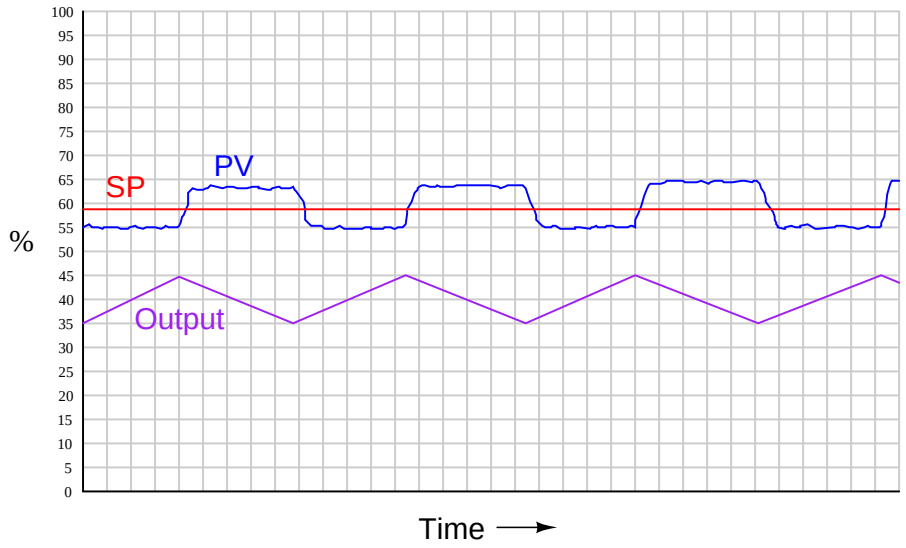
Note how the PV fails to respond to the first upward change in output, and also how it fails to respond to the first reversal of output signal direction. In this particular open-loop test, the loop exhibits a hysteresis of about 5%, since that is how much the output signal may be reversed without seeing any change in PV.

[17] A 10% hysteresis value means that the signal must be changed by 10% following a reversal of direction before any movement is seen from the valve stem.

Hysteresis is a problem in feedback control because it essentially acts like a variable dead time. Recall that "dead time" was defined as a period of time during which a change in manipulated variable produces no effect in the process variable: the process appears "dead" for some amount of time before showing a response. If a change in controller output (manipulated variable) is insufficient to overcome the hysteresis inherent to a control valve or other component in a loop, the process variable will not respond to that output signal change at all. Only when the manipulated variable signal continues to change sufficiently to overcome hysteresis will there be a response from the process variable, and the time required for that to take place depends on how soon the controller's output happens to reach that critical value. If the controller's output moves quickly, the "dead time" caused by hysteresis will be short. If the controller's output changes slowly over time, the "dead time" caused by hysteresis will be longer.

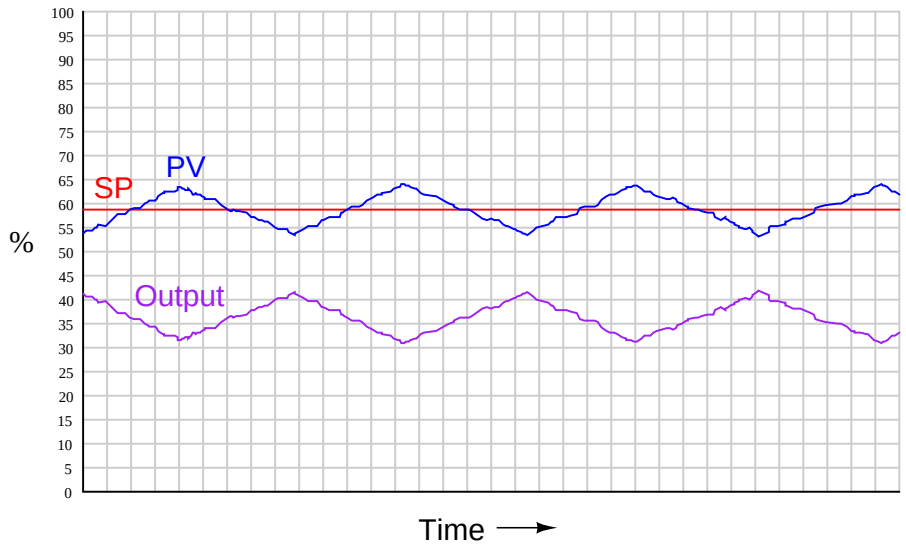
Another problem caused by hysteresis in a feedback loop occurs in combination with integral action, whether it be programmed into the controller or is inherent to the process (i.e. an *integrating* process). It is highly unlikely that a "sticky" control valve will happen to "stick" at exactly the right stem position required to equalize PV and SP. Therefore, the probability at any time of an error developing between PV and SP, or of an offset developing between the valve position and the equilibrium position required by an integrating process, is very great. This leads to a condition of guaranteed instability. For a self-regulating process with integral action in the controller, the almost guaranteed existence of PV − SP error means the controller output will ceaselessly ramp up and down as the valve first slips and sticks to give a positive error, then slips and sticks to give a negative error. For an integrating process with proportional action in the controller, the process variable will ceaselessly ramp up and down as the valve first sticks too far open, then too far closed to equalize process in-flow and out-flow which is necessary to stabilize the process variable. In either case, this particular form of instability is called a *slip-stick cycle*.

The following process trend shows a slip-stick cycle in a self-regulating process, controlled by an integral-only controller:



Note how the output ceaselessly ramps in a futile attempt to drive the process variable to setpoint. Once sufficient pressure change accumulates in the valve actuator to overcome static stem friction, the valve "slips" to a new position (not equal to setpoint), and the controller begins to ramp the output in the other direction.

The next trend shows a slip-stick cycle in an integrating process, controlled by a proportional-only controller:



It is very important to note that the problems created by a "sticky" control valve *cannot* be overcome by tuning the controller. For instance, if one were to de-tune the integral-only controller (i.e. longer time constant, or fewer repeats per minute) in the self-regulating process, it would *still* exhibit the same slip-stick behavior, only over a longer period (lower frequency) than before. If one were to de-tune the proportional-only controller (i.e. greater proportional band, or less gain) in the integrating process, the exact same thing would happen: a decrease in cycle frequency, but no elimination of the cycling. Furthermore, de-tuning the controller in either process would also result in less responsive (poorer) corrective action to setpoint and load changes. The only solution[18] to either one of these problems is to reduce or eliminate the friction inside the control valve.

27.2 Before you tune . . .

Much has been written about the benefits of robust PID control. Increased productivity, decreased equipment strain, and increased process safety are some of the advantages touted of proper PID tuning. What is often overlooked, though, are the negative consequences of poor PID controller tuning. If robust PID control can increased productivity, then poor PID control can decrease productivity. If a well-tuned system helps equipment run longer and safer, then a badly tuned system may increased failure frequency and safety incidents. The instrumentation professional should be mindful of this dichotomy when proceeding to tune a PID control system. One should never think there is "nothing to lose" by trying different PID settings. Tuning a PID controller is as serious a matter as the productivity and safety impact of the process itself.

PID tuning parameters are easy to access, which makes them a tempting place to begin for technicians looking to improve the performance of a feedback loop. Another temptation driving technicians to focus on controller tuning as a first step is the prestige associated with being able to tame an unruly feedback loop with a few magical adjustments to the controller's PID tuning constants. For those who do not understand PID control (and this constitutes the vast majority of the human population, even in the industrial world), there is something "magic" about being able to achieve robust control behavior simply by making small adjustments to numbers in a computer (or to knobs in an analog controller). The reality, though, is that many poorly-behaving control systems are that way not due (at least purely) to a deficit of proper PID tuning values, but rather to problems external to the controller which no amount of "tuning" will solve. Adjusting PID tuning constants as a *first step* is almost always a bad idea.

This section aims to describe and explain some of the recommended considerations prior to making adjustments to the tuning of a loop controller. These considerations include:

- Identifying operational needs (i.e. "How do the operators want the system to respond?")
- Identifying process and system hazards before manipulating the loop
- Identifying whether it is a tuning problem, a field instrument problem, and/or a design problem

[18] An alternate solution is to install a positioner on the control valve, which acts as a secondary (cascaded) controller seeking to equalize stem position with the loop controller's output signal at all times. However, this just "shifts" the problem from one controller to another. I have seen examples of control valves with severe packing friction which will *self-oscillate* their own positioners (i.e. the positioner will "hunt" back and forth for the correct valve stem position given a constant signal from the loop controller)! If valve stem friction can be minimized, it should be minimized.

27.2.1 Identifying operational needs

As defined elsewhere in this book, "robust" control is a stability of the process variable despite changes in load, fast response to changes in setpoint, minimal oscillation following either type of change, and minimal offset (error between setpoint and process variable) over time. However, these criteria are not equally valued in all processes, and neither are they equally attainable with simple PID control in all processes. It may be critical, for example, in a boiler water level control process to have fast response to changes in load, but minimal offset over time is not as important. It may be completely permissible to have a persistent 5% error between PV and SP in such a system, so long as the water level does not deviate much over 20% for any length of time due to load changes. In another process, such as liquid level control inside one stage ("effect") of a multi-stage ("multi-effect") evaporator system, a priority may be placed upon relatively steady flow control through the valve rather than steady level in the process. A level controller tuned for aggressive response to setpoint changes will cause large fluctuations in liquid flow rate to all successive stages ("effects") of the evaporator process in the event of a sudden load or setpoint change, which would be more detrimental to product quality than some deviation from setpoint in that one effect.

Thus, we see the need for whomever intends to tune a control system to determine what the operational needs of the system are. The operations personnel (operators, unit managers, process engineers) are your best resources here. Ultimately, they are your "customers." Your task is to give the customers the system performance they need to do their jobs best.

Keep in mind the following process control objectives, knowing that it will likely be impossible to achieve *all* of them with any particular PID tuning. Try to rank the relative importance of these objectives, then concentrate on achieving those most important, at the expense of those least important:

- Minimum change in PV (dynamic stability) with changes in load
- Fast response to setpoint changes (minimum dynamic error)
- Minimum overshoot/undershoot/oscillation following sudden load or setpoint changes
- Minimum error (PV − SP) over time
- Minimum valve velocity

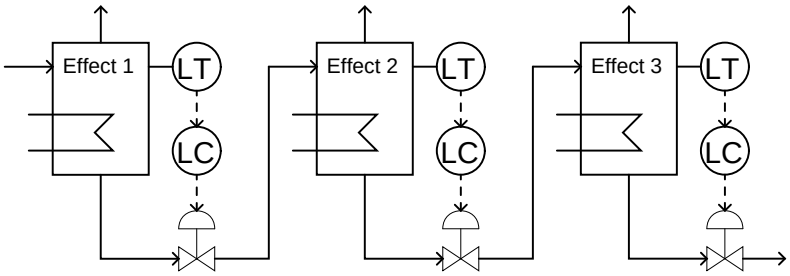
The control actions best suited for rapid response to load and/or setpoint changes are proportional and derivative. Integral action takes effect only *after* error has had time to develop, and as such cannot act as immediately as either proportional or derivative.

If the priority is to minimize overshoot, undershoot, and/or oscillations, the controller's response will likely need to be more sluggish than is typical. New setpoint values will take longer to achieve, and load changes will not be responded to with quite the same vigor. Derivative action may be helpful in some applications to "tame" the oscillatory tendencies of proportional and integral actions.

Minimum error over time can really only be addressed by integral action. No other controller action pays specific "attention" to the magnitude and duration of error. This is not to say that the process will work well on integral-only control, but rather that integral action will be absolutely necessary (i.e. a P-only or PD controller will not suffice).

Minimum valve velocity is a priority in processes where the manipulated variable has an effect on some *other* process in the system. For example, liquid level control in a multi-stage (multi-"effect") evaporator system where the discharge flow from one evaporator becomes the incoming flow for another evaporator, is a system where sharp changes in the manipulated variable of one control loop can upset downstream processes:

A multi-effect evaporator system



In other words, an aggressively-tuned level controller on an upstream evaporator (e.g. Effect 1) may achieve its goal of holding liquid level very steady in that evaporator by varying its out-going flow, but it will do so at the expense of causing level variations in all downstream evaporators. Cases such as this call for controller tuning (at least in the upstream effects) responding slowly to errors. Proportional action will very likely be limited to low gain values (high proportional band values), and derivative action (if any is used at all) should be set to respond only to the process variable, not to error (PV − SP). Understand that tuning a PID loop with the goal of minimizing valve motion *will* result in more deviation from setpoint than if the controller were tuned more aggressively.

27.2.2 Identifying process and system hazards

When students practice PID control in an Instrumentation program, they usually do so using computer simulation software and/or "toy" processes constructed in a lab environment. A potential disadvantage to this learning environment is a failure to recognize real problems that may develop when tuning an actual production process. Rarely will you find a completely isolated feedback loop in industry: generally there are interactive effects between control loops in a process, which means one cannot proceed to tune a loop with impunity.

A very important question to ask the operations personnel before tuning a loop is, "How far and how fast am I allowed to let the process variable increase and decrease?" Processes and process equipment may become dangerously unstable, for example, if certain temperatures become to high (or too low, as is the case in process liquids that solidify when cold). It is not uncommon for certain control loops in a process to be equipped with alarms, either hard or soft, that automatically *shut down* equipment if exceeded. Clearly, these "shutdown" limits must be avoided during the tuning of the process loop.

One should also examine the control strategy before proceeding to tune. Is this a cascaded loop? If so, the slave controller needs to be tuned before the master. Does this loop have feedforward added incorporated to handle load changes? If so, the effectiveness of that feedforward loop (gain, dynamic compensation) should be checked and adjusted before the feedback loop is tuned. Are there limits in this loop? Is this a selector or override control strategy? If so, you need to be able to clearly tell which loop components are selected, and which signals are being limited, at any given time.

Another consideration is whether or not the process is in a "normal" condition before you attempt to improve its performance. Ask the operations personnel if this is a typical day, or if there is some abnormal condition in effect (equipment shutdown, re-routing of flows, significantly different production rates, etc.) that might skew the response of the process loop to be tuned. Once again we see a need for input from the operations personnel, because they know the day-to-day behavior of the system better than anyone else.

27.2.3 Identifying the problem(s)

One of the questions I advise instrument technicians to ask of operators when diagnosing any process problem is simply, "How long has this problem existed?" The age of a problem can be a very important indicator of possible causes. If you were told that a problem suddenly developed after the last night shift, you would be inclined to suspect an equipment failure, or something else that could happen *suddenly* (e.g. a hand valve someone opened or shut when they shouldn't have). Alternatively, if you were told a problem has been in existence since the day the process was constructed, you would be more inclined to suspect an issue with system design or improper installation. This same diagnostic technique – obtaining a "history" of the "patient" – applies to loop tuning as well. A control loop that suddenly stopped working as it should might be suffering from an instrument failure (or an unauthorized change of controller parameters), whereas a chronically misbehaving loop would more likely be suffering from poor design, bad instrument installation, or a controller that was never tuned properly.

In either case, poor control is just as likely to be caused by field instrument problems as it is by incorrect PID tuning parameters. No PID settings can possibly compensate for faulty instrumentation, but it is possible for some instrument problems to be "masked" by controller tuning. Your first step in actually manipulating the control loop should be a check of instrument health. Thankfully, this is relatively easy to do by performing a series of "step-change" tests with the controller in manual mode. By placing the controller in manual and making small changes in output signal (remember to check with operations to see how far you are allowed to move the output, and how far you can let the PV drift!), you can determine much about the process and the loop instrumentation, including:

- Whether the PV signal is "noisy" (first turn off all damping in the controller and transmitter)
- How much "stiction" is in the control valve
- Whether the process is integrating or self-regulating
- Process gain (and whether this gain is stable or if it changes as PV changes)
- Process lag time and lag degree (first-order versus multiple-order)
- Process dead time

Such an open-loop test might reveal potential problems without pinpointing the exact nature or location of those problems. For example, a large lag time may be intrinsic to the process, or it may be the result of a poorly-installed sensor (e.g. a thermocouple not pushed to touch the bottom of its thermowell) or even a control valve in need of a volume booster or positioner. Dead time measured in an open-loop test may also be intrinsic to the process, or it could be the result of stiction in the valve. The only way to definitively identify the problem is to test the instruments themselves, ideally in the field location.

In order to obtain the best "view" of the process, you need to make sure the graphing trend display has sufficient resolution and responsiveness. If the trend fails to show fine details such as noise in the process, it is possible that the graphing device will be insufficient for your needs.

If this is the case, you may still perform response tests of the loop, but you will have to use some other instrument(s) to graph the controller and process actions. A modern tool useful for this

purpose is a portable computer with a data acquisition device connected, giving the computer the ability to read instrument signal voltages. Many data graphing programs exist for taking acquired data and plotting it over the time domain. Data acquisition modules with sample rates in the thousands of samples per second are available for very modest prices.

27.2.4 Final precautions

Be prepared to document your work! This means capturing and recording "screen shot" images of process trend graphs, both for the initial open-loop tests and the closed-loop PID trials. It also means documenting the original PID settings of the controller, and all PID setting values attempted during the tuning process (linked to their respective trend graphs, so it will be easy to tell which sets of PID constants produced which process responses). If there are any instrument configuration settings (e.g. damping time values in process transmitters) changed during the tuning exercise, both the original values and all your changes need to be documented as well.

As a final word, I would like to cast a critical vote against auto-tuning controllers. With all due respect for the engineers who work hard to make controllers "intelligent" enough to adjust their own PID settings, there is no controller in the world that can account for all the factors described in this "Before you tune . . ." section. Feel free to use the automatic tuning feature of a controller, but only *after* you have ensured all instrument and process problems are corrected, and *after* you have confirmed the tuning goal of the controller matches the behavioral goal of the control loop as defined by the operators (e.g. fast response versus minimum overshoot, etc.). Some people in the automation business are over-confident with regard to the capabilities of auto-tuning controllers. We would all do well to recognize this feature as a *tool*, not unlike any other tool in that it has specific purposes and limitations, and is certainly no panacea for a lack of understanding how PID control works.

27.3 Quantitative PID tuning procedures

A *quantitative* PID tuning procedure is a step-by-step approach leading directly to a set of numerical values to be used in a PID controller. These procedures may be split into two categories: *open loop* and *closed loop*. An "open loop" tuning procedure is implemented with the controller in manual mode: introducing a step-change to the controller output and then mathematically analyzing the results of the process variable response to calculate appropriate PID settings for the controller to use when placed into automatic mode. A "closed loop" tuning procedure is implemented with the controller in automatic mode: adjusting tuning parameters to achieve an easily-defined result, then using those PID parameter values and information from a graph of the process variable over time to calculate new PID parameters.

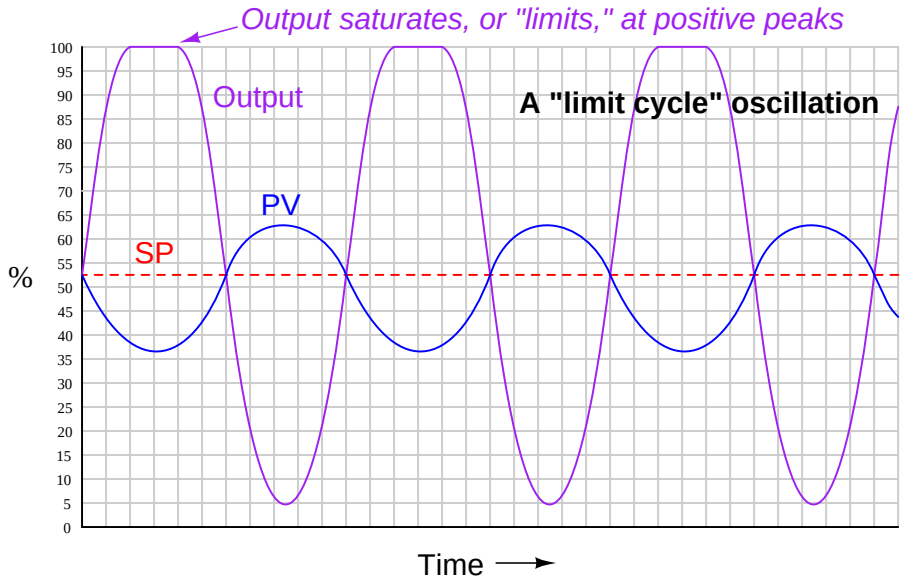
Quantitative PID tuning got its start with a paper published in the November 1942 *Transactions of the American Society of Mechanical Engineers* written by two engineers named Ziegler and Nichols. "Optimum Settings For Automatic Controllers" is a seminal paper, and deserves to be read by any serious student of process control theory. That Ziegler's and Nichols' recommendations for PID controller settings may still be found in modern references more than 60 years after publication is a testament to its impact in the field of industrial control. Although dated in its terminology and references to pneumatic controller technology (some controllers mentioned as not even having adjustable proportional response, and others as having only discrete degrees of reset adjustment rather than continuously variable settings!), the PID algorithm described by its authors and the effects of P, I, and D adjustments on process control behavior are as valid today as they were then.

This section is devoted to a discussion of quantitative PID tuning procedures in general, and the "Ziegler-Nichols" methods in specific. It is the opinion of this author that the Ziegler-Nichols tuning methods are useful primarily as historical references, and indeed suffer from serious practical impediments. Perhaps the most serious problem I have with the Ziegler-Nichols methods (and in fact any algorithmic procedure for PID tuning) is that it tends to absolve the practitioner of responsibility for understanding the process they intend to tune. Any time you provide people with step-by-step instructions to perform complex tasks, there will be a great many readers of those instructions tempted to mindlessly follow the instructions, even to their doom. PID tuning is one of these "complex tasks," and there is significant likelihood for a person to do more harm than good if all they do is implement a step-by-step approach rather than understand what they are doing, why they are doing it, and what it means if the results do not meet with satisfaction. Please bear this in mind as you study any PID tuning procedure, Ziegler-Nichols or otherwise.

27.3.1 Ziegler-Nichols closed-loop ("Ultimate Gain")

Closed-loop refers to the operation of a control system with the controlling device in "automatic" mode, where the flow of the information from sensing element to transmitter to controller to control element to process and back to sensor represents a continuous ("closed") feedback loop. If the total amount of signal amplification provided by the instruments is too much, the feedback loop will self-oscillate. While oscillation is almost always considered undesirable in a control system, it may be used as an exploratory test of process dynamics if the controller acts purely on proportional action (no integral or derivative action): providing data useful for calculating effective PID controller settings. Thus, a "closed-loop" PID tuning procedure entails disabling any integral or derivative actions in the controller, then raising the gain value of the controller until self-sustaining oscillations of constant amplitude are achieved in the process variable. The amount of controller gain necessary to sustain process oscillations of consistent amplitude is called the *ultimate sensitivity* (S_u) of the process, while the time (period) between successive oscillation peaks is called the *ultimate period* (P_u) of the process.

When performing such a test on a process loop, it is important to ensure the oscillation peaks do not reach the limits of the instrumentation, either measurement or final control element. In other words, in order for the oscillation to accurately reveal the process characteristics of ultimate sensitivity and ultimate period, the oscillations must be naturally limited and not artificially limited by either the transmitter or the control valve saturating. Oscillations characterized by either the transmitter or the final control element reaching their range limits should be avoided in order to obtain the best closed-loop oscillatory test results. An illustration is shown here as a model of what to avoid:



If the controller in question is proportional-only (i.e. capable of providing no integral or derivative control actions), Ziegler and Nichols' recommendation is to set the controller gain[19] to one-half the value of the ultimate sensitivity, which I will call ultimate *gain* (K_u) from now on:

$$K_p = 0.5K_u$$

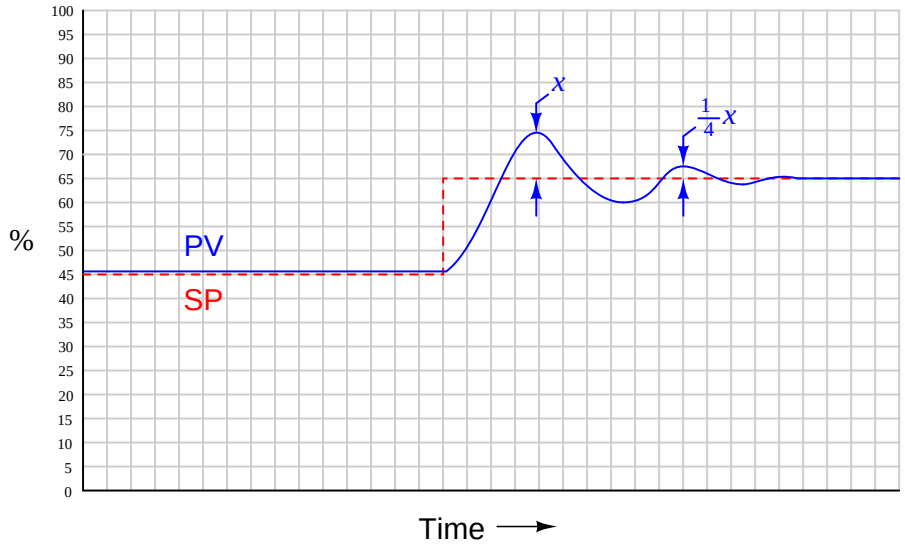
Where,

K_p = Controller gain

K_u = "Ultimate" gain determined by increasing controller gain until self-sustaining oscillations are achieved

Generally, a controller gain of one-half the experimentally determined "ultimate" gain results in reasonably quick response to setpoint and process load changes. Oscillations of the process variable following such setpoint and load changes typically damp with each successive wave peak being approximately one-quarter the amplitude of the one preceding. This is known as *quarter-wave damping*. While certainly not ideal, it is a compromise between fast response and stability.

The following process trend shows what "quarter-wave damping" looks like, with the process variable (PV) oscillating a bit following a step-change in setpoint (SP):



Ziegler and Nichols were careful to qualify quarter-wave damping as less than optimal for some applications. In their own words (page 761):

> "The statement that a sensitivity setting of one half the ultimate with attendant 25 per cent amplitude ratio gives optimum control must be modified in some cases. For example, the actual level maintained by a liquid-level controller might not be nearly as important as the effect of sudden valve movements on further portions of the process. In

[19] Note that this is truly the *gain* of the controller, not the *proportional band*. If you were to enter a proportional band value one-half the proportional band value necessary to sustain oscillations, the controller would (obviously) oscillate completely out of control!

> this case the sensitivity should be lowered to reduce the amplitude ratio even though the offset is increased by so doing. On the other hand, a pressure-control application giving oscillations with very short period could be set to give an 80 or 90 per cent amplitude ratio. Due to the short period, a disturbance would die out in reasonable time, even though there were quite a few oscillations. The offset would be reduced somewhat though it should be kept in mind that it can never be reduced to less than one half of the amount given at our previously defined optimum sensitivity of one half the ultimate."

Some would argue (myself included) that quarter-wave damping exercises the control valve needlessly, causing undue stem packing wear over time and consuming large quantities of compressed air to operate. Given the fact that all modern process controllers have integral (reset) capability, unlike the simple pneumatic controllers of Ziegler and Nichols' day, there is really no need to tolerate prolonged offset (failure of the process variable to exactly equalize with setpoint over time) as a necessary cost of avoiding valve oscillation.

If the controller in question has integral (reset) action in addition to proportional, Ziegler and Nichols' recommendation is to set the controller gain to slightly less than one-half the value of the ultimate sensitivity, and to set the integral time constant[20] to a value slightly less than the ultimate period:

$$K_p = 0.45K_u$$

$$\tau_i = \frac{P_u}{1.2}$$

Where,

K_p = Controller gain

K_u = "Ultimate" gain determined by increasing controller gain until self-sustaining oscillations are achieved

τ_i = Controller integral setting (minutes per repeat)

P_u = "Ultimate" period of self-sustaining oscillations (minutes)

[20] Either minutes per repeat or seconds per repeat. If the controller's integral rate is expressed in units of repeats per minute (or second), the formula would be $K_i = \frac{1.2}{P_u}$.

If the controller in question has all three control actions present (full PID), Ziegler and Nichols' recommendation is to set the controller tuning constants as follows:

$$K_p = 0.6K_u$$

$$\tau_i = \frac{P_u}{2}$$

$$\tau_d = \frac{P_u}{8}$$

Where,

K_p = Controller gain

K_u = "Ultimate" gain determined by increasing controller gain until self-sustaining oscillations are achieved

τ_i = Controller integral setting (minutes per repeat)

τ_d = Controller derivative setting (minutes)

P_u = "Ultimate" period of self-sustaining oscillations (minutes)

An important caveat with any tuning procedure based on ultimate gain is the potential to cause trouble in a process while experimentally determining the ultimate gain. Recall that "ultimate" gain is the amount of controller gain (proportional action) resulting in self-sustaining oscillations of constant amplitude. In order to precisely determine this gain setting, one must spend some time provoking the process with sudden setpoint changes (to induce oscillation) and experimenting with greater and greater gain settings until constant oscillation amplitude is achieved. Any more gain than the "ultimate" value, of course, leads to ever-*growing* oscillations which may be brought under control only by decreasing controller gain or switching to manual mode (thereby stopping all feedback in the system). The problem with this is, one never knows for certain when ultimate gain is achieved until this critical value has been exceeded, as evidenced by ever-growing oscillations. In other words, *the system must be brought to the brink of total instability in order to determine its ultimate gain value.* Not only is this time-consuming to achieve – especially in systems where the natural period of oscillation is long, as is the case with many temperature and composition control applications – but potentially hazardous to equipment and certainly detrimental to process quality[21]. In fact, one might argue that any process tolerant of such abuse probably doesn't need to be well-tuned!

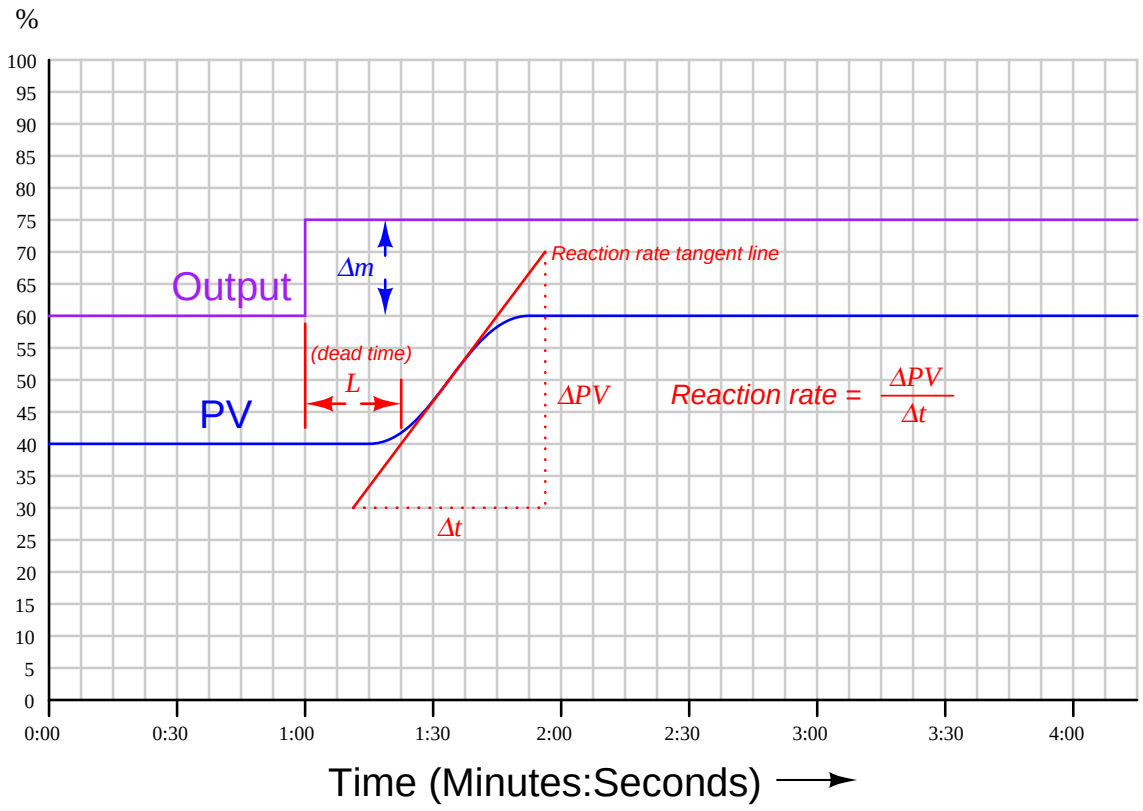
Despite its practical limitations, the rules given by Ziegler and Nichols do shed light on the relationship between realistic P, I, and D tuning parameters and the operational characteristics of the process. Controller gain should be some fraction of the gain necessary for the process to self-oscillate. Integral time constant should be proportional to the process time constant; i.e. the "slower" the process is to respond, the "slower" (less aggressive) the controller's integral response should be. Derivative time constant should likewise be proportional to the process time constant, although this has the opposite meaning from the perspective of aggressiveness: a "slow" process deserves a long derivative time constant; i.e. *more aggressive* derivative action.

[21] For a graphic demonstration of just how impractical this may be, just try telling a lead operations manager or a unit supervisor in a chemical processing facility you wish to over-tune the temperature controller in the main reaction furnace or the pressure controller in one of the larger distillation columns until it nearly oscillates out of control, and that doing so may necessitate hours of unstable operation before you find the perfect gain setting. Consider yourself lucky if you are not escorted to the control room door by security personnel following this declaration of intent.

27.3.2 Ziegler-Nichols open-loop

In contrast to the first tuning technique presented by Ziegler and Nichols in their landmark 1942 paper where the process was made to oscillate using proportional-only automatic control and the parameters of that oscillation served to define PID tuning parameters, their second tuning technique did not even rely on the presence of a controller. Instead, this second technique consisted of making a manual "step-change" of the control element (valve) and analyzing the resulting effect on the process variable, much the same way as described in the Process Characterization section of this chapter (section 27.1, beginning on page 1506).

After making the step-change in output signal with the controller in manual mode, the process variable trend is closely analyzed for two salient features: the *dead time* and the *reaction rate*. Dead time (L)[22] is the amount of time delay between the output step-change and the first indication of process variable change. Reaction rate is the maximum rate at which the process variable changes following the output step-change (the maximum time-derivative of the process variable):



Dead time and reaction rate are responses common to self-regulating and integrating processes alike. Whether or not the process variable ends up stabilizing at some new value, its rate of rise will

[22] Unfortunately, Ziegler and Nichols chose to refer to dead time by the word *lag* in their paper. In modern technical parlance, "lag" refers to a first-order inverse-exponential function, which is fundamentally different from dead time.

reach some maximum value following the output step-change, and this will be the reaction rate of the process[23]. The unit of measurement for reaction rate is *percent per minute*:

$$R = \frac{\Delta \text{PV}}{\Delta t} = \frac{[\text{Percent rise}]}{[\text{Minutes run}]}$$

While dead time in a process tends to be constant regardless of the output step-change magnitude, reaction rate tends to vary directly with the magnitude of the output step-change. For example, an output step-change of 10% will generally cause the PV to rise at a rate twice as steep compared to the effects of a 5% output step-change. In order to ensure our predictive calculations capture only what is inherent to the process and not our own arbitrary open-loop tuning actions, we must include the output step-change magnitude (Δm) in those calculations as well[24].

If the controller in question is proportional-only (i.e. capable of providing no integral or derivative control actions), Ziegler and Nichols' recommendation is to set the controller gain as follows:

$$K_p = \frac{\Delta m}{RL}$$

Where,
K_p = Controller gain (unitless)
Δm = Output step-change magnitude made while testing in open-loop (manual) mode (percent)
R = Process reaction rate = $\frac{\Delta \text{PV}}{\Delta t}$ (percent per minute)
L = Process dead time (minutes)

If the controller in question has integral (reset) action in addition to proportional, Ziegler and Nichols' recommendation is to set the controller gain to 90% of the proportional-only value, and to set the integral time constant to a value just over three times the dead time value:

$$K_p = 0.9\frac{\Delta m}{RL}$$

$$\tau_i = 3.33L$$

Where,
K_p = Controller gain (unitless)
Δm = Output step-change magnitude made while testing in open-loop (manual) mode (percent)
R = Process reaction rate = $\frac{\Delta \text{PV}}{\Delta t}$ (percent per minute)
L = Process dead time (minutes)
τ_i = Controller integral setting (minutes per repeat)

[23] Right away, we see a weakness in the Ziegler-Nichols open-loop method: it makes absolutely no distinction between self-regulating and integrating process types. We know this is problematic from the analysis of each process type in sections 27.1.1 and 27.1.2, beginning on page 1507.

[24] Ziegler and Nichols' approach was to define a normalized reaction rate called the *unit reaction rate*, equal in value to $\frac{R}{\Delta m}$. I opt to explicitly include Δm in all the tuning parameter equations in order to avoid the possibility of confusing reaction rate with unit reaction rate.

If the controller has full PID capability, Ziegler and Nichols' recommendation is to set the controller gain to 120% of the proportional-only value, to set the integral time constant to twice the dead time value, and to set the derivative time constant to one-half the dead time value:

$$K_p = 1.2\frac{\Delta m}{RL}$$

$$\tau_i = 2L$$

$$\tau_d = 0.5L$$

Where,

K_p = Controller gain (unitless)

Δm = Output step-change magnitude made while testing in open-loop (manual) mode (percent)

R = Process reaction rate = $\frac{\Delta \text{PV}}{\Delta t}$ (percent per minute)

L = Process dead time (minutes)

τ_i = Controller integral setting (minutes per repeat)

τ_d = Controller derivative setting (minutes)

27.4 Heuristic PID tuning procedures

In contrast to quantitative tuning procedures where definite numerical values for P, I, and D controller settings are obtained through data collection and analysis, a *heuristic* tuning procedure is one where general rules are followed to obtain approximate or qualitative results. The majority of PID loops in the world have been tuned with such methods, for better or for worse. My goal in this section is to optimize the effectiveness of such tuning methods.

When I was first educated on the subject of PID tuning, the only procedure presented for loop tuning was a crude step-by-step procedure:

1. Configure the controller for proportional action only (integral and derivative control actions set to minimum effect), setting the gain near or at 1.
2. Increase controller gain until self-sustaining oscillations are achieved, "bumping" the setpoint value up or down as necessary to provoke oscillations.
3. When the ultimate gain is determined, set the controller gain for half that value.
4. Repeat steps 2 and 3, this time adjusting integral action instead of proportional.
5. Repeat steps 2 and 3, this time adjusting derivative action instead of proportional.

The first three steps of this procedure are identical to the steps recommended by Ziegler and Nichols for closed-loop tuning. The last two steps are someone else's contribution. While this particular procedure may be peculiar to my own personal path of education, it showcases the general spirit of most heuristic tuning methods: adjust each controller parameter to be more and more aggressive until some compromise is reached between fast action and instability.

Much improvement may be made to any "trial-and-terror" PID tuning procedure if one is aware of the process characteristics and recognizes the applicability of P, I, and D actions to particular characteristics. Simply experimenting with P, I, and D parameter values is tedious at best and dangerous at worst if one has no understanding of what each type of control action if useful for, and the limitations of each control action.

27.4.1 Features of P, I, and D actions

Purpose of each action

- **Proportional action** is the "universal" control action, capable of providing at least marginal control quality for any process.
- **Integral action** is useful for eliminating offset caused by load variations and process self-regulation.
- **Derivative action** is useful for canceling lags.

Limitations of each action

- **Proportional action** will cause oscillations if sufficiently aggressive, in the presence of lags and/or dead time. The more lags (higher-order), the worse the problem. It also directly amplifies process noise.
- **Integral action** will cause oscillation if sufficiently aggressive, in the presence of lags and/or dead time. Any amount of integral action will guarantee setpoint overshoot in purely integrating processes.
- **Derivative action** dramatically amplifies process noise, and will cause oscillations in fast-acting processes.

Special applicability of each action

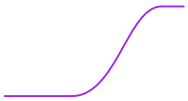
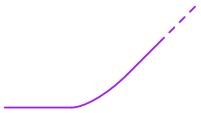
- **Proportional action** works exceptionally well when aggressively applied to self-regulating processes dominated by first-order lag, and to purely integrating processes.
- **Integral action** works exceptionally well when aggressively applied to fast-acting, self-regulating processes. Has the unique ability to ignore process noise.
- **Derivative action** works exceptionally well to speed up the response of processes dominated by large lag times, and to help stabilize runaway processes.

Gain and phase shift of each action

- **Proportional action** acts on the *present*, adding no phase shift to a sinusoidal signal. Its gain is constant for any signal frequency.
- **Integral action** acts on the *past*, adding a -90^o phase shift to a sinusoidal signal. Its gain decreases with increasing frequency.
- **Derivative action** acts on the *future*, adding a $+90^o$ phase shift to a sinusoidal signal. Its gain increases with increasing frequency.

27.4.2 Tuning recommendations based on process dynamics

Knowing which control actions to focus on first is a matter of characterizing the process (identifying whether it is self-regulating, integrating, runaway, noisy, has lag or dead time, or any combination of these traits based on an open-loop response test[25]) and then selecting the best actions to fit those characteristics. The following table shows some general recommendations for fitting PID tuning to different process characteristics

	Pure self-regulating	*May be controlled with aggressive integral action, and perhaps with a bit of proportional action. Use absolutely no derivative action!*
	Self-reg w/ pure 1st order lag	*Responds well to aggressive proportional action, with integral action needed only for recovery from load changes.*
	Self-reg w/ multiple lags	*Proportional action needed for quick response to setpoint changes, integral action needed for recovery from load changes, and derivative needed to prevent overshoot. Proportional and integral actions are limited by tendency to oscillate.*
	Integrating w/ lag(s)	*Proportional action should be aggressive as possible without generating oscillations. Integral action needed only for recovery from load changes.*
	Pure integrating	*Responds well to aggressive proportional action, with integral action needed only for recovery from load changes.*

General rules:

- Use no derivative action if the process signal is "noisy"
- Use proportional action sparingly if the process signal is "noisy"
- The slower the time lag(s), the less integral action to use
- The higher-order the time lag(s), the less proportional action (gain) to use
- Self-regulating processes *need* integral action
- Integrating processes *need* proportional action
- Dead time requires a reduction of all PID constants below what would normally work

[25] Recall that an open-loop response test consists of placing the loop controller in manual mode, introducing a step-change to the controller output (manipulated variable), and analyzing the time-domain response of the process variable as it reacts to that perturbation.

Once you have determined the basic character of the process, and understand from that characterization what the needs of the process will be regarding P, I, and/or D control actions, you may "experiment" with different tuning values of P, I, and D until you find a combination yielding robust control.

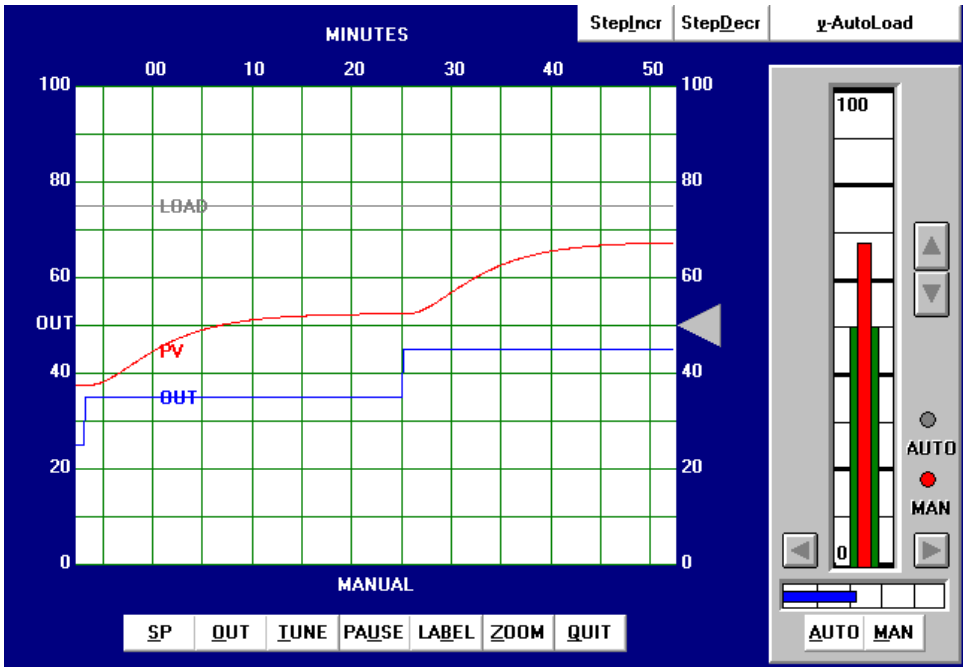
27.5 Tuning techniques compared

In this section I will show screenshots from a process loop simulation program illustrating the effectiveness of Ziegler-Nichols open-loop ("Reaction Rate") and closed-loop ("Ultimate") PID tuning methods, and then contrast them against the results of my own heuristic tuning. As you will see in each case, the results obtained by either Ziegler-Nichols method were quite poor. This is not necessarily an indictment of Ziegler's and Nichols' recommendations as much as it is a demonstration of the power of understanding. Ziegler and Nichols presented a simple step-by-step procedure for obtaining *approximate* PID tuning constant values based on closed-loop and open-loop process responses, which could be applied by anyone regardless of their level of understanding PID control theory. If I were tasked with drafting a procedure to instruct anyone to quantitatively determine PID constant values without an understanding of process dynamics or process control theory, I doubt my effort would be an improvement. Ultimately, robust PID control is attainable only at the hands of someone who understands how PID works, what each mode does (and why), and is able to distinguish between intrinsic process characteristics and instrument limitations. The purpose of this section is to clearly demonstrate the limitations of ignorantly-followed procedures, and contrast this "mindless" approach against the results of simple experimentation directed by qualitative understanding.

Each of the examples illustrated in this section were simulations run on a computer program called *PC-ControLab* developed by Wade Associates, Inc. Although these are simulated processes, in general I have found similar results using both Ziegler-Nichols and heuristic tuning methods on real processes. The control criteria I used for heuristic tuning were fast response to setpoint changes, with minimal overshoot or oscillation.

27.5.1 Tuning a "generic" process

The first process tuned in simulation was a "generic" process, unspecific in its nature or application. Performing an open-loop test (two 10% output step-changes made in manual mode, both increasing) on this process resulted in the following behavior:



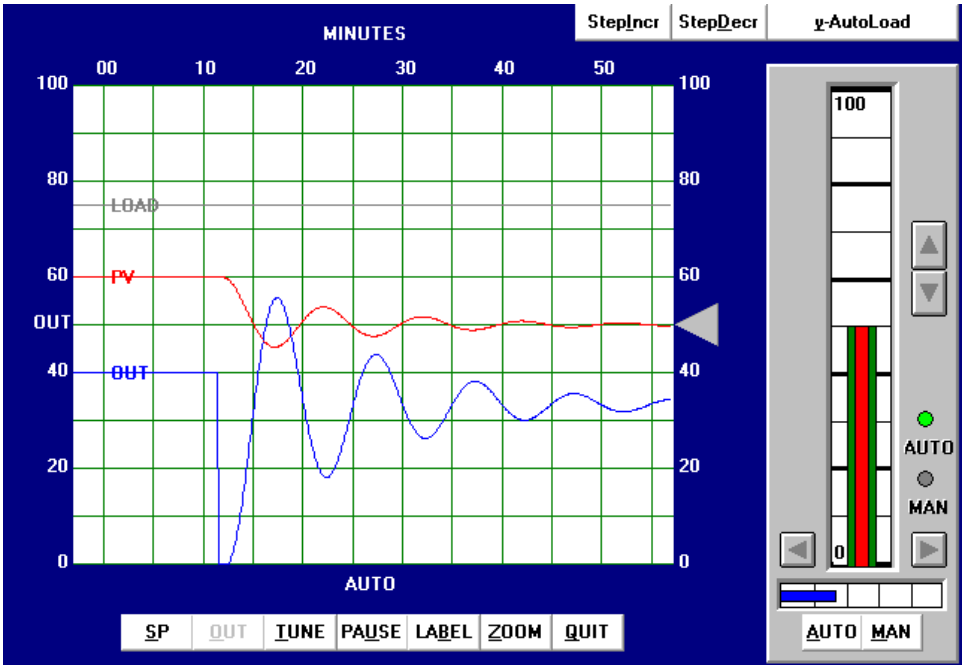
From the trend, we can see that this process is self-regulating, with multiple lags and some dead time. The reaction rate (R) is 20% over 15 minutes, or 1.333 percent per minute. Dead time (L) appears to be approximately 2 minutes. Following the Ziegler-Nichols recommendations for PID tuning based on these process characteristics (also including the 10% step-change magnitude Δm):

$$K_p = 1.2\frac{\Delta m}{RL} = 1.2\frac{10\%}{\frac{20\%}{15 \text{ min}}2 \text{ min}} = 4.5$$

$$\tau_i = 2L = (2)(2 \text{ min}) = 4 \text{ min}$$

$$\tau_d = 0.5L = (0.5)(2 \text{ min}) = 1 \text{ min}$$

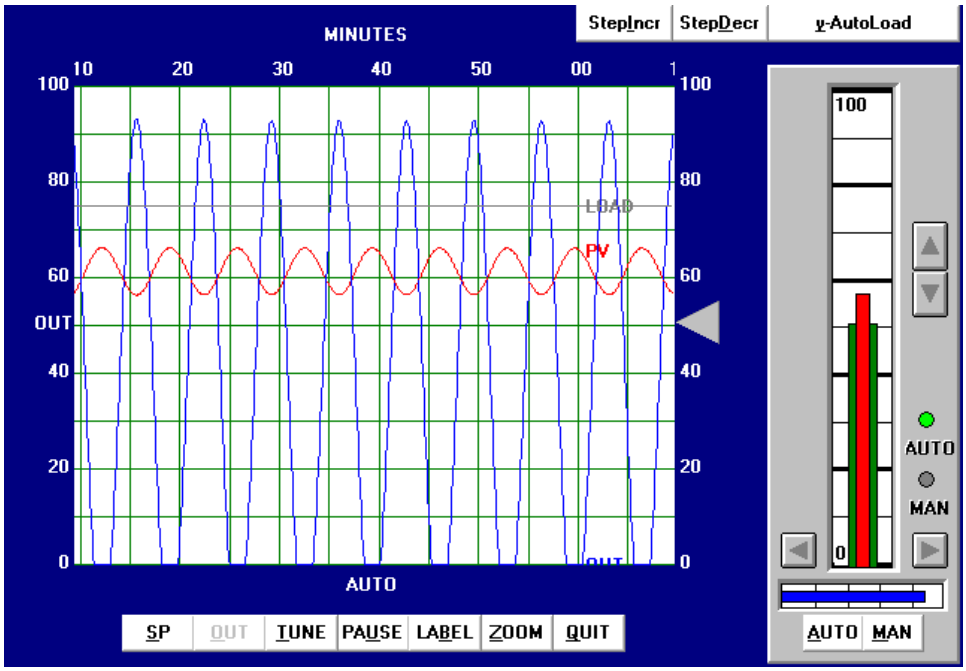
Applying the PID values of 4.5 (gain), 4 minutes per repeat (integral), and 1 minute (derivative) gave the following result in automatic mode:



The process oscillations follow a 10% setpoint change (from 60% to 50%), and take almost an hour to settle. Clearly, this is less-than-robust behavior. The controller is much too aggressive, which is why the process oscillates so much following the setpoint step-change.

Looking closely at the PV and OUT waveforms, we see their phase relationship is nearly 180^o, consistent with what we would expect for a reverse-acting controller with strong proportional action. If integral or derivative action were primarily responsible for the oscillation, we would see the influence of an additional 90^o phase shift characteristic to those actions (integral naturally produces a -90^o phase shift, while derivative naturally produces a $+90^o$ phase shift, for any sinusoidal function). Thus, if the oscillations were primarily the result of excessive integral action, we would expect the OUT wave to lead the PV wave by nearly 90^o (180^o from reverse action $-$ 90^o from integral action $= 90^o$). If excessive derivative action were primarily responsible for the oscillations, we would expect the OUT wave to lag the PV wave by nearly 90^o (180^o from reverse action + 90^o from derivative action $= 270^o = -90^o$). Since we see a phase shift of the oscillations between OUT and PV very close to the 180^o predicted by proportional action, we can be sure this controller's response is dominated by proportional action, which would be a good place to start "taming" this over-exuberant controller if we were inclined to modify the Ziegler-Nichols tuning recommendations.

Next, the closed-loop, or "Ultimate" tuning method of Ziegler and Nichols was applied to this process. Eliminating both integral and derivative control actions from the controller, and experimenting with different gain (proportional) values until self-sustaining oscillations of consistent amplitude[26] were obtained, gave a gain value of 11:



From the trend, we can see that the ultimate period (P_u) is approximately 7 minutes in length. Following the Ziegler-Nichols recommendations for PID tuning based on these process characteristics:

$$K_p = 0.6K_u = (0.6)(11) = 6.6$$

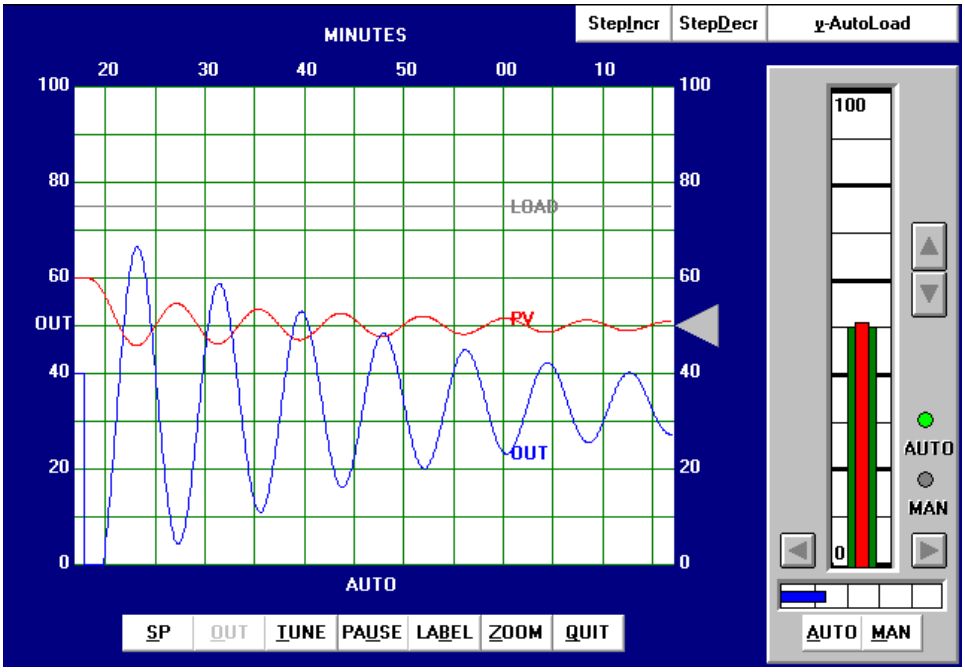
$$\tau_i = \frac{P_u}{2} = \frac{7 \text{ min}}{2} = 3.5 \text{ min}$$

$$\tau_d = \frac{P_u}{8} = \frac{7 \text{ min}}{8} = 0.875 \text{ min}$$

It should be immediately apparent that these tuning parameters will yield poor control. While the integral and derivative values are close to those predicted by the open-loop (Reaction Rate) method, the gain value calculated here is even larger than what was calculated before. Since we know proportional action was excessive in the last tuning attempt, and this one recommends an even higher gain value, we can expect our next trial to oscillate even worse.

[26]The astute observer will note the presence of some limiting (saturation) in the output waveform, as it attempts to go below zero percent. Normally, this is unacceptable while determining the ultimate gain of a process, but here it was virtually impossible to make the process oscillate at consistent amplitude without saturating on the output signal. The gain of this process falls off quite a bit at the ultimate frequency, such that a high controller gain is necessary to sustain oscillations, causing the output waveform to have a large amplitude.

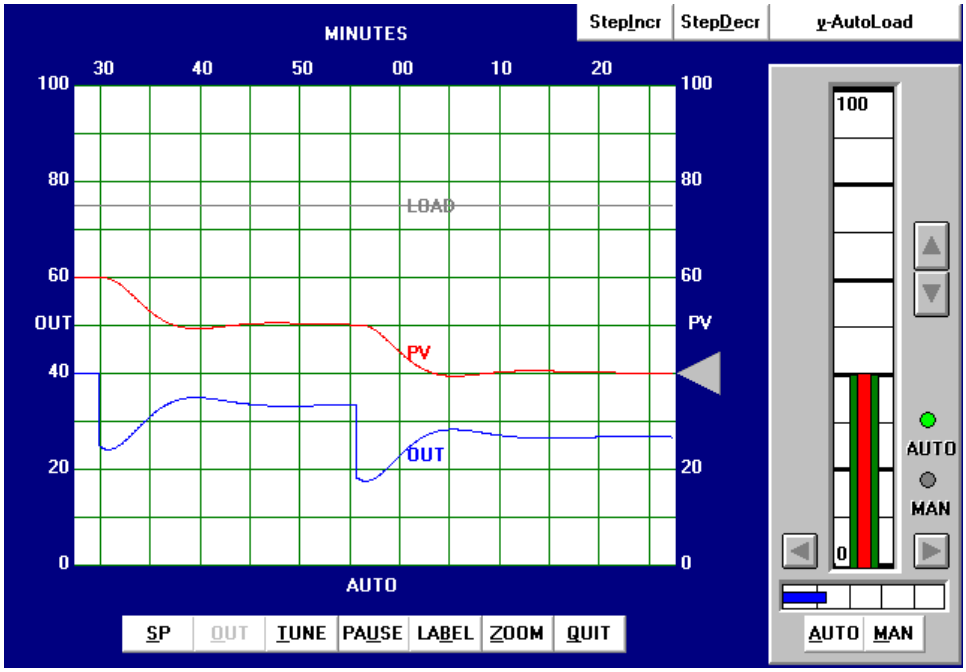
Applying the PID values of 6.6 (gain), 3.5 minutes per repeat (integral), and 0.875 minute (derivative) gave the following result in automatic mode:



The process oscillations follow a 10% setpoint change (from 60% to 50%), and are nowhere near settled after an hour's worth of time. Once again we see the nearly 180^o phase shift between the OUT and the PV waves, indicating proportional reverse-action controller response. As expected, this is far too much gain for robust control!

From the initial open-loop (manual output step-change) test, we could see this process contains multiple lags in addition to about 2 minutes of dead time. Both of these factors tend to limit the amount of gain we can use in the controller before the process oscillates. Both Ziegler-Nichols tuning attempts confirmed this fact, which led me to try much lower gain values in my initial heuristic tests. Given the self-regulating nature of the process, I knew the controller needed integral action, but once again the aggressiveness of this action would be necessarily limited by the lag and dead times. Derivative action, however, would prove to be useful in its ability to help "cancel" lags, so I suspected my tuning would consist of relatively tame proportional and integral values, with a relatively aggressive derivative value.

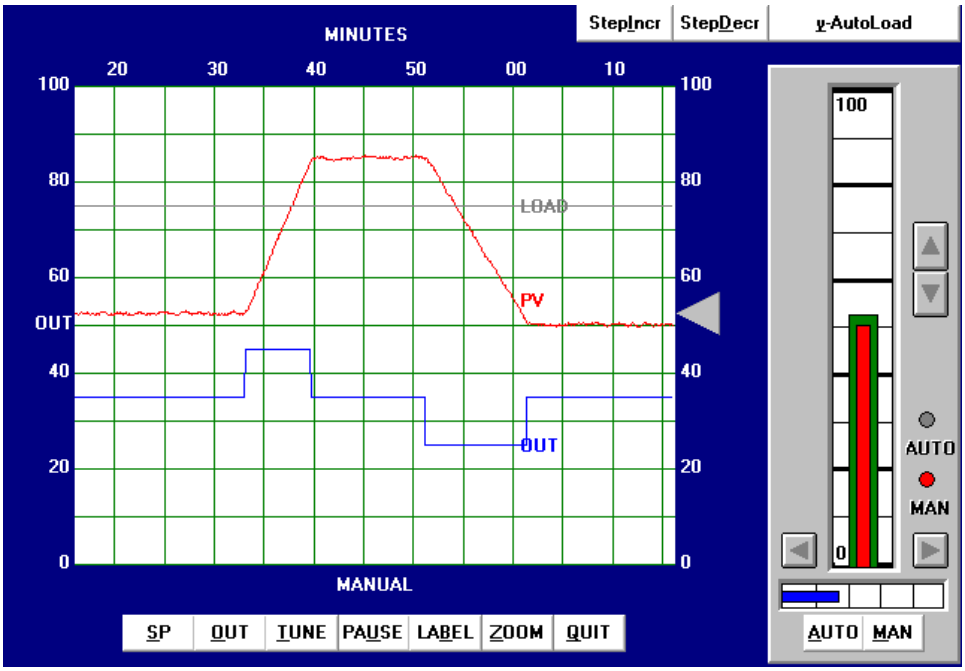
After some experimenting, the values I arrived at were 1.5 (gain), 10 minutes (integral), and 5 minutes (derivative). These tuning values represent a proportional action only one-third as aggressive as the least-aggressive Ziegler-Nichols recommendation, an integral action less than half as aggressive as the Ziegler-Nichols recommendations, and a derivative action *five times* more aggressive than the most aggressive Ziegler-Nichols recommendation. The results of these tuning values in automatic mode are shown here:



With this PID tuning, the process responded well to not just one, but *two* 10% setpoint step-changes within the span of an hour.

27.5.2 Tuning a liquid level process

The next simulated process I attempted to tune was a liquid level-control process. Performing an open-loop test (one 10% increasing output step-change, followed by a 10% decreasing output step-change, both made in manual mode) on this process resulted in the following behavior:



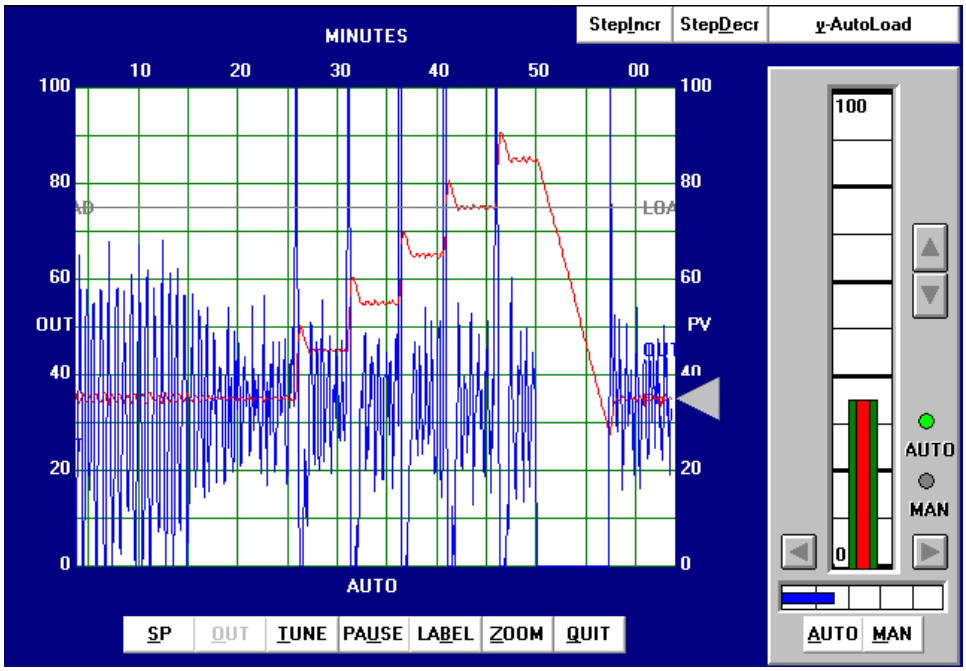
From the trend, the process appears to be purely integrating, as though the control valve were throttling the flow of liquid into a vessel with a constant out-flow. The reaction rate (R) on the first step-change is 50% over 10 minutes, or 5 percent per minute. Dead time (L) appears virtually nonexistent, estimated to be 0.1 minutes simply for the sake of having a dead-time value to use in the Ziegler-Nichols equations. Following the Ziegler-Nichols recommendations for PID tuning based on these process characteristics (also including the 10% step-change magnitude Δm):

$$K_p = 1.2\frac{\Delta m}{RL} = 1.2\frac{10\%}{\frac{50\%}{10\text{ min}}0.1\text{ min}} = 24$$

$$\tau_i = 2L = (2)(0.1\text{ min}) = 0.2\text{ min}$$

$$\tau_d = 0.5L = (0.5)(0.1\text{ min}) = 0.05\text{ min}$$

Applying the PID values of 24 (gain), 0.2 minutes per repeat (integral), and 0.05 minutes (derivative) gave the following result in automatic mode:



The process variable certainly responds rapidly to the five increasing setpoint changes and also to the one large decreasing setpoint change, but the valve action is hopelessly chaotic. Not only would this "jittery" valve motion prematurely wear out the stem packing, but it would also result in vast over-consumption of compressed air to continually stroke the valve from one extreme to the other. Furthermore, we see evidence of "overshoot" at every setpoint change, most likely from excessive integral action.

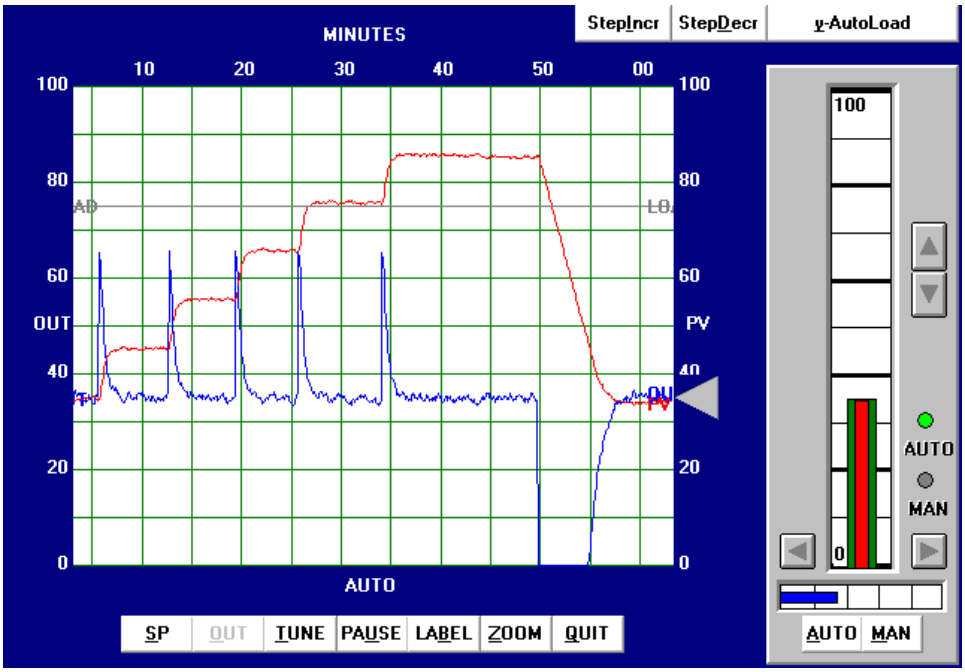
We can see from the valve's wild behavior even during periods when the process variable is holding at setpoint that the problem is not a loop oscillation, but rather the effects of process noise on the controller. The extremely high gain value of 24 is amplifying PV noise by that factor, and reproducing it on the output signal.

Next, I attempted to perform a closed-loop "Ultimate" gain test on this process, but I was not successful. Even the controller's maximum possible gain value would not generate oscillations, due to the extremely crisp response of the process (minimal lag and dead times) and its integrating nature (constant phase shift of -90^o).

From the initial open-loop (manual output step-change) test, we could see this process was purely integrating. This told me it could be controlled primarily by proportional action, with very little integral action required, and no derivative action whatsoever. The presence of some process noise is the only factor limiting the aggressiveness of proportional action. With this in mind, I experimented with increasingly aggressive gain values until I reached a point where I felt the output signal noise

was at a maximum acceptable limit for the control valve. Then, I experimented with integral action to ensure reasonable elimination of offset.

After some experimenting, the values I arrived at were 3 (gain), 10 minutes (integral), and 0 minutes (derivative). These tuning values represent a proportional action only one-eighth as aggressive as the Ziegler-Nichols recommendation, and an integral action *fifty times* less aggressive than the Ziegler-Nichols recommendation. The results of these tuning values in automatic mode are shown here:



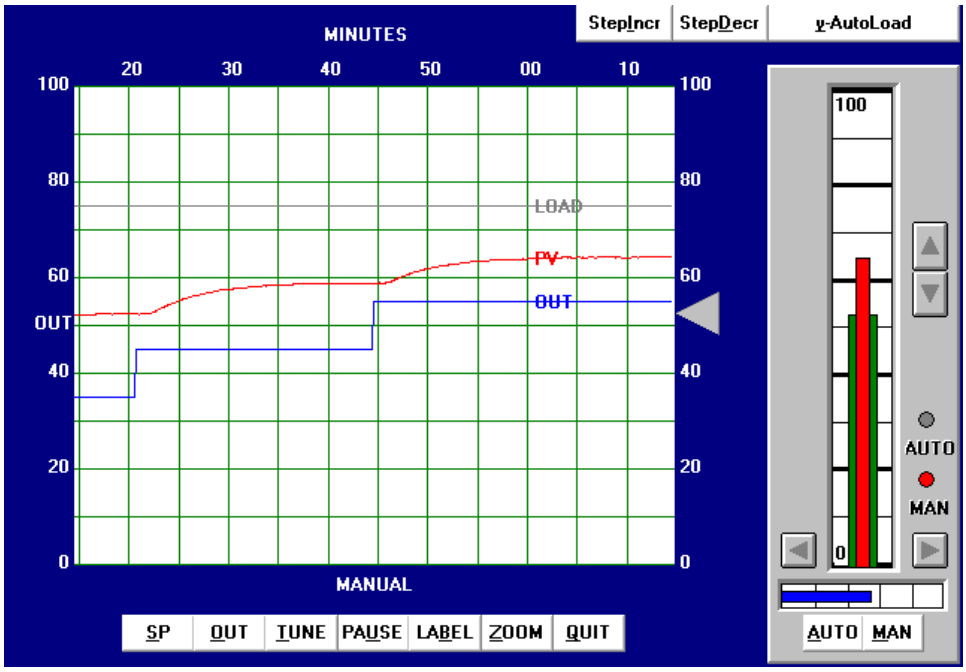
You can see on this trend five 10% increasing setpoint value changes, with crisp response every time, followed by a single 50% decreasing setpoint step-change. In all cases, the process response clearly meets the criteria of rapid attainment of new setpoint values and no overshoot or oscillation.

If it was decided that the noise in the output signal was too much for the valve to handle over time, we would have the option of further reducing the gain value and (possibly) compensating for slow offset recovery with more aggressive integral action. We could also attempt the insertion of a damping constant into either the level transmitter or the controller itself, so long as this added lag did not cause oscillation problems in the loop[27]. The best solution would be to find a way to isolate the level transmitter from noise, so that the process variable signal was much "quieter." Whether or not this is possible depends on the process and on the particular transmitter used.

[27] We would have to be *very* careful with the addition of damping, since the oscillations could create may not appear on the trend. Remember that the insertion of damping (low-pass filtering) in the PV signal is essentially an act of "lying" to the controller: telling the controller something that differs from the real, measured signal. If our PV trend shows us this damped signal and not the "raw" signal from the transmitter, it is possible for the process to oscillate and the PV trend to be deceptively stable!

27.5.3 Tuning a temperature process

This next simulated process is a temperature control process. Performing an open-loop test (two 10% increasing output step-changes, both made in manual mode) on this process resulted in the following behavior:



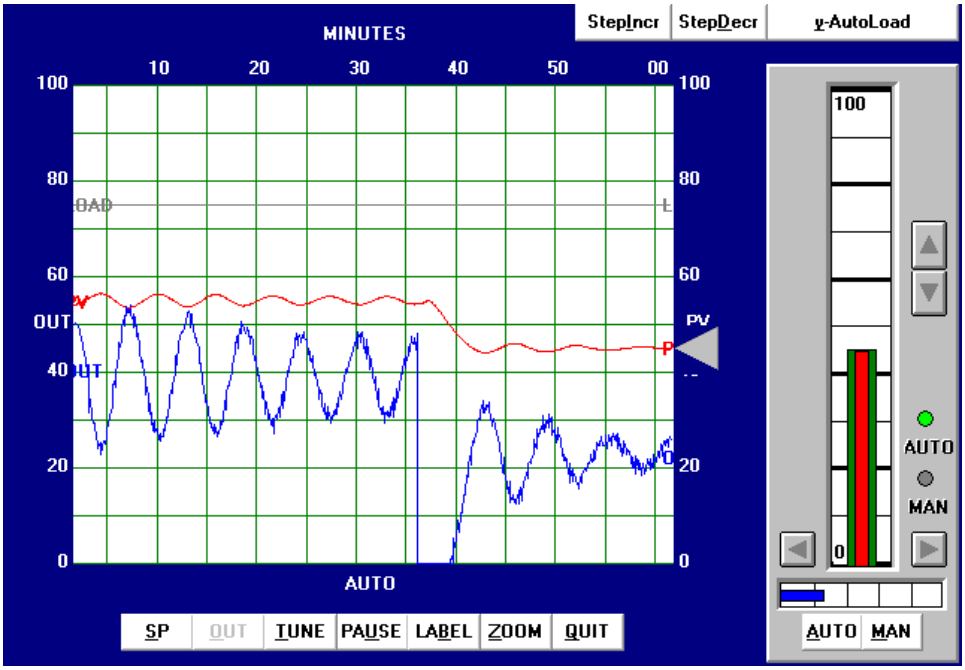
From the trend, the process appears to be self-regulating with a slow time constant (lag) and a substantial dead time. The reaction rate (R) on the first step-change is 30% over 30 minutes, or 1 percent per minute. Dead time (L) looks to be approximately 1.25 minutes. Following the Ziegler-Nichols recommendations for PID tuning based on these process characteristics (also including the 10% step-change magnitude Δm):

$$K_p = 1.2\frac{\Delta m}{RL} = 1.2\frac{10\%}{\frac{30\%}{30 \text{ min}}1.25 \text{ min}} = 9.6$$

$$\tau_i = 2L = (2)(1.25 \text{ min}) = 2.5 \text{ min}$$

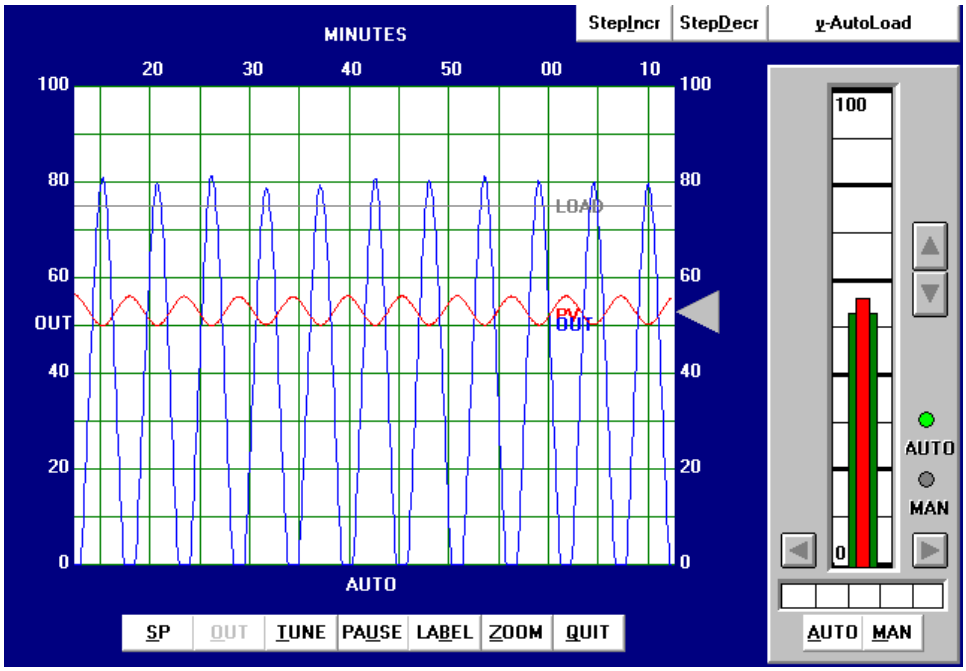
$$\tau_d = 0.5L = (0.5)(1.25 \text{ min}) = 0.625 \text{ min}$$

Applying the PID values of 9.6 (gain), 2.5 minutes per repeat (integral), and 0.625 minutes (derivative) gave the following result in automatic mode:



As you can see, the results are quite poor. The PV is still oscillating with a peak-to-peak amplitude of almost 20% from the last process upset at the time of the 10% downward SP change. Additionally, the output trend is rather noisy, indicating excessive amplification of process noise by the controller.

Next, the closed-loop, or "Ultimate" tuning method of Ziegler and Nichols was applied to this process. Eliminating both integral and derivative control actions from the controller, and experimenting with different gain (proportional) values until self-sustaining oscillations of consistent amplitude were obtained, gave a gain value of 15:



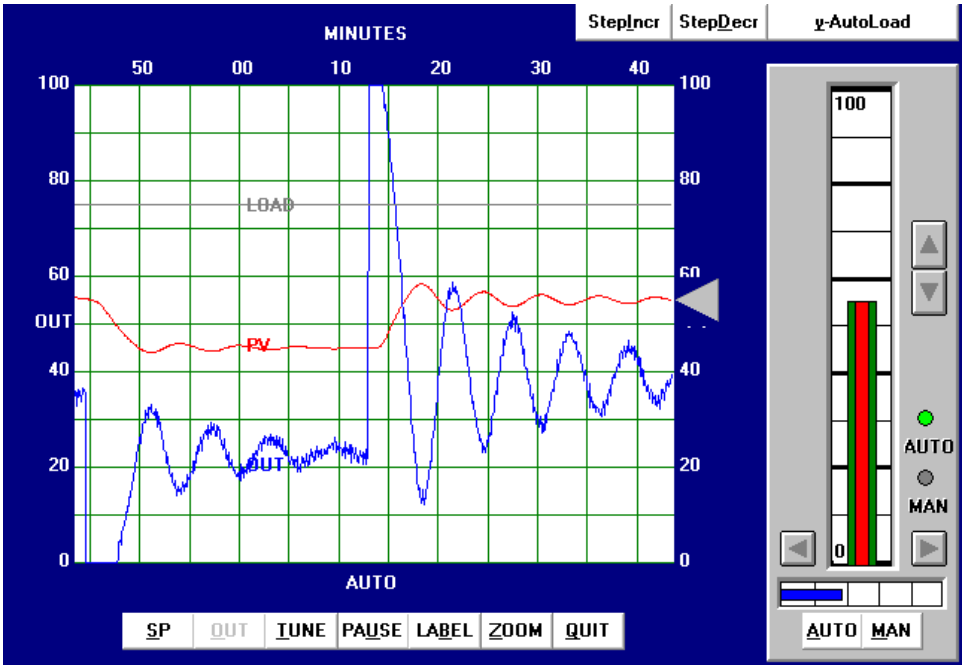
From the trend, we can see that the ultimate period (P_u) is approximately 5.2 minutes in length. Following the Ziegler-Nichols recommendations for PID tuning based on these process characteristics:

$$K_p = 0.6K_u = (0.6)(15) = 9$$

$$\tau_i = \frac{P_u}{2} = \frac{5.2 \text{ min}}{2} = 2.6 \text{ min}$$

$$\tau_d = \frac{P_u}{8} = \frac{5.2 \text{ min}}{8} = 0.65 \text{ min}$$

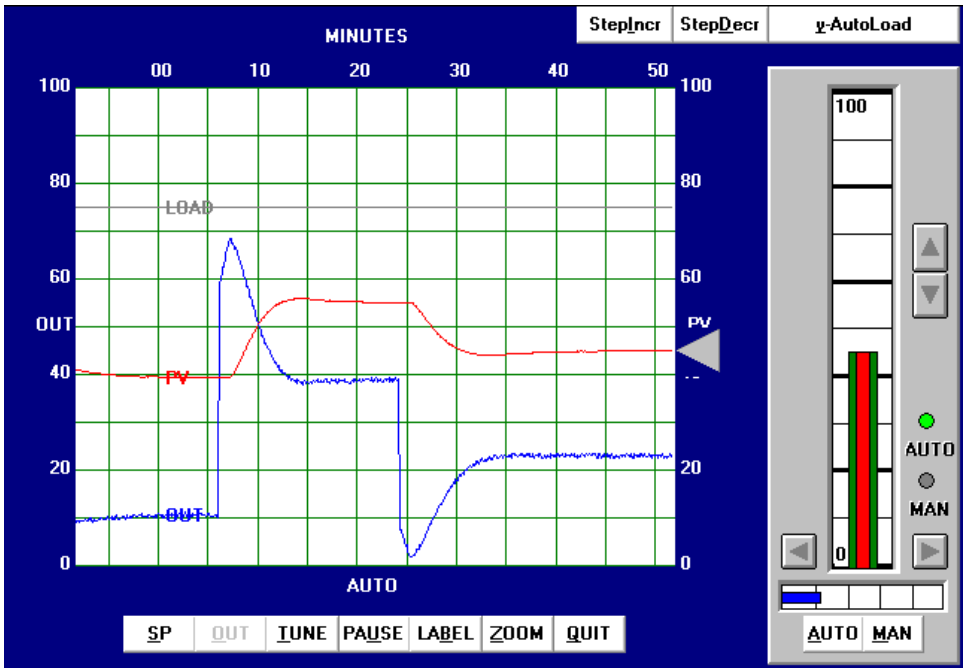
These PID tuning values are quite similar to those predicted by the open loop ("Reaction Rate") method, and so we would expect to see very similar results:



As expected, we still see excessive oscillation following a 10% setpoint change, as well as excessive "noise" in the output trend.

From the initial open-loop (manual output step-change) test, we could see this process was self-regulating with a slow lag and substantial dead time. The self-regulating nature of the process demands at least some integral control action to eliminate offset, but too much will cause oscillation given the long lag and dead times. The existence of over 1 minute of process dead time also prohibits the use of aggressive proportional action. Derivative action, which is generally useful in overcoming lag times, will cause problems here by amplifying process noise. In summary, then, we would expect to use mild proportional, integral, *and* derivative tuning values in order to achieve good control with this process. Anything too aggressive will cause problems for this process.

After some experimenting, the values I arrived at were 3 (gain), 5 minutes (integral), and 0.5 minutes (derivative). These tuning values represent a proportional action only one-third as aggressive as the Ziegler-Nichols recommendation, and an integral action about half as aggressive as the Ziegler-Nichols recommendation. The results of these tuning values in automatic mode are shown here:



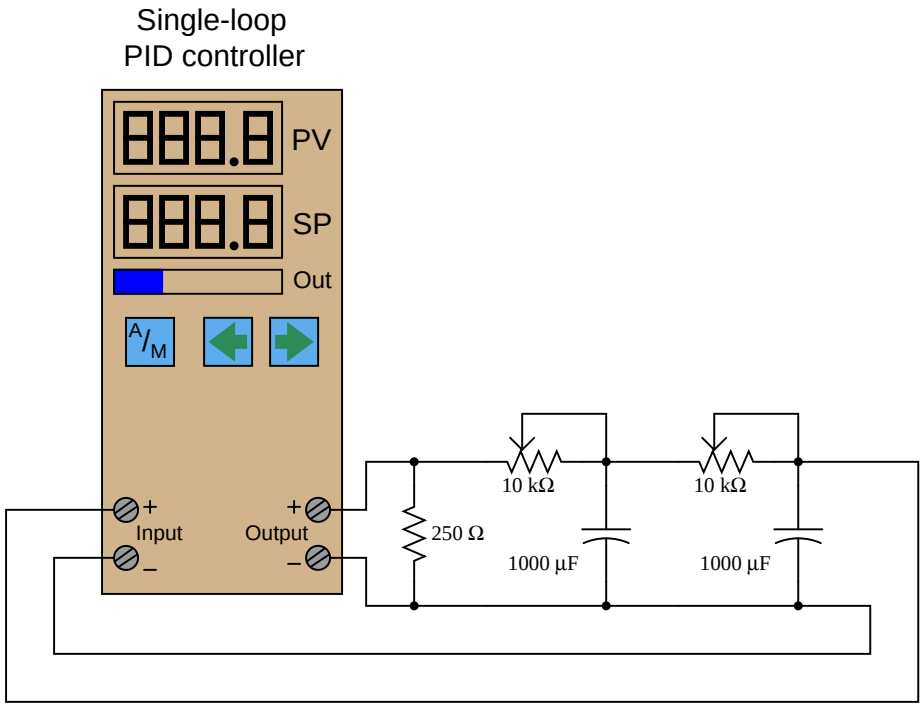
As you can see, the system's response has almost no overshoot (with either a 10% setpoint change or a 15% setpoint change) and very little "noise" on the output trend. Response to setpoint changes is relatively crisp considering the naturally slow nature of the process: each new setpoint is achieved within about 7.5 minutes of the step-change.

27.6 Note to students

Learning how to tune PID controllers is a skill born of much practice. Regardless of how thoroughly you may study the subject of PID control on paper, you really do not understand it until you have spent a fair amount of time actually tuning real controllers.

In order to gain this experience, though, you need access to working processes and the freedom to disturb those processes over and over again. If your school's lab has several "toy" processes built to facilitate this type of learning experience, that is great. However, your learning will grow even more if you have a way to practice PID tuning at your own convenience.

Thankfully, there is a relatively simple way to build your own "process' for PID tuning practice. First, you need to obtain an electronic single-loop PID controller[28] and connect it to a resistor-capacitor network such as this:



The 250 Ω resistor converts the controller's 4-20 mA signal into a 1-5 VDC signal, which then drives the passive integrator (lag) RC networks. The two stages of RC "lag" simulate a self-regulating process with a second-order lag and a steady-state gain of 1. The potentiometers establish the lag times for each stage, providing a convenient way to alter the process characteristics for more tuning

[28] Many instrument manufacturers sell simple, single-loop controllers for reasonable prices, comparable to the price of a college textbook. You need to get one that accepts 1-5 VDC input signals and generates 4-20 mA output signals, and has a "manual" mode of operation in addition to automatic – these features are *very important!* Avoid controllers that can only accept thermocouple inputs, and/or only have time-proportioning (PWM) outputs. Additionally, I strongly recommend you take the time to experimentally learn the actions of proportional, integral, and derivative as outlined in section 26.14 beginning on page 1497 before you embark on any PID tuning exercises.

practice. Feel free to extend the circuit with additional RC lag networks for even more delay (and an even harder-to-tune process!).

Since this simulated "process" is direct-acting (i.e. increasing manipulated variable signal results in an increasing process variable signal), the controller must be configured for *reverse* action (i.e. increasing process variable signal results in a decreasing manipulated variable signal) in order to achieve negative feedback. You are welcome to configure the controller for direct action just to see what the effects will be, but I assure you control will be impossible: the PV will saturate beyond 100% or below 0% no matter how the PID values are set.

References

Lipták, Béla G., *Instrument Engineers' Handbook – Process Control Volume II*, Third Edition, CRC Press, Boca Raton, FL, 1999.

Mollenkamp, Robert A., *Introduction to Automatic Process Control*, Instrument Society of America, Research Triangle Park, NC, 1984.

Palm, William J., *Control Systems Engineering*, John Wiley & Sons, Inc., New York, NY, 1986.

Shinskey, Francis G., *Energy Conservation through Control*, Academic Press, New York, NY, 1978.

Shinskey, Francis G., *Process-Control Systems – Application / Design / Adjustment*, Second Edition, McGraw-Hill Book Company, New York, NY, 1979.

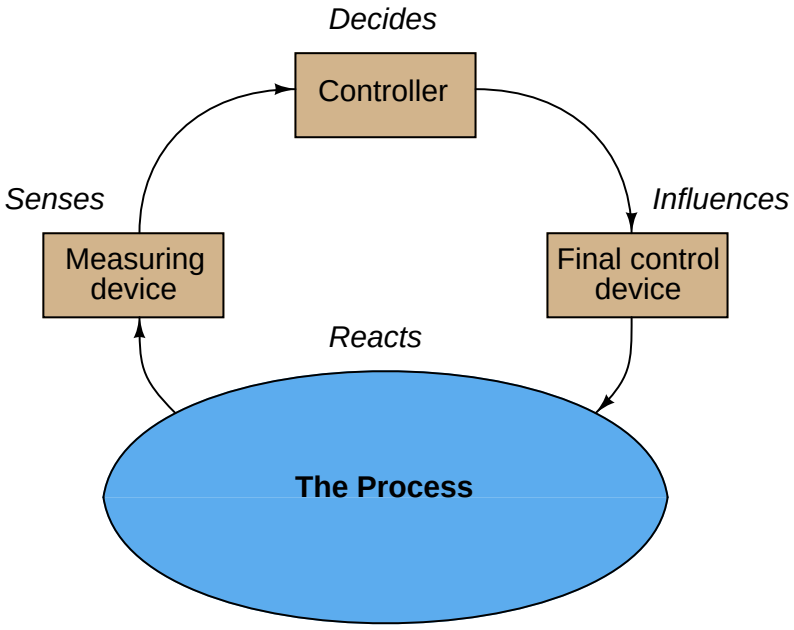
St. Clair, David W., *Controller Tuning and Control Loop Performance, a primer*, Straight-Line Control Company, Newark, DE, 1989.

Ziegler, J. G., and Nichols, N. B., *Optimum Settings for Automatic Controllers*, Transactions of the American Society of Mechanical Engineers (ASME), Volume 64, pages 759-768, Rochester, NY, November 1942.

Chapter 28

Basic process control strategies

In a simple control system, a process variable (PV) is measured and compared with a setpoint value (SP). A manipulated variable (MV, or output) signal is generated by the controller and sent to a final control element, which then influences the process variable to achieve stable control. The algorithm by which the controller develops its output signal is typically PID (Proportional-Integral-Derivative), but other algorithms may be used as well:

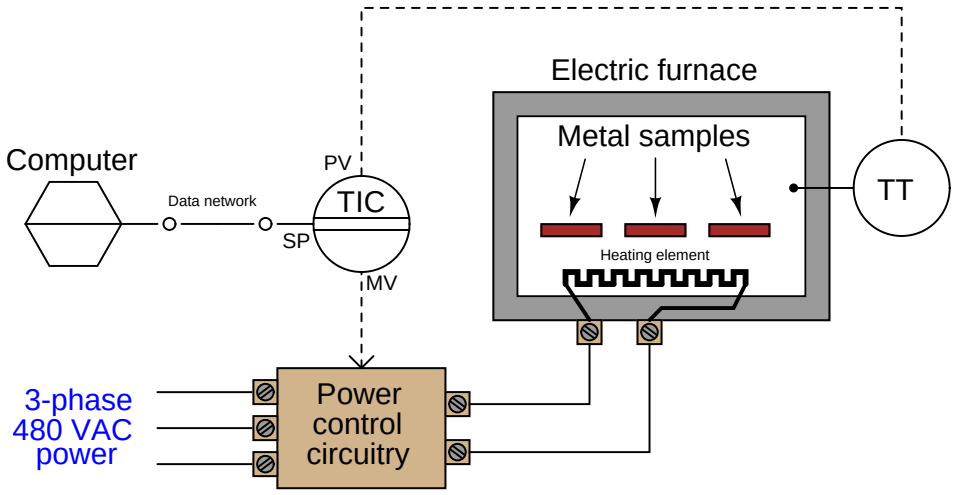


This form of simple control may be improved upon and expanded for a greater range of process applications by interconnecting multiple controllers and/or redirecting measurement and control signals in more complex arrangements. An exploration of some of the more common control system configurations is the subject of this chapter.

28.1 Supervisory control

In a manually-controlled process, a human operator directly actuates some form of final control element (usually a valve) to influence a process variable. Simple automatic ("regulatory") control relieves human operators of the need to continually adjust final control elements by hand, replacing this task with the occasional adjustment of setpoint values. The controller then manipulates the final control element to hold the process variable at the setpoint value determined by the operator.

The next step in complexity after simple automatic control is to automate the adjustment of the setpoint for a process controller. A common implementation of this concept is the automatic cycling of setpoint values according to a timed schedule. An example of this is a temperature controller for a heat-treatment furnace used to temper metal samples:



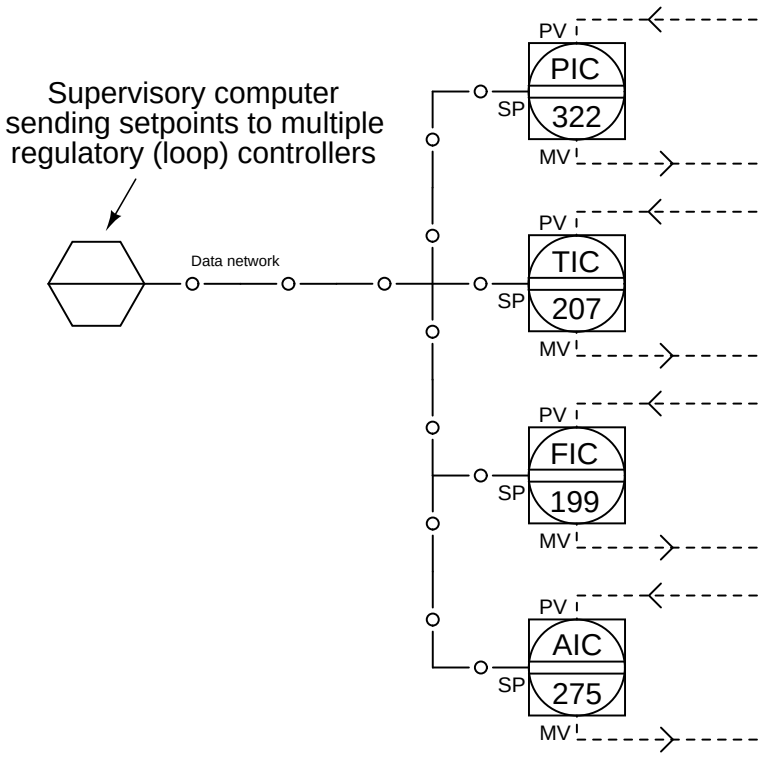
Here, a computer communicates setpoint values to the temperature indicating controller (TIC) over a digital network interface such as Ethernet. From the temperature controller's perspective, this is a *remote* setpoint signal, as opposed to a *local* setpoint value which would be set by a human operator at the controller faceplate. Since the heat-treatment of metals requires particular temperature ranges and rates of change over time, this control system relieves the human operator of having to manually adjust setpoint values again and again during heat-treatment cycles. Instead, the computer schedules different setpoint values at different times (even setpoint values that change steadily at a certain rate over a period of time) according to the needs of the particular metal type and treatment type. Such a control scheme is quite common for heat-treating processes, and it is referred to as *ramp and soak*[1].

Supervisory setpoint control is also used in the chemical processing industries to optimize production by having a powerful computer provide setpoint adjustments to regulatory controls based on mathematical models of the process and optimization constraints. Dividing the control system into two layers – the upper-layer supervisory control computer and the lower-layer process controllers

[1] In honor of the system's ability to slowly "ramp" temperature up or down at a specified rate, then "soak" the metal at a constant temperature for set periods of time. Many single-loop process controllers have the ability to perform ramp-and-soak setpoint scheduling without the need of an external "supervisory" computer.

– provides a flexible means of achieving basic control and advanced (optimization) controls without burdening any one piece of control hardware with too much responsibility over the process. If the supervisory computer fails for whatever reason, for example, the regulatory controls (either panel-mounted controllers or more likely a distributed control system) may default to local setpoint control.

Such optimization control systems are usually built over a digital network for reasons of convenience. A single network cable not only is able to handle the frequent setpoint changes from the supervisory computer to the multitude of process loop controllers, but it may also carry process variable information from those controllers back to the supervisory computer so it has data for its optimization algorithms to operate on:

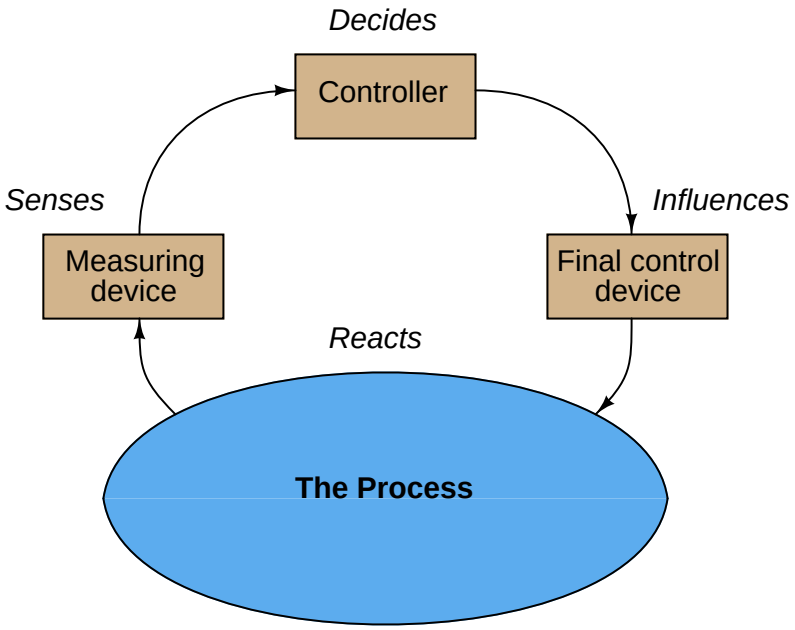


The complexity of these optimization algorithms is limited only by the computational power of the supervisory computer and the creativity of the programmers and engineers who implement it. A modern trend in process optimization for industries able to produce varying proportions of different products from the same raw material feed is to have computer algorithms select and optimize production not only for maximum cost efficiency, but also for maximum market sales and minimum storage of volatile product[2].

[2]I once attended a meeting of industry representatives where one person talked at length about a highly automated lumber mill where logs were cut into lumber not only according to minimum waste, but also according to the real-time market value of different board types and stored inventory. The joke was, if the market value of wooden toothpicks suddenly spiked up, the control system would shred every log into toothpicks in an attempt to maximize profit!

28.2 Cascade control

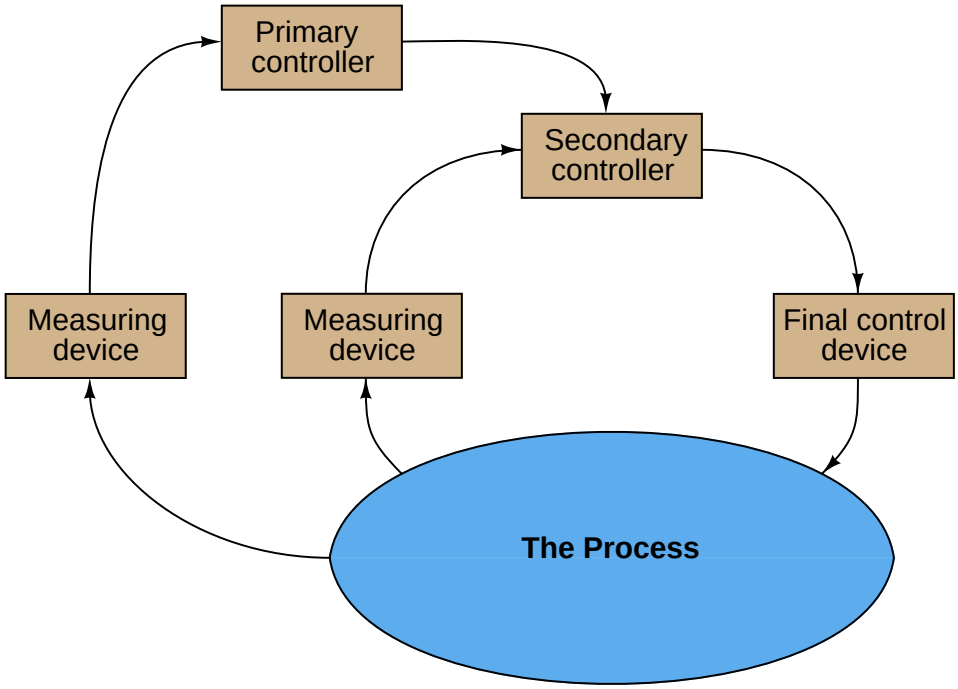
A simple control system drawn in block diagram form looks like this:



Information from the measuring device (e.g. transmitter) goes to the controller, then to the final control device (e.g. control valve), influencing the process which is sensed again by the measuring device. The controller's task is to inject the proper amount of negative feedback such that the process variable stabilizes over time. This flow of information is collectively referred to as a feedback control *loop*.

To *cascade* controllers means to connect the output signal of one controller to the setpoint of another controller, with each controller sensing a different aspect of the same process. The first controller (called the *primary*, or *master*) essentially "gives orders" to the second controller (called the *secondary* or *slave*) via a *remote setpoint* signal.

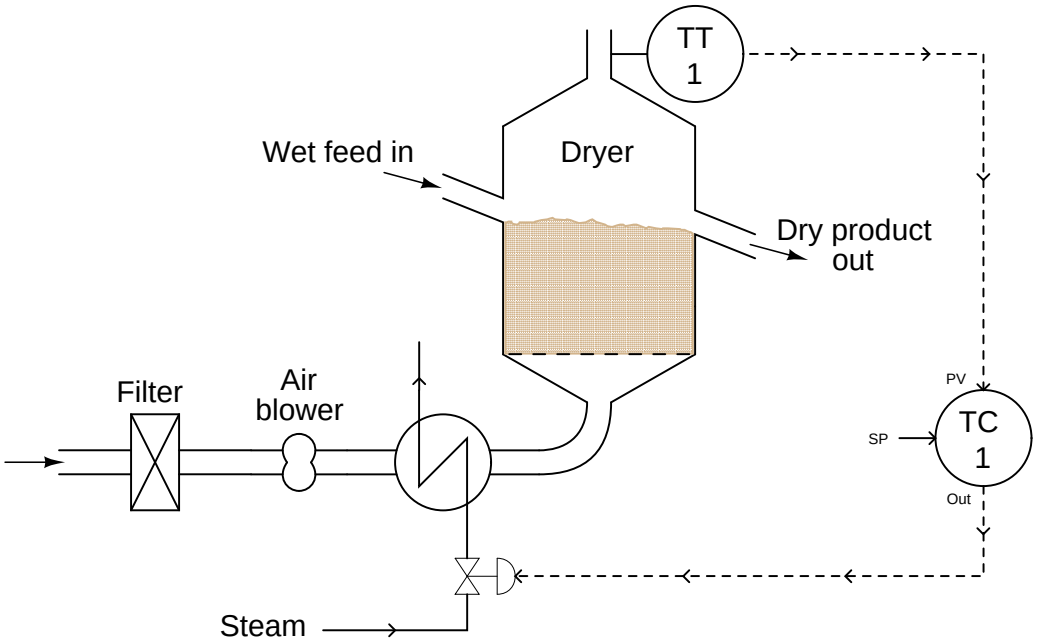
Thus, a cascade control system consists of two feedback control loops, one nested inside the other:



A very common example of cascade control is a *valve positioner*, which receives a command signal from a regular process controller, and in turn works to ensure the valve stem position precisely matches that command signal. The control valve's stem position is the process variable (PV) for the positioner, just as the command signal is the positioner's setpoint (SP). Valve positioners therefore act as "slave" controllers to "master" process controllers controlling pressure, temperature, flow, or some other process variable.

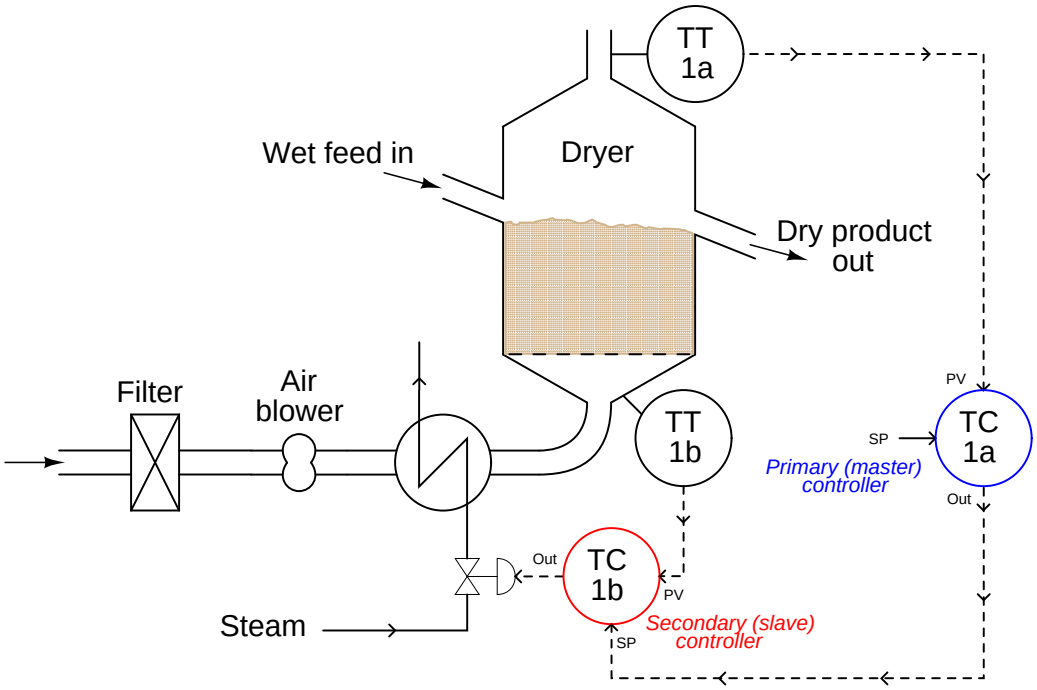
The purpose of cascade control is to achieve greater stability of the primary process variable by regulating a secondary process variable in accordance with the needs of the first. An essential requirement of cascaded control is that the secondary process variable be faster-responding (i.e. less lag time) than the primary process variable.

For example, consider the following dryer system where heated air is used to force-evaporate water from a granular solid. The primary process variable is the outlet air exiting the dryer, which should be maintained at a high enough temperature to ensure water will not remain in the upper layers of the solid material. This outlet temperature is fairly slow to react, as the solid material mass creates a large lag time:



There are several parameters influencing the temperature of the outlet air other than the moisture content of the drying material. These include air flow, ambient air temperature, and variations in steam temperature. Each one of these variables is a *load* on the process variable we are trying to control (outlet air temperature). If any of these parameters were to suddenly change, the effect would be slow to register at the outlet temperature even though there would be immediate impact at the bottom of the dryer where the heated air enters. Correspondingly, the control system would be slow to correct for any of these changing loads.

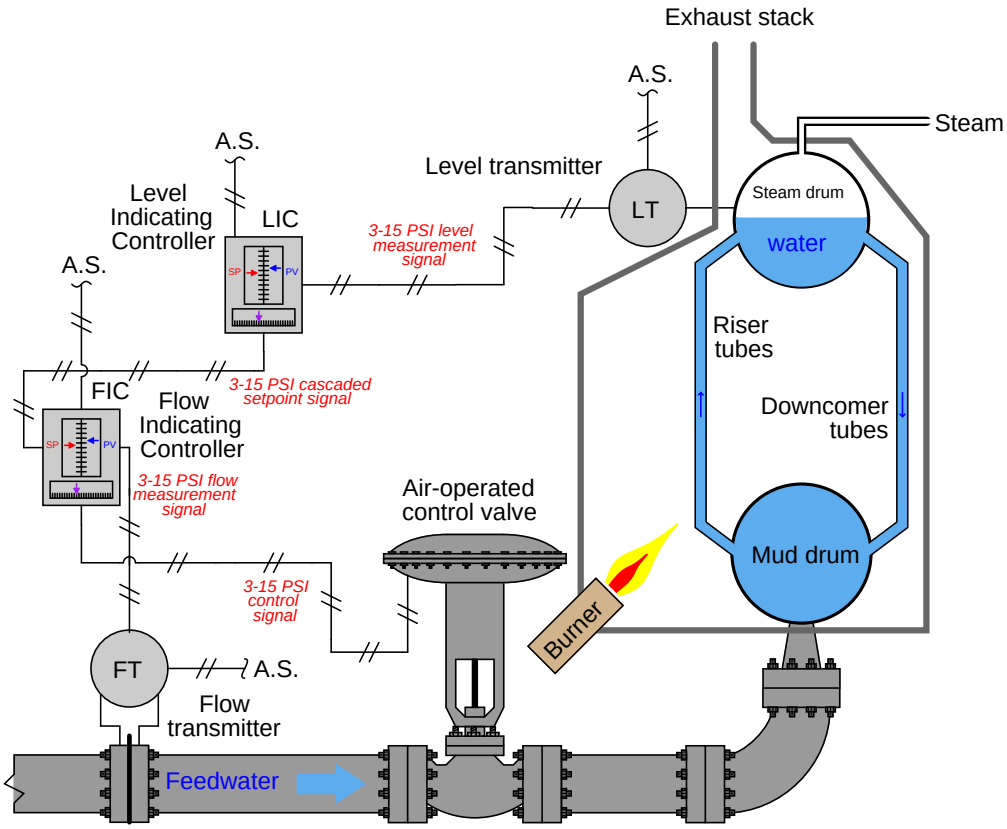
One way to help gain better control over this dryer system is to install a second temperature transmitter at the inlet duct of the dryer, with its own controller to adjust steam flow at the command of the primary controller:



Now, if any of the loads related to incoming air flow or temperature were to change, the secondary controller (TC-1b) would *immediately* compensate by adjusting steam flow through the heat exchanger to maintain a constant air temperature entering the dryer. Thus, the "slave" control loop (1b) helps stabilize the "master" control loop (1a) by reacting to changes in one of the variables influencing it.

A helpful way to think of this is to consider the slave controller as *shielding* the master controller from the loads previously mentioned (incoming air flow, ambient temperature, and steam temperature). Of course, these variables still act as loads to the slave controller, as it must continuously adjust the steam valve to compensate for changes in air flow, ambient air temperature, and steam temperature. However, so long as the slave controller does a good job of stabilizing the air temperature entering the dryer, the master controller will never "see" the effects of those load changes. Responsibility for incoming air temperature has been delegated to the slave controller, and as a result the master controller is conveniently isolated from the loads impacting that loop.

A common implementation of cascade control is where a flow controller receives a setpoint from some other process controller (pressure, temperature, level, analytical, etc.), fluid flow being one of the fastest-responding process types in existence. A feedwater control system for a steam boiler – shown here in pneumatic form – is a good example:



The "secondary" flow controller works to maintain feedwater flow to the boiler at whatever flow rate desired by the level controller. If feedwater pressure happens to increase or decrease, any resulting changes in flow will be quickly countered by the flow controller without the level controller having to act from a consequent change in steam drum water level. Thus, cascade control works to guard against steam drum level instability resulting from changes in the feedwater flow caused by factors outside the control system.

It is worth noting that the inclusion of a flow control "slave" loop to this boiler water level control system helps to overcome a potential problem of the control valve: nonlinear behavior. In the control valves chapter, we explore the phenomenon of *installed valve characteristics* (Section 25.1.13, on page 1349), specifically noting how changes in pressure drop across a control valve influences its throttling behavior. The result of these pressure changes is a non-linearization of valve response, such that the valve tends to be more responsive near its closed position and less responsive near its open position. One of the benefits of cascaded flow control is that this problem becomes confined to the secondary

(flow control) loop, and is effectively removed from the primary control loop. To phrase it simply, distorted valve response becomes "the flow controller's problem" rather than something the level controller must manage. The result is a level control system with more predictable response.

An analogy for considering cascade control is that of *delegation* in a work environment. If a supervisor delegates some task to a subordinate, and that subordinate performs the task without further need of guidance or assistance from the supervisor, the supervisor's job is made easier. The subordinate takes care of all the little details that would otherwise burden the supervisor if the supervisor had no one to delegate to.

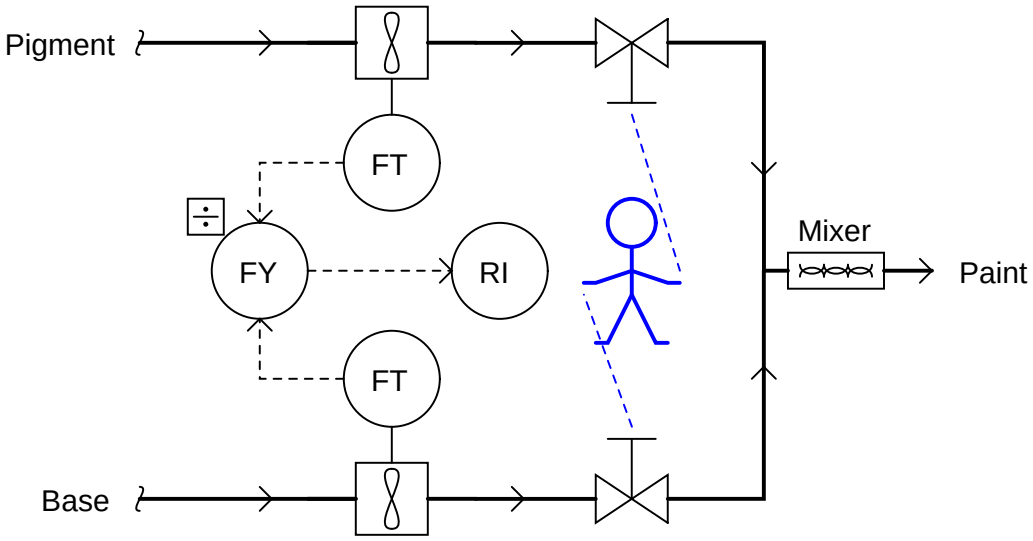
A necessary step in implementing cascade control is to ensure the secondary ("slave") controller is well-tuned *before* any attempt is made to tune the primary ("master") controller. Just a moment's thought is all that is needed to understand why this precedence in tuning must be: it is a simple matter of dependence. The slave controller does not depend on good tuning in the master controller in order to control the slave loop. If the master controller were placed in manual (effectively turning off its automatic response), the slave controller would simply control to a constant setpoint. However, the master controller most definitely depends on the slave controller being well-tuned in order to fulfill the master's "expectations." If the slave controller were placed in manual mode, the master controller would not be able to exert any control over its process variable whatsoever. Clearly then, the slave controller's response is essential to the master controller being able to control its process variable, therefore the slave controller must be the first one to tune.

28.3 Ratio control

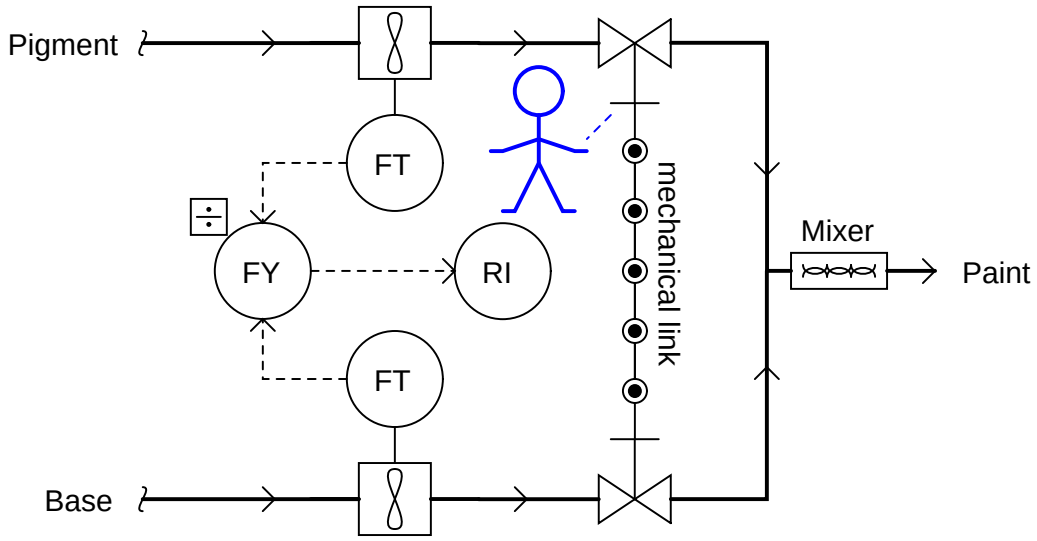
Most people reading this book have likely had the experience of adjusting water temperature as they took a shower with dual valves: one valve controlling hot water and one valve controlling cold water. In order to adjust water temperature, the *proportion* of one valve opening to the other must be changed. Increasing or decreasing total water flow rate without upsetting the outlet temperature is a matter of adjusting both valves in the same direction, maintaining that same proportion of hot to cold water flow.

Although you may not have given it much thought while taking your shower, you were engaged in a control strategy known as *ratio control*, where the ratio of one flow rate to another is controlled for some desired outcome. Many industrial processes also require the precise mixing of two or more ingredients to produce a desired product. Not only do these ingredients need to be mixed in proper proportion, but it is usually desirable to have the total flow rate subject to arbitrary increases and decreases so production rate as a whole may be altered at will.

A simple example of ratio control is in the production of paint, where a base liquid must be mixed with one or more pigments to achieve a desired consistency and color. A manually controlled paint mixing process, similar to the hot and cold water valve "process" in some home showers, is shown here. Two flowmeters, a ratio calculating relay, and a display provide the human operator with a live measurement of pigment-to-base ratio:

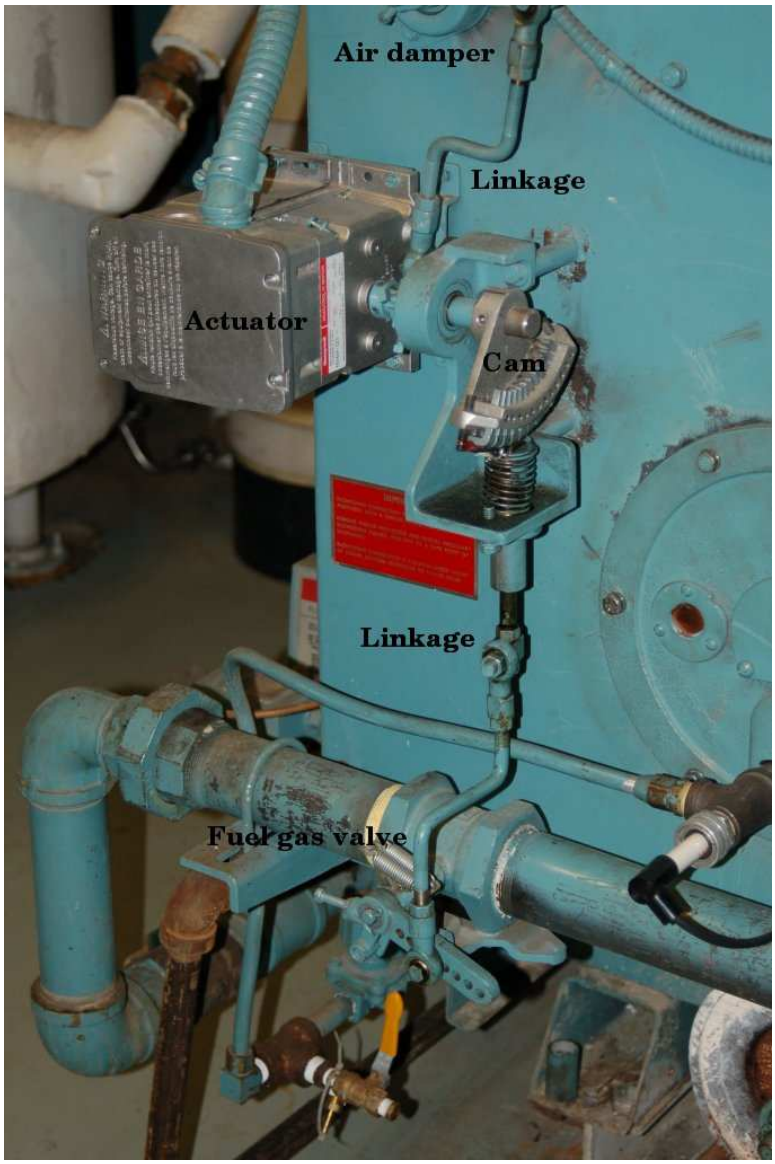


One alteration we could make to this mixing system is to link the two manual control valve handles together in such a way that the ratio of base to pigment was *mechanically* established. All the human operator needs to do now is move the one link to increase or decrease mixed paint production:



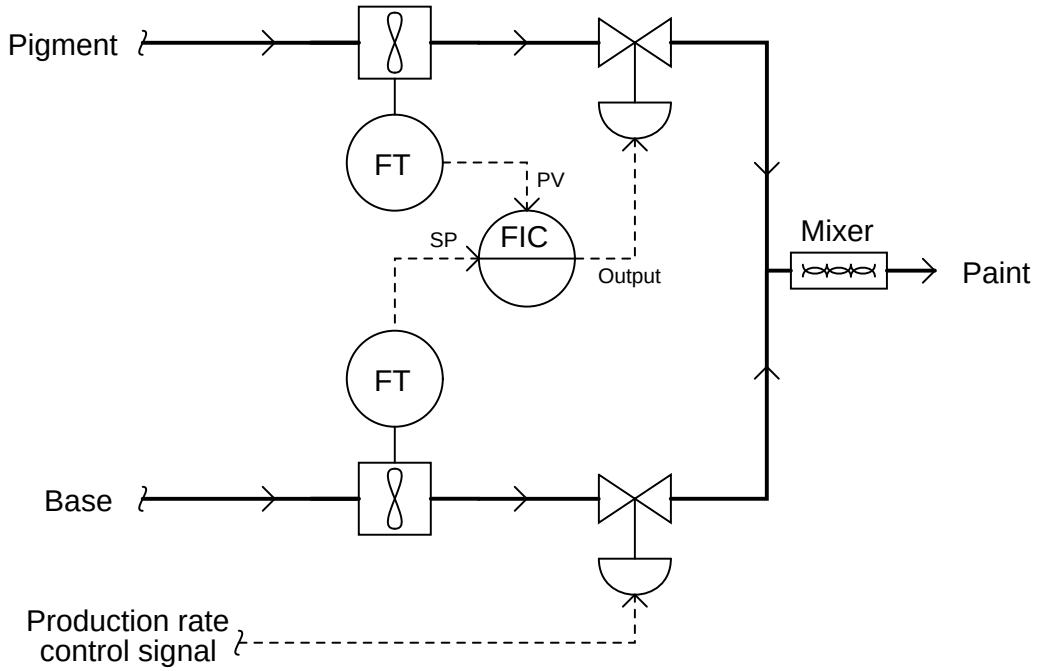
Adjusting the pigment-to-base ratio is now a matter of adjusting the linkage ratio, a task most likely performed by a mechanic or someone else knowledgeable in the operation of mechanical linkages. The convenience of total flow adjustment gained by the link comes at the price of inconvenient ratio adjustment.

Mechanical link ratio-control systems are commonly used to manage simple burners, proportioning the flow rates of fuel and air for clean, efficient combustion. A photograph of such a system appears here, showing how the fuel gas valve and air damper motions are coordinated by a single rotary actuator:



As you can see in this photo, the fuel gas valve is actuated by means of a cam, allowing precise "tuning" of the valve characteristics for consistent fuel/air ratio across a wide range of firing rates. Making ratio adjustments in such a linkage system is obviously a task for a skilled mechanic or technician.

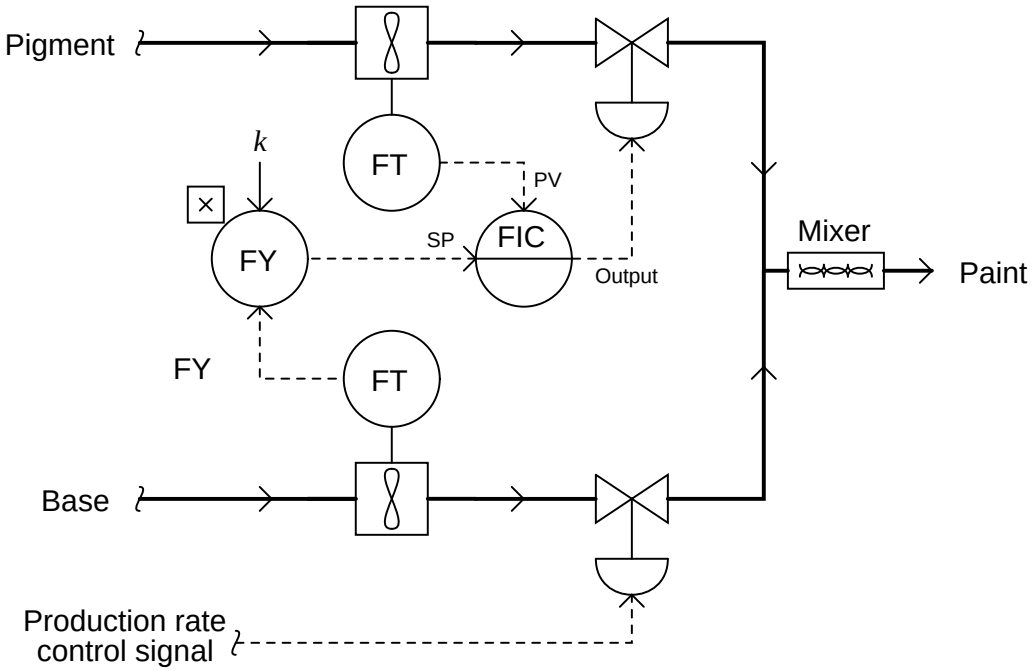
A more automated approach to the general problem of ratio control involves the installation of a flow control loop on one of the lines, while keeping just a flow transmitter on the other line. The signal coming from the uncontrolled flow transmitter becomes the setpoint for the flow control loop:



Here, the flow transmitter on the uncontrolled line measures the flow rate of base, sending a flow rate signal to the pigment flow controller which acts to match flow rates. If the calibrations of each flow transmitter are precisely equal to one another, the ratio of pigment to base will be 1:1 (equal). The flow of base liquid into the mixing system is called a *wild flow* or *wild variable*, since this flow rate is not controlled by the ratio control system. The only purpose served by the ratio control system is to match the pigment flow rate to the wild (base) flow rate, so the same ratio of pigment to base will always be maintained regardless of total flow rate. Thus, the flow rate of pigment will be held *captive* to match the "wild" base flow rate, which is why the controlled variable in a ratio system is sometimes called the *captive variable* (in this case, a captive *flow*).

As with the mechanically-linked manual ratio mixing system, this ratio control system provides convenient total flow control, but inconvenient control over mixing ratio. In order to alter the ratio of pigment to base, someone would have to re-calibrate one (or both) flow transmitters. To achieve a 2:1 ratio of base to pigment, for example, the base flow transmitter's range would have to be double that of the pigment flow transmitter. This way, an equal percentage of flow registered by both flow transmitters (as the ratio controller strives to maintain equal percentage values of flow between pigment and base) would actually result in twice the amount of base flow than pigment flow.

We may incorporate convenient ratio adjustment into this system by adding another component (or function block) to the control scheme: a device called a *signal multiplying relay* (or alternatively, a *ratio station*). This device (or computer function) takes the flow signal from the base (wild) flow transmitter and multiplies it by some constant value (k) before sending the signal to the pigment (captive) flow controller as a setpoint:



With identical flow range calibrations in both flow transmitters, this multiplying constant k directly determines the base-to-pigment ratio (i.e. the ratio will be 1:1 when $k = 1$; the ratio will be 2:1 when $k = 2$, etc.). If the k value is easily adjusted by a human operator, mixing ratio becomes a very simple parameter to change at will, just as the total production rate is easy to adjust by moving the base flow control valve.

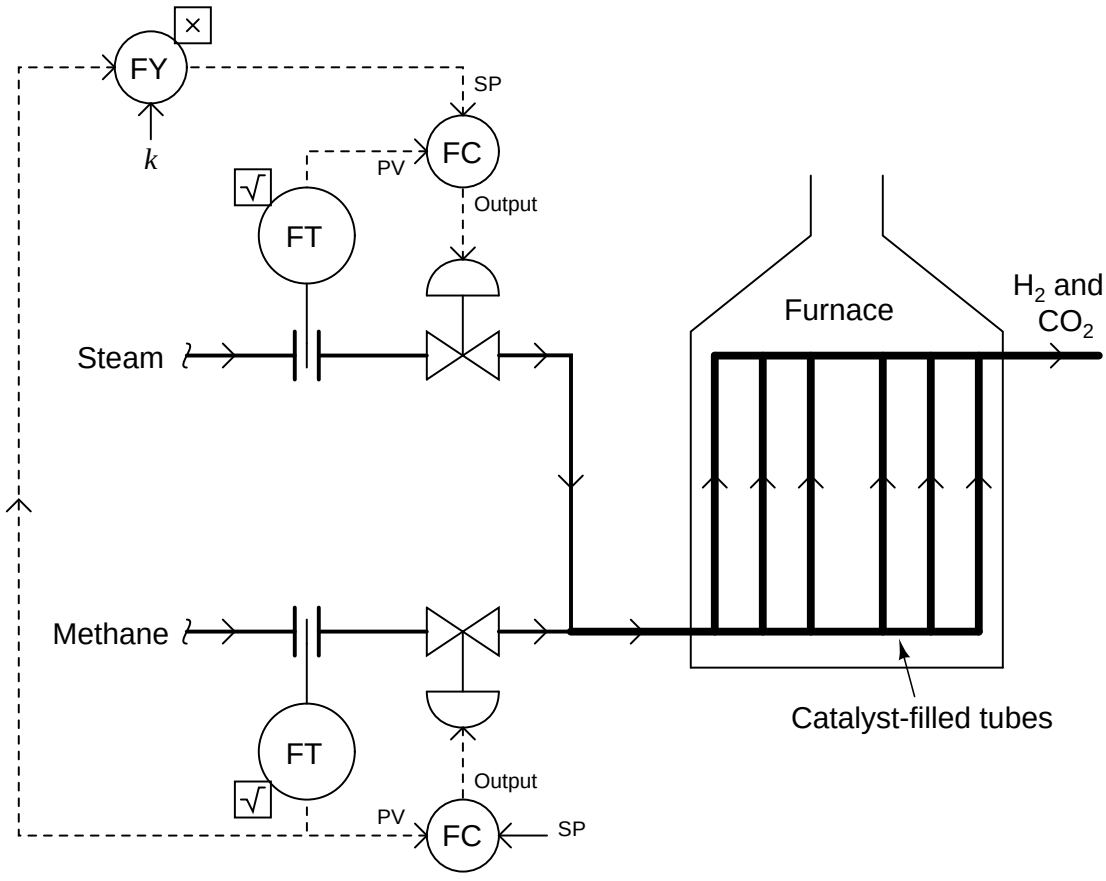
Another example of ratio control at work is in a process whereby hydrocarbon gases (usually methane) are converted into hydrogen gas and carbon dioxide gas. This is known as the *steam-hydrocarbon reforming process*, and it is one of the more popular ways of generating hydrogen gas for industrial use. The overall reaction for this process with methane gas (CH_4) and steam (H_2O) as the reactants is as follows[3]:

$$CH_4 + 2H_2O \rightarrow 4H_2 + CO_2$$

[3]The conversion from hydrocarbon and steam to hydrogen and carbon dioxide is typically a two-stage process: the first (*reforming*) stage produces hydrogen gas and carbon monoxide, while a second (*water-gas-shift*) stage adds more steam to convert the carbon monoxide into carbon dioxide with more hydrogen liberated. Both reactions are endothermic, with the reforming reaction being more endothermic than the water-gas-shift reaction.

This is an endothermic reaction, which means a net input of energy is required to make it happen. Typically, the hydrocarbon gas and steam are mixed together in a heated environment in the presence of a catalyst (to reduce the activation energy requirements of the reaction). This usually takes the form of catalyst-packed metal tubes inside a gas-fired furnace. It is important to control the proportion of gas to steam flow into this process. Too much hydrocarbon gas, and the result will be "coking" (solid hydrocarbon deposits) inside the heated tubes and on the surface of the catalyst beads, decreasing the efficiency of the process over time. Too much steam and the result is wasted energy as unreacted steam simply passes through the heater tubes, absorbing heat and carrying it away from the catalyst where it would otherwise do useful work.

One way to achieve the proper ratio of hydrocarbon gas to steam flow is to install a normal flow control loop on one of these two reactant feed lines, then use that process variable (flow) signal as a setpoint to a flow controller installed on the other reactant feed line. This way, the second controller will maintain a proper balance of flow to proportionately match the flow rate of the other reactant. An example P&ID is shown here, where the methane gas flow rate establishes the setpoint for steam flow control:



Note how the methane gas flow transmitter signal goes both to the methane flow controller

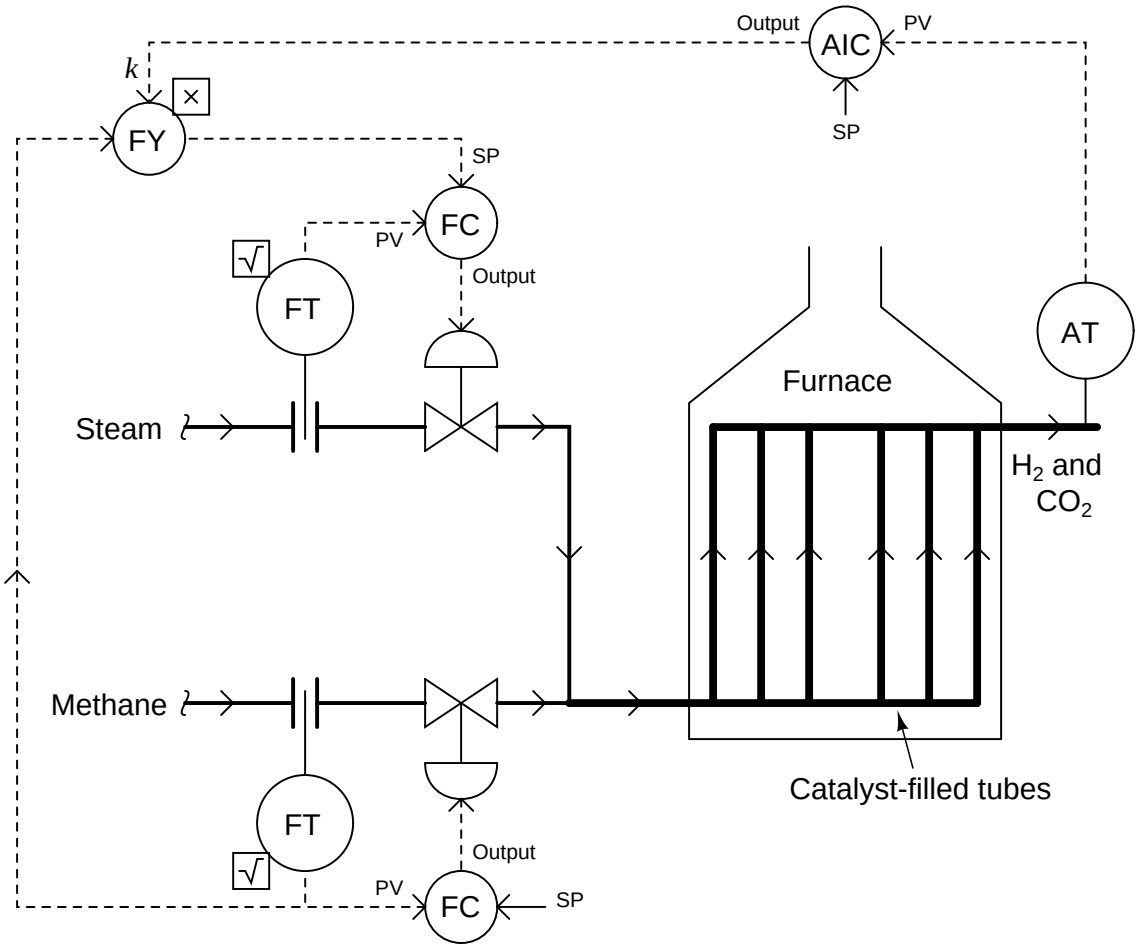
and to a *multiplying relay* that multiplies this signal by a constant value (k) before passing it on to the steam flow controller as a setpoint. This k value sets the *ratio* of steam flow to methane flow. Although this might appear to be a cascade control system at first glance, it is actually quite different. In a cascade system, the *output* of one controller becomes the setpoint for another. Here in a ratio control system, the *process variable* of one controller becomes the setpoint for another, such that two process variables remain in constant proportion (ratio) to one another.

If the two flow transmitters are compensated to measure mass flow, the ideal value of k should be set such that two molecules of steam vapor (H_2O) enter the reforming furnace for every one molecule of methane (CH_4). With a 2-to-1 molecular ratio of steam to methane (2 moles of steam per one mole of methane), this equates to a 9-to-4 mass flow ratio once the formula weights of steam and methane are calculated[4]. Thus, if the methane and gas flowmeters are calibrated for equal mass flow ranges, the ideal value for k should be $\frac{9}{4}$, or 2.25. Alternatively, the flow transmitter calibrations could be set in such a way that the ideal ratio is intrinsic to those transmitters' ranges (i.e. the methane flow transmitter has 2.25 times the mass flow range of the steam flow transmitter), with k set to an ideal value of 1. This way a 9:4 ratio of methane mass flow to steam mass flow will result in equal percentage output values from both flow transmitters. In practice, the value for k is set a bit higher than ideal, in order to ensure just a little excess steam to guard against coking inside the reaction heater tubes[5].

[4] Steam has a formula weight of 18 amu per molecule, with two hydrogen atoms (1 amu each) and one oxygen atom (16 amu). Methane has a formula weight of 16 amu per molecule, with one carbon atom (12 amu) and four hydrogen atoms (1 amu each). If we wish to have a molecular ratio of 2:1, steam-to-methane, this makes a formula weight ratio of 36:16, or 9:4.

[5] It is quite common for industrial control systems to operate at ratios a little bit "skewed" from what is stoichiometrically ideal due to imperfect reaction efficiencies. Given the fact that no chemical reaction ever goes to 100% completion – simply because 100% mixing is virtually impossible – a decision must be made as to which form of incompleteness is worse. In a steam-hydrocarbon reforming system, we must ask ourselves which is worse: excess (unreacted) steam at the outlet, or excess (unreacted) hydrocarbon at the outlet. Excess hydrocarbon content will "coke" the catalyst and heater tubes, which is very bad for the process over time. Excess steam merely results in a bit more operating energy loss, with no degradation to equipment life. The choice, then, is clear: it is better to operate this process "lean" (more steam than ideal) than "rich" (less steam than ideal).

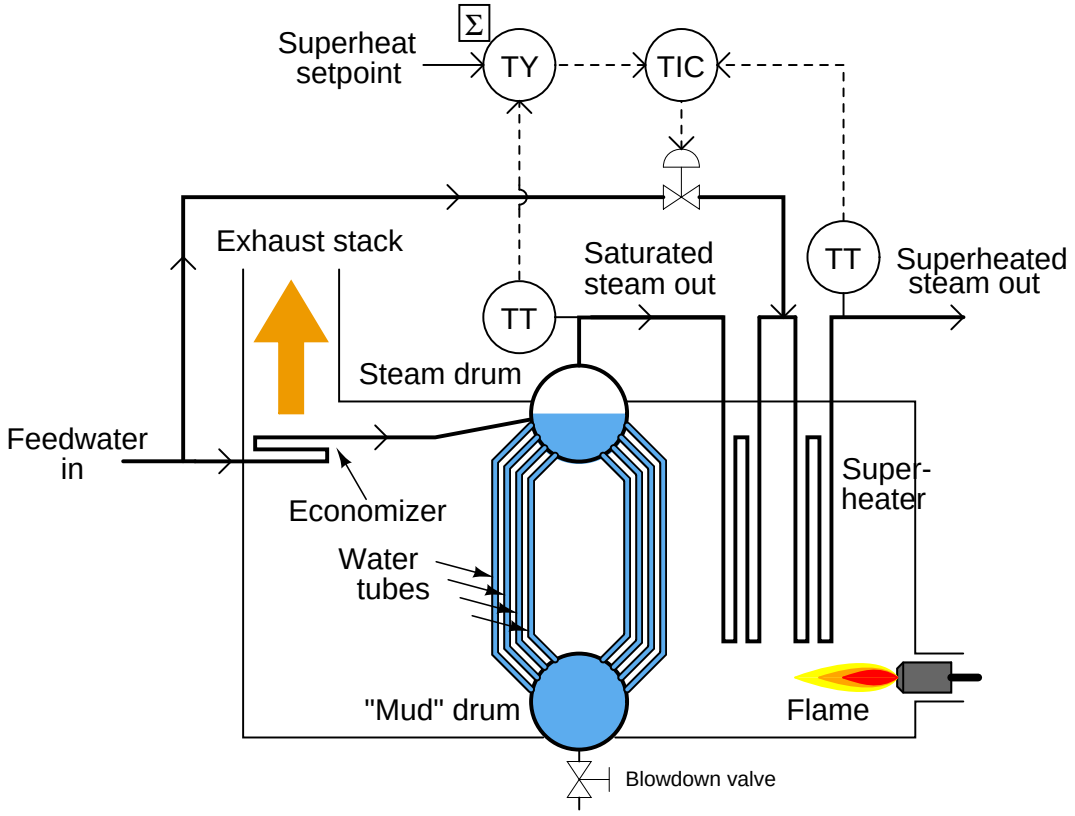
We could add another layer of sophistication to this ratio control system by installing a gas analyzer at the outlet of the reaction furnace designed to measure the composition of the product stream. This analyzer's signal could be used to adjust the value of k so the ratio of steam to methane would automatically vary to ensure optimum production quality even if the feedstock composition (i.e. percentage concentration of methane in the hydrocarbon gas input) changes:



28.4 Relation control

A control strategy similar to ratio control is *relation* control. This is similar to ratio control in that a "wild" variable determines the setpoint for a captive variable, but with relation control the mathematical relationship between the wild and captive variables is one of addition (or subtraction) rather than multiplication (or division). In other words, a relation control system works to maintain a specific *difference* between wild and captive flow values, whereas a ratio control system works to maintain a specific *ratio* between wild and captive flow values.

An example of relation control appears here, where a temperature controller for a steam superheater on a boiler receives its setpoint from the biased output of a temperature transmitter sensing the temperature of saturated steam (that is, steam exactly at the boiling point of water) in the steam drum:



It is a basic principle of thermodynamics that the vapor emitted at the surface of a boiling liquid will be at the same temperature as that liquid. Furthermore, any heat lost from that vapor will cause at least some of that vapor to condense back into liquid. In order to ensure the vapor is "dry" (i.e. it may lose substantial heat energy without condensing), the vapor must be heated beyond the liquid's boiling point at some later stage in the process.

Steam within the steam drum of a boiler is *saturated* steam: at the same temperature as the boiling water. Any heat lost from saturated steam causes at least some of it to immediately condense back into water. In order to ensure "dry" steam output from the boiler, the saturated steam taken from the steam drum must be further heated through a set of tubes called a *superheater*. The resulting "dry" steam is said to be *superheated*, and the difference between its temperature and the temperature of the boiling water (saturated steam) is called *superheat*.

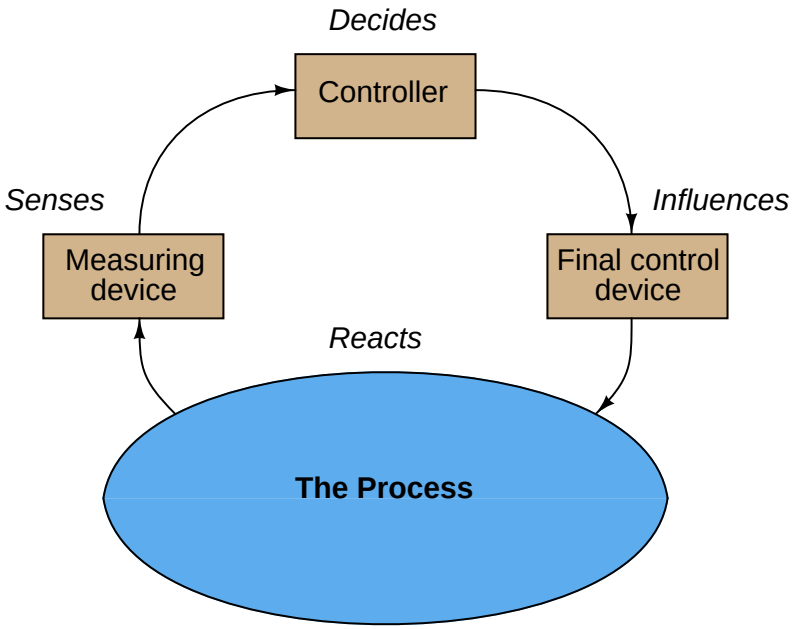
This control system maintains a set amount of superheat by measuring the saturated steam's temperature (within the steam drum), adding a "superheat setpoint" bias value to that signal, then passing the biased signal to the temperature indicating controller (TIC) where the superheated steam temperature is regulated by adding water[6] to the superheated steam. With this system in place, the boiler operator may freely define how much superheat is desired, and the controller attempts to maintain the superheated steam at that much higher temperature than the saturated steam in the drum, over a wide range of saturated steam temperatures.

A ratio control system would not be appropriate here, since what we desire in this process is a controlled *offset* (rather than a controlled *ratio*) between two steam temperatures. The control strategy looks very much like a ratio control, except for the substitution of a summing function instead of a multiplying function.

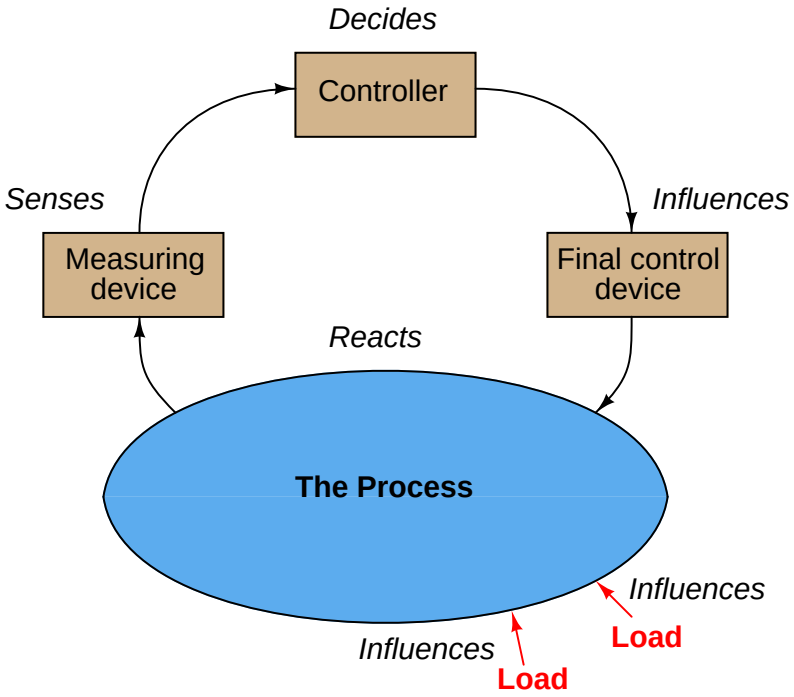
[6] This mixing of superheated steam and cold water happens in a specially-designed device called a *desuperheater*. The basic concept is that the water will absorb heat from the superheated steam, turning that injected water completely into steam and also reducing the temperature of the superheated steam. The result is a greater volume of steam than before, at a reduced temperature. So long as some amount of superheat remains, the de-superheated steam will still be "dry" (above its condensing temperature). The desuperheater control merely adds the appropriate amount of water until it achieves the desired superheat value.

28.5 Feedforward control

Feedback control works on the principle of information from the outlet of a process being "fed back" to the input of that process for corrective action. A block diagram of feedback control looks like a loop:



The reason any control system is necessary at all[7] to maintain a process variable at some stable value is the existence of something called a *load*. A "load" is some variable influencing a process that is not itself under direct control, and may be represented in the block diagram as an arrow entering the process, but not within the control loop:



For example, consider the problem of controlling the speed of an automobile. In this scenario, vehicle speed is the process variable being measured and controlled, while the final control device is the accelerator pedal controlling engine power output. If it were not for the existence of hills and valleys, head-winds and tail-winds, air temperature changes, road surface variations, and a host of other "load" variables affecting car speed, it would be an elementary matter to drive at a constant speed: simply hold the accelerator position constant.

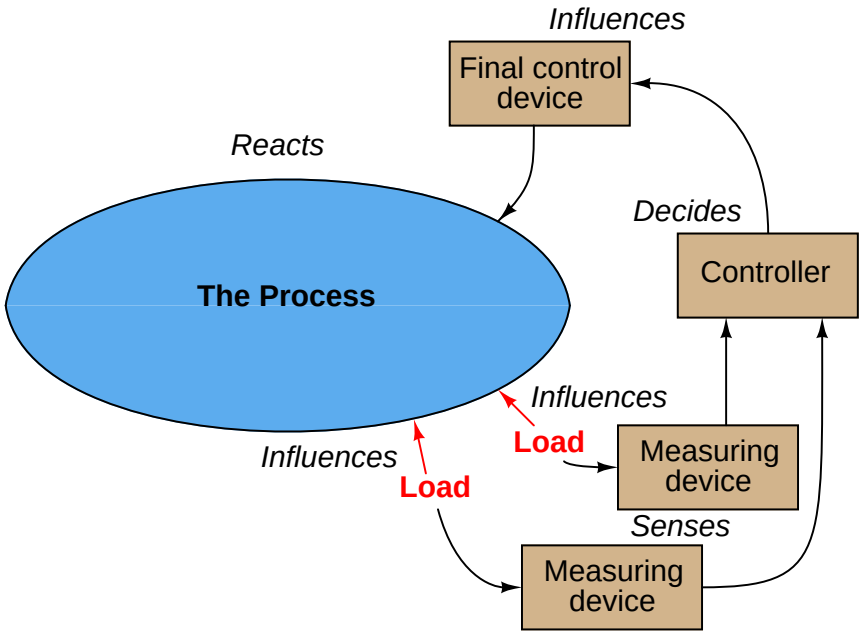
However, the presence of these "load" variables makes it necessary for the human driver of the automobile (or a *cruise control* system) to continually adjust engine power in order to maintain constant speed regardless of load variations. Using the car's measured speed as feedback, the driver (or cruise control) adjusts the accelerator pedal position as necessary based on whether or not the car's speed matches the desired "setpoint" value.

An inherent weakness of any feedback control system is that it can never be *pro-active*. The best any feedback control system can ever do is make adjustments to a process *after* some disturbance in process variable is detected. This makes deviations from setpoint inevitable, if only for short

[7] This statement is true only for self-regulating processes. Integrating and "runaway" processes require control systems to achieve stability even in the complete absence of any loads. However, since self-regulation typifies the vast majority of industrial processes, we may conclude that the fundamental purpose of most control systems is to counteract the effects of loads.

periods of time. In the context of our automobile cruise control system, this means the car cannot maintain perfectly constant speed because the control system does not have the ability to anticipate hills, wind gusts, changes in air temperature, or changes in road surface.

Feedforward control addresses this weakness by taking a fundamentally different approach, basing final control decisions on the states of load variables rather than the process variable. In other words, a feedforward control system monitors all the factors influencing a process and decides how to compensate for these factors *ahead of time* before they have the opportunity to affect the process variable. If all loads are accurately measured, and the control algorithm realistic enough to predict process response for these known load values, the process variable does not even need to be measured at all:



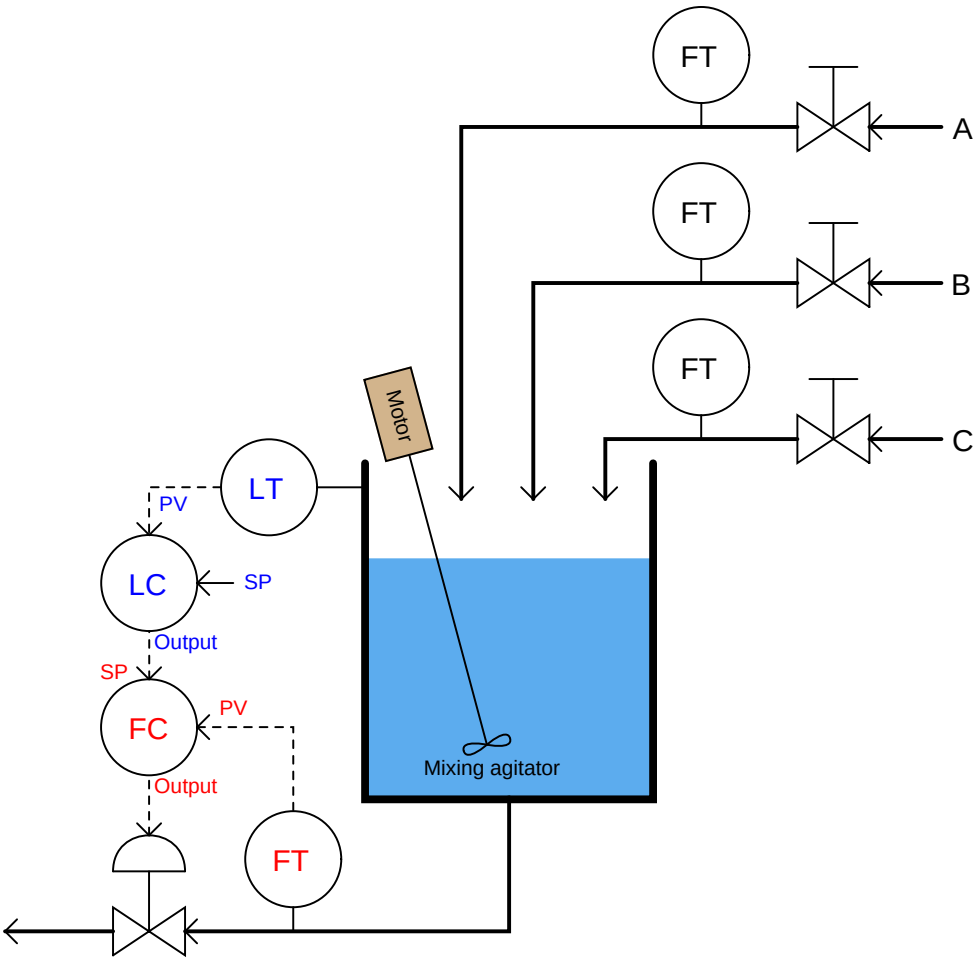
A purely feedforward automobile cruise control system would be interfaced with topographical maps, real-time weather monitoring stations, and road surface sensors to decide how much engine power is necessary at any given time to attain the desired speed[8]. Assuming all relevant load variables are accounted for, the cruise control would be able to maintain constant speed regardless of conditions, and without the need to even monitor the car's speed.

If you are feeling skeptical of this feedforward principle and its ability to control a process variable without even measuring it, this is a good thing – you are thinking critically. In practice, it is nearly impossible to accurately account for *all* loads influencing a process and to both anticipate and counter-act their combined effects, and so pure feedforward control systems are rare. Instead, the

[8]The load variables I keep mentioning that influence a car's speed is an incomplete list at best. Many other variables come into play, such as fuel quality, engine tuning, and tire pressure. In order for a purely feedforward (i.e. no feedback monitoring of the process variable) control system to work, *every single load variable* must be accurately monitored and factored into the system's output signal. This is impractical for a great many applications, which is why we usually find feedforward control used in conjunction with feedback control.

feedforward principle finds use as an augment to normal feedback control. To understand feedforward control better, we will consider its pure application before exploring how it may be combined with feedback control.

First, let us consider a liquid level control system on an open tank, where three different fluid ingredients (shown in the following P&ID simply as A, B, and C) are mixed to produce a final product. A level transmitter (LT) measures liquid level, while a level controller (LC) compares this level to a setpoint value, and outputs a signal calling for a certain amount of discharge flow. A cascaded (slave) flow controller (FC) senses outgoing flow via a flow transmitter (FT) and works to maintain whatever rate of flow is "asked" for by the level controller:

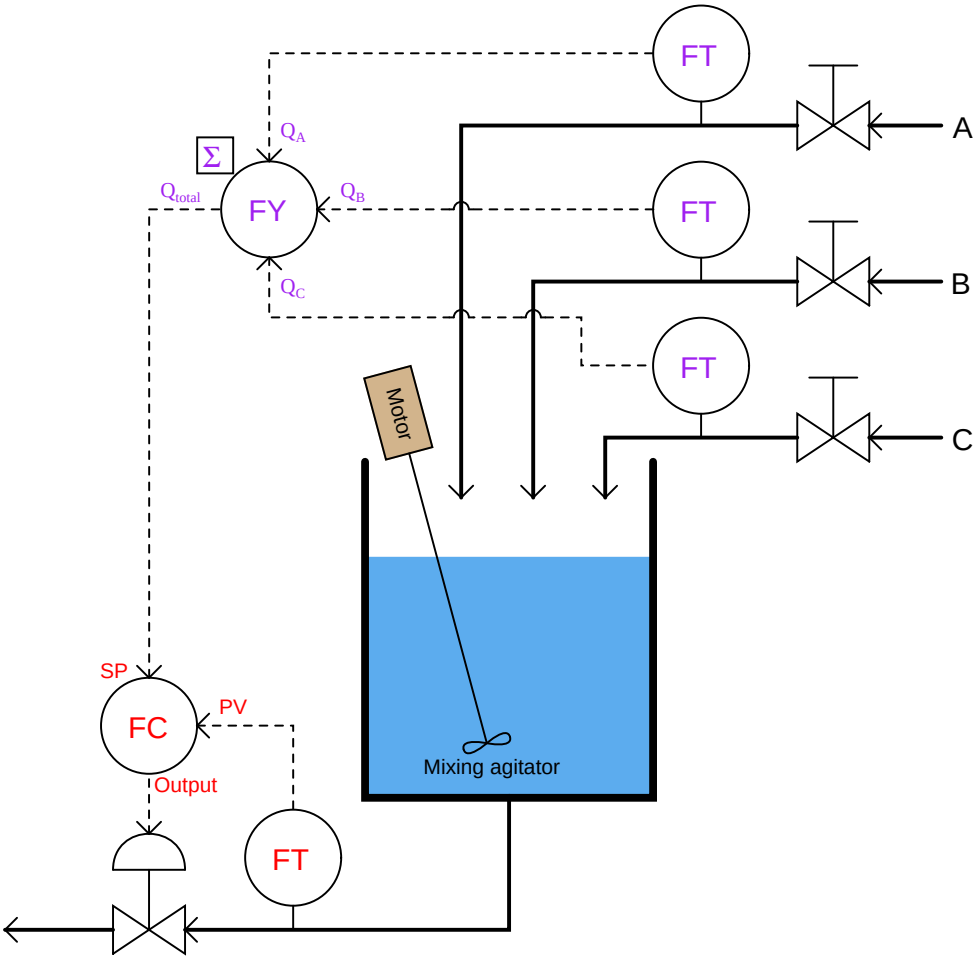


The level control system acts to keep liquid level constant in the vessel, ensuring adequate mixing of the three ingredients[9]. Being a feedback level control system, it adjusts the discharge flow rate in

[9]If the liquid level drops too low, there will be insufficient retention time in the vessel for the fluids to mix before

response to measured changes in liquid level. Like all feedback control systems, this one is *reactive* in nature: it can only take corrective action *after* a deviation between process variable (level) and setpoint is detected. As a result, temporary deviations from setpoint are guaranteed to occur with this control system every time the combined flow rate of the three ingredients increases or decreases.

Let us now change the control system strategy from feedback to feedforward. It is clear what the loads are in this process: the three ingredient flows entering the vessel. If we measure and sum these three flow rates[10], then use the total incoming flow signal as a setpoint for the discharge flow controller, the outlet flow should (ideally) match the inlet flow, resulting in a constant liquid level. Being a purely feedforward control system, there is no level transmitter (LT) any more, just flow transmitters to measure the three loads:



they exit the product line at the bottom. If liquid level is too high, the mixing action will be damped to the point where it ceases to be effective.

[10]The device or computer function performing the summation is shown in the P&ID as a bubble with "FY" as the label. The letter "F" denotes *Flow*, while the letter "Y" denotes a signal relay or transducer.

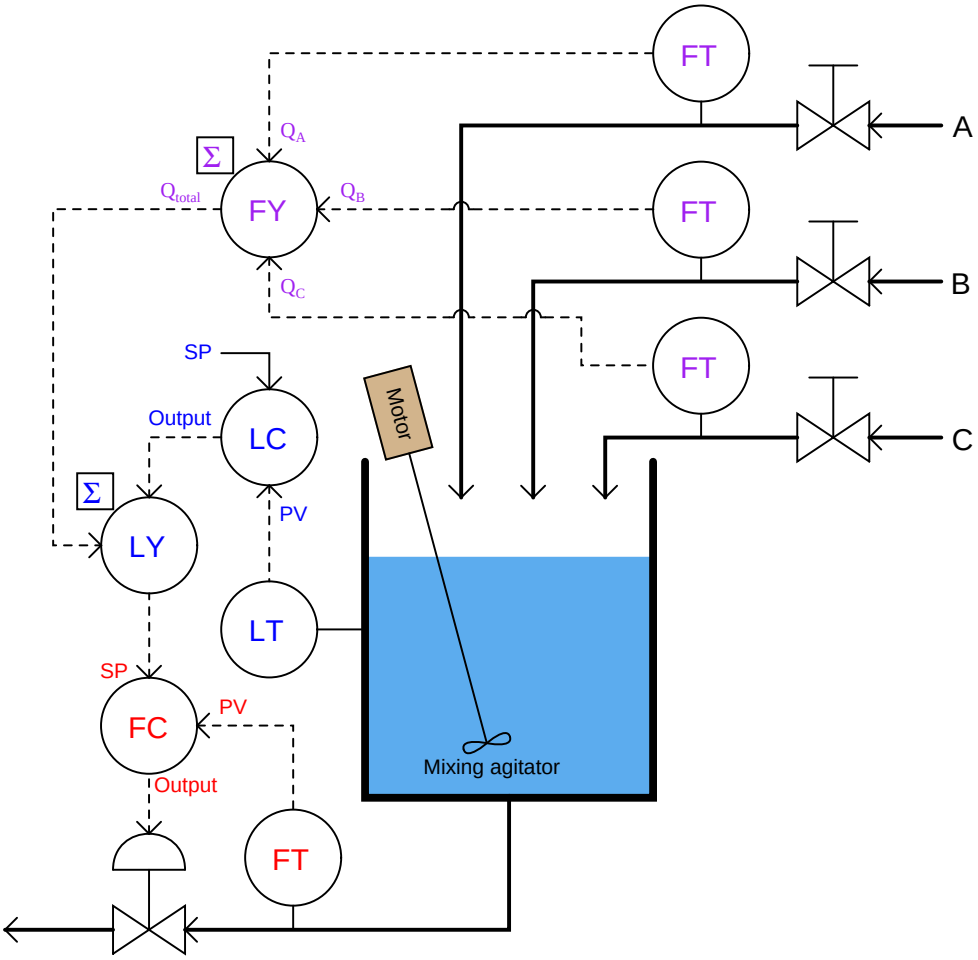
If all flow transmitter calibrations are perfect, the summing of flow rates flawless, and the flow controller's tuning robust, this level control system should control liquid level in the vessel by pro-action rather than by reaction. Any change in the flow rate of ingredients A, B, and/or C will quickly result in a matching adjustment to the discharge flow rate. So long as total volumetric flow out of the vessel is held equal to total volumetric flow into the vessel, the liquid level inside the vessel *cannot* change[11].

An interesting property of feedforward control systems is that they cannot generate oscillations as is the case with an over-tuned (excessive gain) feedback system. Since a feedforward system never monitors the effects of its actions, it will never react to something it did to the process, which is the foundation of feedback oscillation. While it is entirely possible for a feedforward control system to be configured with too much gain, the effect of this will be *overcompensation* for a load change rather than oscillation. In the case of the mixing tank feedforward level control process, improper instrument scaling and/or offsets will cause the discharge and inlet flows to mis-match, resulting in a liquid level that either continues to increase or decrease over time ("integrate"). However, no amount of mis-adjustment can cause this feedforward system to produce oscillations.

In reality, this pure feedforward control system is impractical even if all calibrations and calculations within are perfect. There are still loads unaccounted for: evaporation of liquid from the vessel, for example, or the occasional pipe fitting leak. Furthermore, since the control system has no "knowledge" of the actual liquid level, it cannot make adjustments to that level. If an operator, for instance, desired to decrease the liquid level to achieve a more vigorous mixing action, he or she would have to manually drain liquid out of the vessel, or temporarily place the discharge flow controller in "manual" mode and increase the flow there (then place back into "cascade" mode where it follows the remote setpoint signal again). The advantage of pro-active control and minimum deviation from setpoint over time comes at a fairly high price of practicality and convenience.

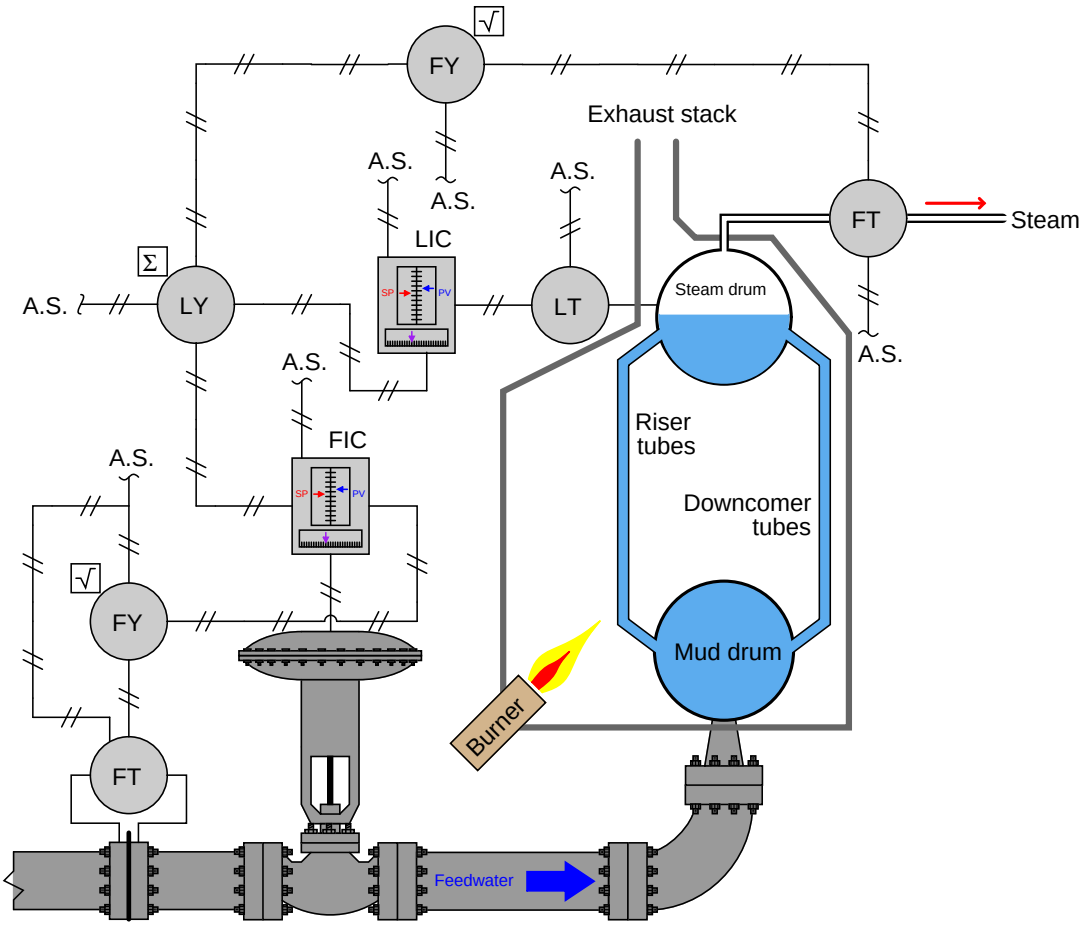
[11] Incidentally, this is a good example of an *integrating* process, where the rate of process variable change over time is proportional to the imbalance of flow rates in and out of the process. Stated another way, total accumulated (or lost) mass in a mass-balance system such as this is the time-integral of the difference between incoming and outgoing flow rates: $m = \int_0^T (Q_{in} - Q_{out})\, dt$.

For these reasons, feedforward control is most often found in conjunction with feedback control. To show how this would work in the liquid level control system, we will incorporate a level transmitter and level controller back into the system, the output of that level controller being summed with the feedforward flow signal (by the LY summing relay) before going to the cascaded setpoint input of the discharge flow controller:



This hybrid control strategy is sometimes called *feedforward with trim*. In this context, "trim" refers to the level controller's (LC) output signal contributing to the discharge flow setpoint, helping to compensate for any unaccounted loads (evaporation, leaks) and provide for level setpoint changes. This "trim" signal should do very little of the control work in this system, the bulk of the liquid level stability coming from the feedforward signals provided by the incoming flow transmitters.

A very similar control strategy commonly used on large steam boilers for the precise control of steam drum water level goes by the name of *three-element feedwater control*. The following illustration shows an example of this control strategy implemented with pneumatic (3-15 PSI signal) instruments:



Such a control system is called "three-element" because it makes use of three process measurements:

- Feedwater flow rate
- Steam drum water level
- Steam flow rate

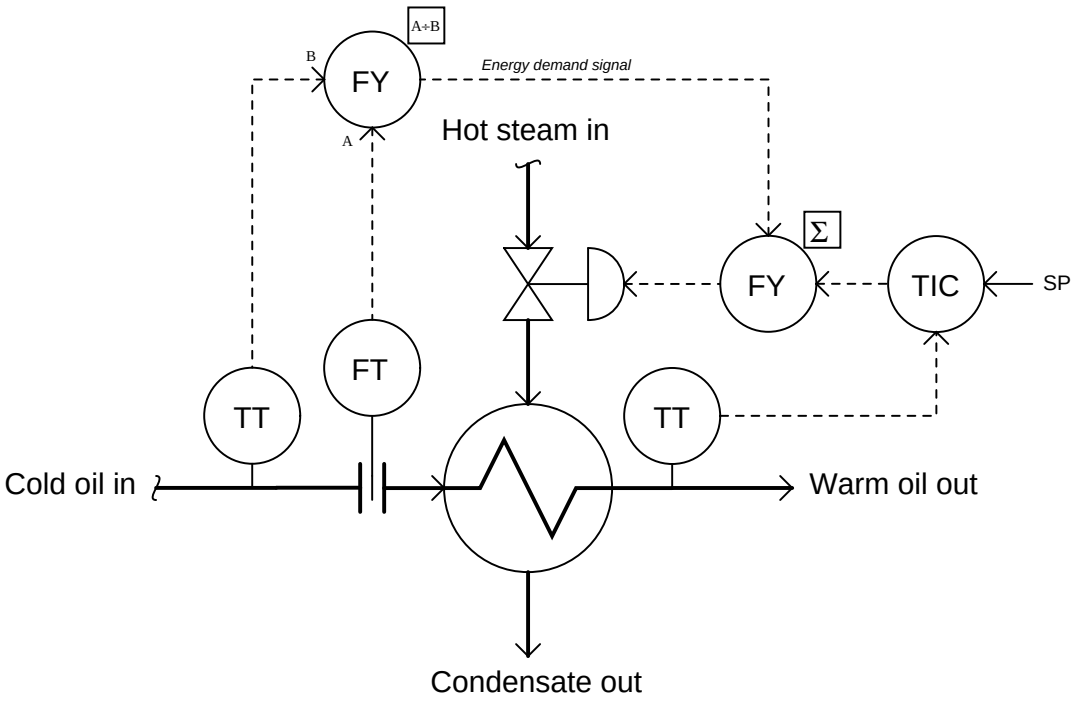
Feedwater flow is controlled by a dedicated flow controller (FIC), receiving a remote setpoint signal from a summing relay (LY). The summer receives two inputs: a steam flow signal and the

output signal (trim) from the level controller (LIC). The feedforward portion of this system (steam flow feeding forward to water flow) is intended to match the mass flow rates of water into the boiler with steam flow out of the boiler. If steam demand suddenly increases, this feedforward portion of the system immediately calls for a commensurate increase in water flow, since every molecule of steam must come from one molecule of water. The level controller and transmitter act as a feedback control loop, supplementing the feedforward signal to the cascaded water flow controller to make up for ("trim") any shortcomings of the feedforward loop.

A three-element boiler feedwater control system is a good example of a feedforward strategy designed to ensure *mass balance*, defined as a state of equality between all incoming mass flow rates and all outgoing mass flow rates. The steam flow transmitter measures outgoing mass flow, its signal being used to adjust incoming water mass rate. Since mass cannot be created or destroyed (the Law of Mass Conservation), every unit of steam mass leaving the boiler must be accounted for as an equivalent unit of water mass entering the boiler. If the control system perfectly balances these mass flow rates, water level inside the boiler *cannot* change.

In processes where the process variable is affected by energy flow rates rather than mass, the balance maintained by a feedforward control system will be *energy balance* rather than mass balance. Like mass, energy cannot be created or destroyed, but must be accounted for. A feedforward control system monitoring all incoming energy flows into a process and adjusting the outgoing energy flow rate (or visa-versa) will ensure no energy is depleted from or accumulated within the process, thus ensuring the stability of the processes' internal energy state.

An example of an energy-balance feedforward control system is found applied to the control of this heat exchanger:



The two transmitters on the incoming (cold oil) line measure oil temperature and oil flow rate, respectively. The quotient of these two variables represents the *energy demand* of the incoming oil (i.e. how much energy will be required to elevate the oil flow's temperature to some higher value). This "energy demand" signal is summed with the temperature controller's output signal to set the steam valve position (adding energy to the process).

28.6 Feedforward with dynamic compensation

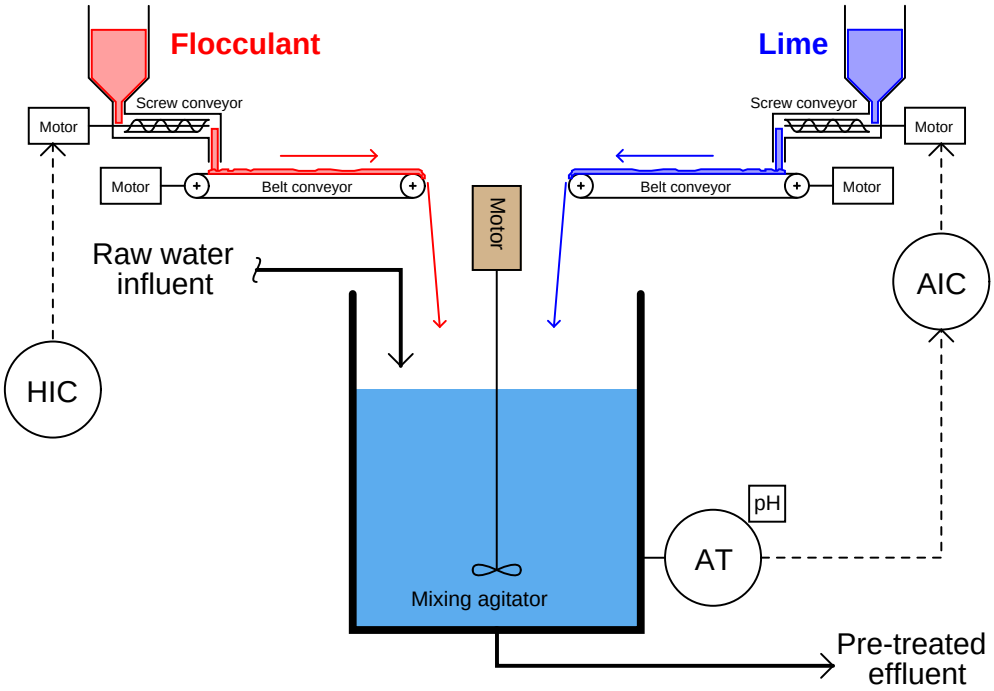
As we have seen, feedforward control is a way to improve the stability of a feedback control system in the face of changing loads. Rather than rely on feedback to make corrective changes to a process only *after* some load change has driven the process variable away from setpoint, feedforward systems monitor the relevant load(s) and use that information to preemptively make stabilizing changes to the final control element such that the process variable will not be affected. In this way, the feedback loop's role is to merely "trim" the process for factors lying outside the realm of the feedforward system.

At least, this is how feedforward control is *supposed* to work. One way feedforward controls commonly fail to live up to their promise is if the effects of load changes and of manipulated variable changes possess different time lags in their respective effects on the process variable. This is a problem in feedforward control systems because it means the corrective action called for in response to a change in load will not affect the process variable at the same time, or in the same way over time, as the load will. In order to correct this problem, we must intelligently insert time lags into the control system to equalize the time lags of load and feedforward correction. This is called *dynamic compensation*.

The following subsections will explore illustrative examples to make both the problem and the solution(s) clear.

28.6.1 Dead time compensation

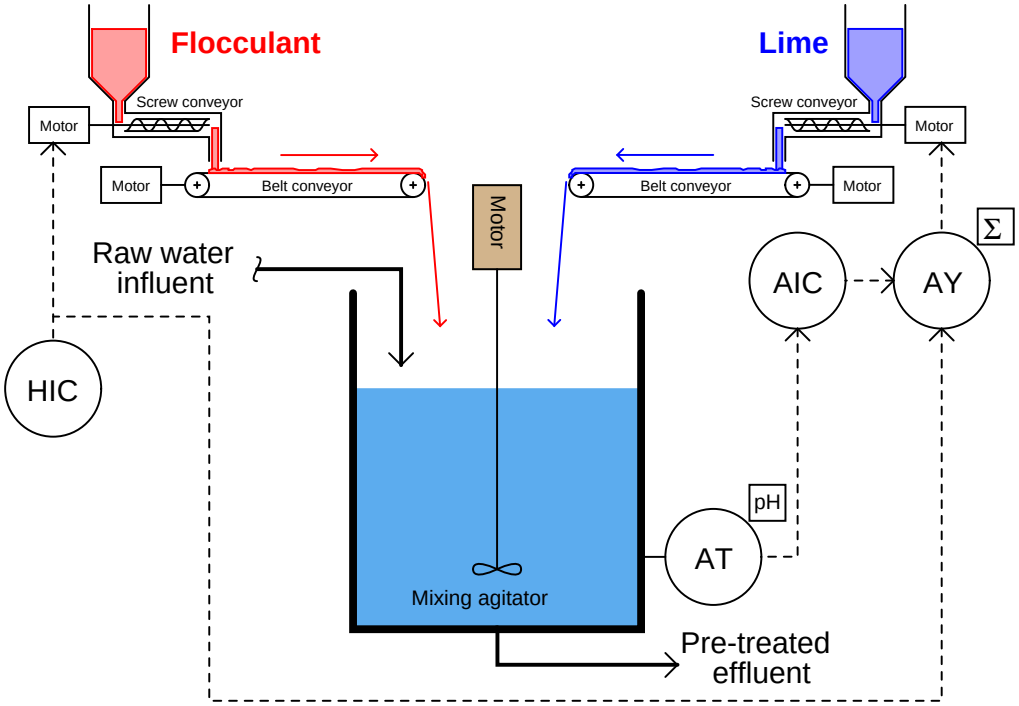
Examine the following control system P&ID showing the addition of *flocculant* (a chemical compound used in water treatment to help suspended solids clump together for easier removal by filtering and/or gravity clarification) and *lime* for pH balance. Flocculant is necessary to expedite the removal of impurities from the water, but some flocculation compounds have the unfortunate effect of decreasing the pH value of the water (turning it more acidic). If the water's pH value is too low, the flocculant ironically loses its ability to function. Thus, lime (an alkaline substance) must be added to the water to counter-act the flocculant's effect on pH to ensure efficient flocculation. Both substances are powders in this water pre-treatment system, metered by variable-speed screw conveyors and carried to the mixing tank by belt-style conveyors:



The control system shown in this P&ID consists of a pH analyzer (AT) transmitting a signal to a pH indicating controller (AIC), adjusting the speed of the lime screw conveyor. The flocculant screw conveyor speed is manually set by a *hand indicating controller* (HIC) – sometimes known as a *manual loading station* – adjusted when necessary by experienced water treatment operators who periodically monitor the effectiveness of flocculation in the system.

This simple feedback control system will work fine in steady-state conditions, but if the operator suddenly changes flocculant flow rate into the mixing vessel, there will be a temporary deviation of pH from setpoint before the pH controller is able to find the correct lime flow rate into the vessel to compensate for the change in flocculant flow. In other words, flocculant feed rate into the mixing tank is a *load* which the pH control loop must compensate for.

Dynamic response could be greatly improved with the addition of feedforward control to this system:



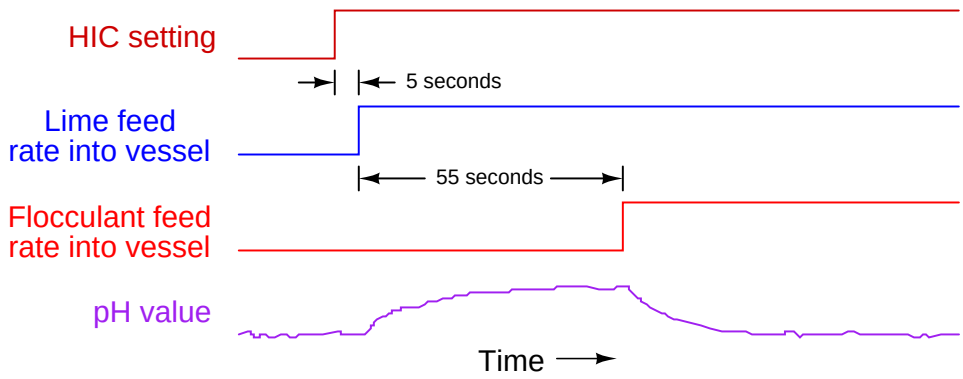
Here, the hand controller's signal goes to a signal summing relay (or function block in a digital control system) in addition to acting as a control signal to the flocculant screw conveyor motor drive. If an operator changes the flocculant feed rate, the lime feed rate will immediately adjust to compensate, *before* any change in pH value takes place in the water. Ideally, the pH controller need only make minor "trim" adjustments to lime feed rate, while the feedforward signal does most of the work in maintaining a steady pH value[12].

Even if all components in the feedforward system have been calibrated and configured properly, however, a potential problem still lurks in this system which can cause the pH value to temporarily deviate from setpoint following flocculant feed rate changes. This problem is the *transport delay* – otherwise known as *dead time* – inherent to the two belt conveyors transporting both flocculant and lime powder from their respective screw conveyors to the mixing vessel. If the rotational speed of a screw conveyor changes, the flow rate of powder exiting that screw conveyor will immediately and proportionately change. However, the belt conveyor imposes a time delay before the new powder

[12] I have deliberately omitted much of the complexity which would normally accompany a feedforward control scheme in a pH system, including gain and bias relays (function blocks), nonlinear functions, limits, and other signal-manipulation algorithms likely necessary to balance lime addition with flocculant addition. pH is a very nonlinear process to control, and as such a simple 1:1 ratio system like this would certainly yield poor results. However, the basic concept will still serve to illustrate the problem of differential time lags in a feedforward system, which I still have not directly revealed at this point!

feed rate enters the mixing vessel. In other words, the water in the vessel will not "see" the effects of a change in flocculant or lime feed rate until after this time delay has elapsed. This is not a problem if the dead times of both belt conveyors are exactly equal, since this means any compensatory change in lime feed rate initiated by the feedforward system will reach the water at exactly the same time the new flocculant rate reaches the water. So long as flocculant and lime feed rates are precisely balanced with one another at the point in time they reach the mixing vessel, pH should remain stable. But what if their arrival times are not coordinated – what will happen to pH then?

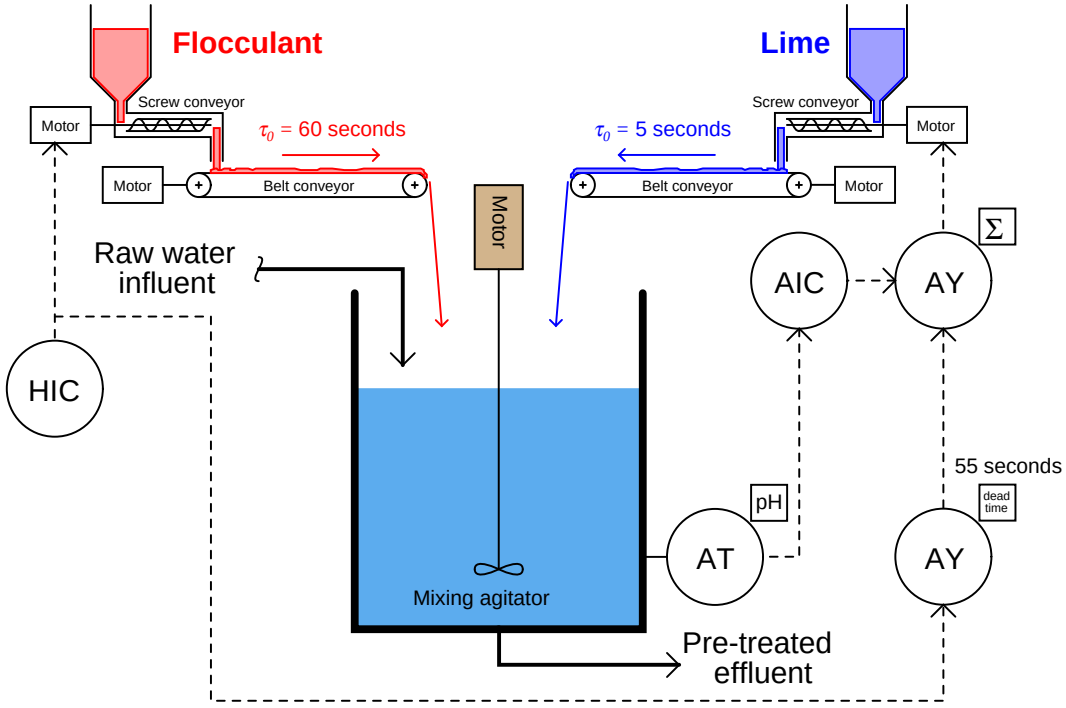
Let us engage in a "thought experiment" to explore the consequences of the flocculant conveyor belt moving much slower than, and/or being much longer than, the lime conveyor belt. Suppose the flocculant belt imposed a dead time of 60 seconds on flocculant powder making it to the vessel, while the lime belt only delayed lime powder transit by 5 seconds from screw conveyor to mixing tank. This would mean changes in flocculant flow (set by the hand controller) would compensate with changes in lime flow *55 seconds too soon*. Now imagine the human operator making a sudden increase to the flocculant powder feed rate. The lime feed rate would immediately increase thanks to the efforts of the feedforward system. However, since the increased flow rate of lime powder will reach the mixing vessel 55 seconds before the increased flow rate of flocculant powder, the effect will be a temporary increase in pH value beginning about 5 seconds after the operator's change, and then a settling of pH value back to setpoint[13], as shown in this timing diagram:



The obvious solution to this problem is to mechanically alter the belt conveyor systems for equal transport times of flocculant and lime powders. If this is impractical, we may achieve a similar result by incorporating another signal relay (or digital function block) inserting dead time into the feedforward control system. In other words, we can modify the control system in such a way to emulate what would be impractical to modify in the process itself.

[13] This "thought experiment" assumes no compensating action on the part of the feedback pH controller for the sake of simplicity. However, even if we include the pH controller's efforts, the problem does not go away. As pH rises due to the premature addition of extra lime, the controller will try to reduce the lime feed rate. This will initially reduce the degree to which pH deviates from setpoint, but then the reverse problem will occur when the increased flocculant enters the vessel 55 seconds later. Now, the pH will drop below setpoint, and the feedback controller will have to ramp up lime addition (to the amount it was before the additional lime reached the vessel) to achieve setpoint.

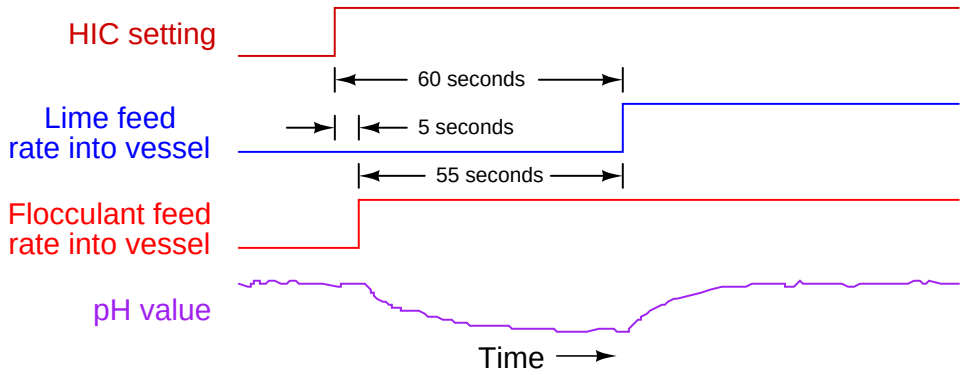
This new function will add a dead time of 55 seconds to the feedforward signal before it enters the summer, thus delaying the lime feed rate's response to feedforward effect by just the right amount of time such that any lime feed rate changes called for by feedforward action will arrive at the vessel *simultaneously* with the changed flocculant feed rate:



This solution to different time lags in a physical system, of purposely adding time lag functions to the signal path(s) of a control system, is called *dynamic compensation*. The proper selection and placement of such dynamic compensation elements in a control system is best done after all the time delays of the process are understood, which is why this scenario began with a detailed explanation of delay time between the two conveyor belts. Only after understanding time lags intrinsic to the process, and understanding the effect these differing time lags will have on the efficacy of feedforward control, can a dynamic compensation element be intelligently placed in a control system for beneficial effect.

Note how the feedback pH controller's loop was purposely spared the effects of the added dead time function, by placing the function outside of that controller's feedback loop. This is important, as dead time in any form is the bane of feedback control. The more dead time within a feedback loop, the easier that loop will tend to oscillate. By strategically placing the dead time function before the summing relay rather than after (between the summer and the lime screw conveyor motor drive), the feedback control system still achieves minimum response time while only the feedforward signal gets delayed.

Let us now consider the same flocculant and lime powder control system, this time with transport delays reversed between the two belt conveyors. If the flocculant conveyor belt is now the fast one (5 seconds dead time) and the lime belt slow (60 seconds), the effects of flocculant feed rate changes will be reversed. An increase in flocculant powder feed rate to the vessel will result in a drop in pH beginning 5 seconds after the HIC setting change, followed by a rise in pH value after the additional lime feed rate finally reaches the vessel:

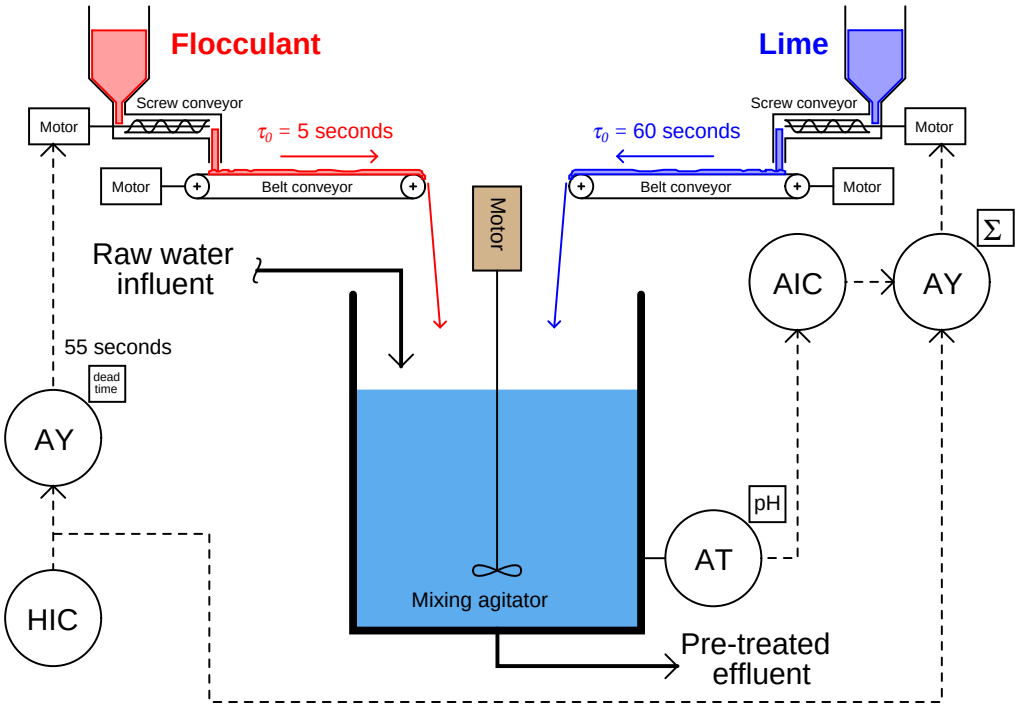


It would be possible to compensate for the difference in conveyor belt transport times using a special relay in the same location of the feedforward signal path as before, if only there was such a thing as a relay that could *predict the future exactly 55 seconds in advance!*[14]. Since no such device exists (or ever will exist), we must apply dynamic compensation elsewhere in the feedforward control system.

If a time delay is the only type of compensation function at our disposal, then the only thing we can delay in this system to make the two dead times equal is the flocculation feed rate. Thus, we should place a 55-second dead time relay in the signal path between the hand indicating controller (HIC) and the flocculant screw conveyor motor drive.

[14] Let me know if you are ever able to invent such a thing. I'll even pay your transportation costs to Stockholm, Sweden so you can collect your Nobel prize. Of course, I will demand to see the prize before buying tickets for your travel, but with your time-travel device that should not be a problem for you.

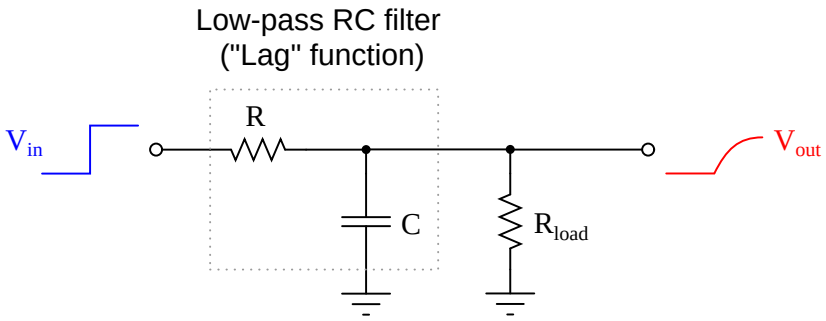
This diagram shows the proper placement of the dead time function:



With this dead time relay in place, any change in flocculation feed rate initiated by a human operator will immediately adjust the feed rate of lime powder, but delay an adjustment to flocculant powder feed rate by 55 seconds, so the two powders' feed rate changes arrive at the mixing vessel simultaneously.

28.6.2 Lag time compensation

Process time delays characterized by pure transport delay (dead time) are less common in industry than other forms of time delays, most notably *lag times*[15]. A simple "lag" time is the characteristic exhibited by a low-pass RC filter circuit, where a step-change in input voltage results in an output voltage asymptotically rising to the new voltage value over time:



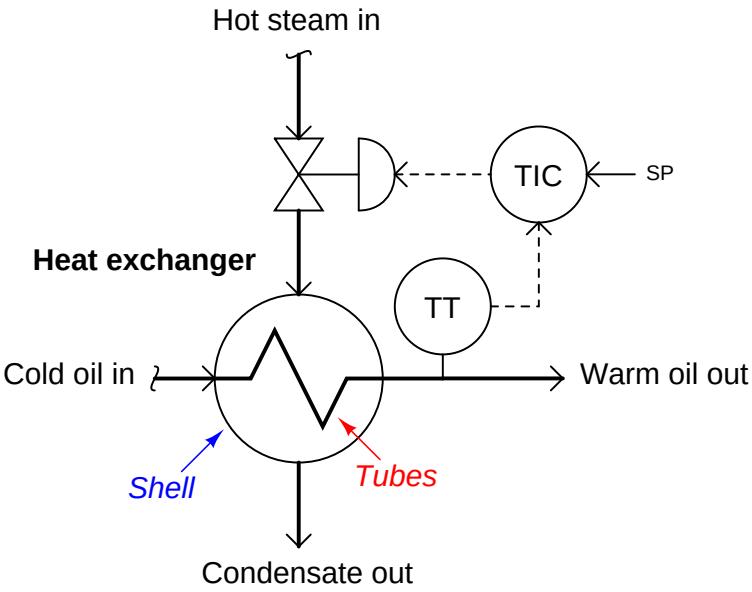
The *time constant* (τ) of such a system – be it an RC circuit or some other physical process – is the time required for the output to move 63.2% of the way to its final value ($1 - e^{-1}$). For an RC circuit such as the one shown, $\tau = RC$ (assuming $R_{load} >> R$ so the load resistance will have negligible effect on timing).

Lag times differ fundamentally from dead times. With a dead time, the effect is simply time-delayed by a finite amount from the cause, like an echo. With a lag time, the effect begins at the exact same time as the cause, but does not follow the same rapid change over time as the cause. Like dead times in a feedforward system, it is quite possible (and in fact usually the case) for loads and final control variables to have differing lag times regarding their respective effects on the process variable. This presents another form of the same problem we saw in the two-conveyor water pre-treatment system, where an attempt at feedforward control is not completely successful because the corrective feedforward action does not happen with the same amount of *lag* as the load.

To illustrate, we will analyze a heat exchanger used to pre-heat fuel oil before being sent to a combustion furnace. Hot steam is the heating fluid used to pre-heat the oil in the heat exchanger. As steam gives up its thermal energy to the oil through the walls of the heat exchanger tubes, it undergoes a phase change to liquid form (water), where it exits the shell of the exchanger as "condensate" ready to be re-boiled back into steam.

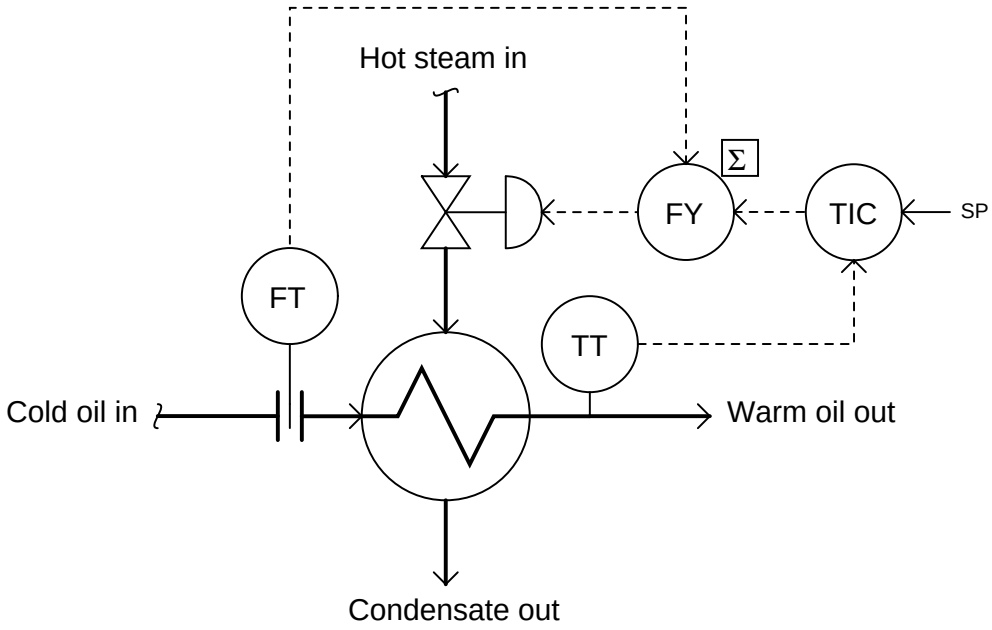
[15] For a more detailed discussion of lag times and their meaning, see section 27.1.5 on page 1525.

A simple feedback control system regulates steam flow to the heat exchanger, maintaining the discharge temperature of the oil at a constant setpoint value:



Once again, it should come as no surprise to us that the outlet temperature will suffer temporary deviations from setpoint if load conditions happen to change. The feedback control system may be able to *eventually* bring the exiting oil's temperature back to setpoint, but it cannot begin corrective action until *after* a load has driven the oil temperature off setpoint. What we need for improved control is *feedforward* action in addition to feedback action. This way, the control system can take corrective action in response to load changes *before* the process variable gets affected.

Suppose we know that the dominant load in this system is oil flow rate[16], caused by changes in demand at the combustion furnace where this oil is being used as fuel. Adapting this control system to include feedforward is as simple as installing an oil flow transmitter and a summing relay (or summing function block):



With feedforward control action in place, the steam flow rate will immediately change with oil flow rate, preemptively compensating for the increased or decreased heat demand of the oil. In other words, the feedforward system acts to maintain precise *energy balance* in the process, so heat energy never accumulates in the exchanger or bleeds away from the exchanger, causing changes in temperature.

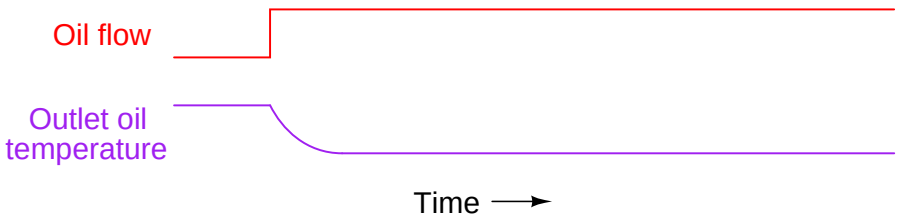
There is a problem of time delay in this system, however: a change in oil flow rate has a *faster* effect on outlet temperature than a proportional change in steam flow rate. This is due to the relative masses impacting the temperature of each fluid. The oil's temperature is primarily coupled to the temperature of the tubes, whereas the steam's temperature is coupled to both the tubes and the shell of the heat exchanger. So, the steam has a greater mass to heat than the oil has to cool, giving the steam a larger thermal time constant than the oil.

For the sake of illustration, we will assume transport delays are short enough to ignore[17], so what we are dealing with is a different in *lag* times between the oil flow's effect on temperature and the steam flow's effect on temperature.

[16] Knowing this allows us to avoid measuring the incoming cold oil temperature and just measure incoming cold oil flow rate as the feedforward variable. If the incoming oil's temperature were known to vary substantially over time, we would be forced to measure it as well as flow rate, combining the two variables together to calculate the *energy demand* and use this abstract variable as the feedforward variable.

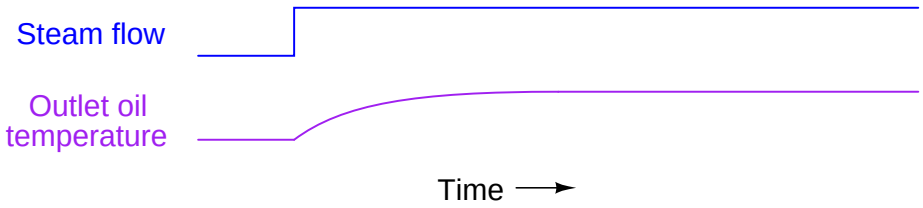
[17] Transport delay (dead time) in heat exchanger systems can be a thorny problem to overcome, as they they tend to change with flow rate! For reasons of simplicity in our illustration, we will treat this process as if it only possessed lag times, not dead times.

Here is what would happen to the heated oil temperature if steam flow were held constant and oil flow were suddenly increased:



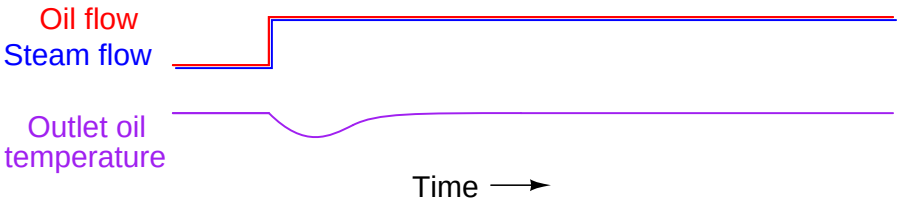
Increased oil flow convects heat away from the steam at a faster rate than before, resulting in decreased oil temperature. This drop in temperature is fairly quick, and is self-regulating.

For contrast, here is what would happen to the heated oil temperature if oil flow were held constant and steam flow were suddenly increased:



Increased steam flow convects heat into the oil at a faster rate than before, resulting in increased oil temperature. This rise in temperature is also self-regulating, but much slower than the temperature change resulting from a proportional adjustment in oil flow. In other words, the time constant (τ) of the process with regard to steam flow changes is greater than the time constant of the process with regard to oil flow changes ($\tau_{steam} > \tau_{oil}$).

If we superimpose these two effects, as will be the case when the feedforward system is working (without the benefit of feedback "trim" control), what we will see when oil flow suddenly increases is a "fight" between the cooling effect of the increased oil flow and the heating effect of the increased steam flow. However, it will not be a fair fight initially: the oil flow's effect will temporarily win over the steam's effect because of the oil's faster time constant. The result will be a momentary dip in outlet temperature before the system achieves equilibrium again:

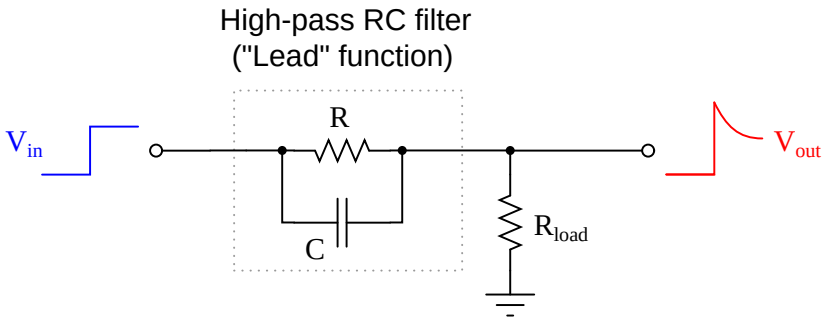


The solution to this problem is not unlike the solution we applied to the water treatment system: we must somehow make these two time delays more equal so their superimposed effects will directly cancel, resulting in an undisturbed process variable. An approximate solution for equalizing two

different lag times is to cascade two lags together in order to emulate one larger lag time[18]. This may be done by inserting a lag time relay or function block in the feedforward system.

When we look at our P&ID, though, a problem is immediately evident. The lag time we need to slow down is the lag time of the oil flow's effect on temperature. In this system, oil flow is a wild variable, not something we have the ability to control. Our feedforward control system can only manipulate the steam valve position in response to oil flow, not influence oil flow in order to give the steam time to "catch up."

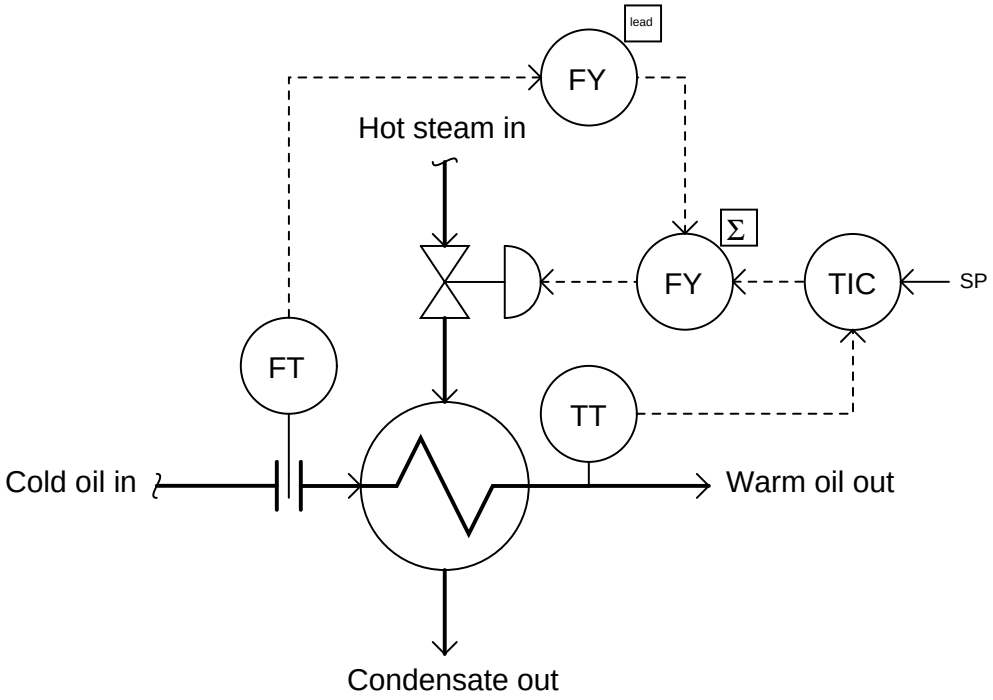
If we cannot slow down the time constant inherent to the wild variable (oil flow), then the best we can do is speed up the time constant of the variable we do have influence over (steam flow). The solution is to insert something called a *lead function* into the feedforward signal driving the steam valve. A "lead" is the mathematical inverse of a lag. If a lag is modeled by an RC low-pass filter circuit, then a "lead" is modeled by an RC high-pass filter circuit:



Being mathematical inverses of each other, a lead function should perfectly cancel a lag function when the output of one is fed to the input of the other, and when the time constants of each are equal. If the time constants of lead and lag are not equal, their cascaded effect will be a partial cancellation. In our heat exchanger control application, this is what we need to do: partially cancel the steam valve's slow time constant so it will be more equal with the oil flow's time constant. Therefore, we need to insert a lead function into the feedforward signal path.

[18]Technically, two cascaded lag times is not the same as one large lag time, not matter what the time constant values. Two first-order lags in series with one another create a *second-order lag*, which is a different effect. However imperfect as the added lag solution is, it is still better than nothing at all!

A lead function will take the form of either a physical signal relay or (more likely with modern technology) a function block executed inside a digital control system. The proper place for the lead function is between the oil flow transmitter and the summation function:



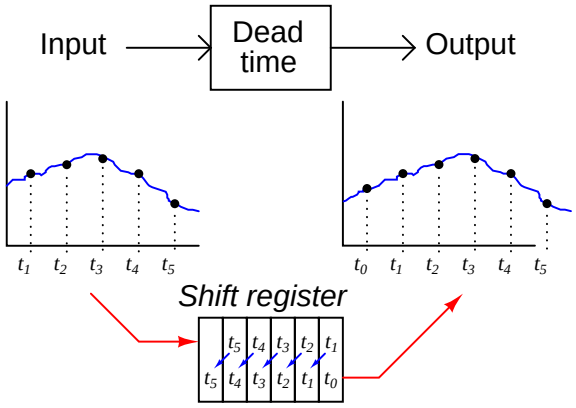
Now, when the oil flow rate to this heat exchanger suddenly changes, the lead function will add a "surge" to the feedforward signal before it goes to the summing function, quickly opening the steam valve further than usual and sending a surge of steam to the exchanger to help overcome the naturally sluggish response of the oil temperature to changes in steam flow. The dynamic compensation will not be perfect, but it will be substantially better than no dynamic compensation at all.

28.6.3 Lead/Lag and dead time function blocks

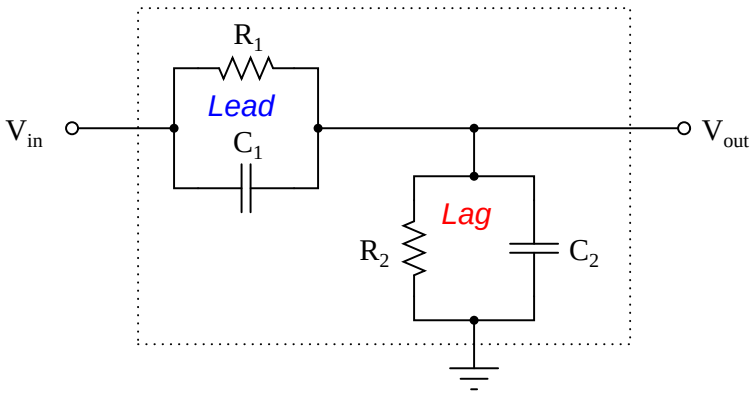
The addition of dynamic compensation in a feedforward control system may require a lag function, a lead function, and/or a dead time function, depending on the nature of the time delay differences between the relevant process load and the system's corrective action. Modern control systems provide all these functions as digital *function blocks*. In the past, these functions could only be implemented in the form of individual instruments with these time characteristics, called *relays*. As we have already seen, lead and lag functions may be rather easily implemented as simple RC filter circuits. Pneumatic equivalents also exist, which were the only practical solution in the days of pneumatic transmitters and controllers. Dead time is notoriously difficult to emulate using analog components of any kind, and so it was common to use lag-time elements (sometimes more than one connected in series) to provide an approximation of dead time.

With digital computer technology, all these dynamic compensation functions are easy to implement and readily available in a control system. Some single-loop controllers even have these capabilities programmed within, ready to use when needed.

A dead time function block is most easily implemented using the concept of a *first-in, first-out shift register*, sometimes called a *FIFO*. With this concept, successive values of the input variable are stored in a series of registers (memory), their progression to the output delayed by a certain amount of time:

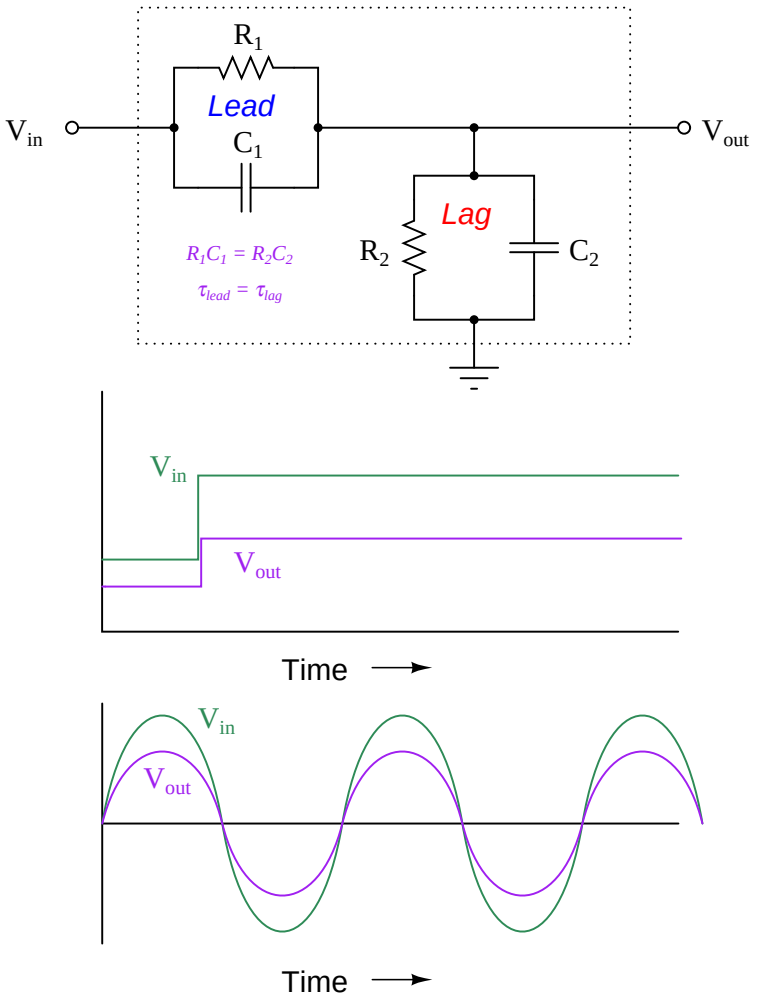


Lead and lag functions are also implemented digitally in modern controllers and control systems, but they are actually easier to comprehend in their analog (RC circuit) forms. The most common way lead and lag functions are found in modern control systems is in combination as the so-called *lead/lag function*, merging both lead and lag characteristics in a single function block (or relay):



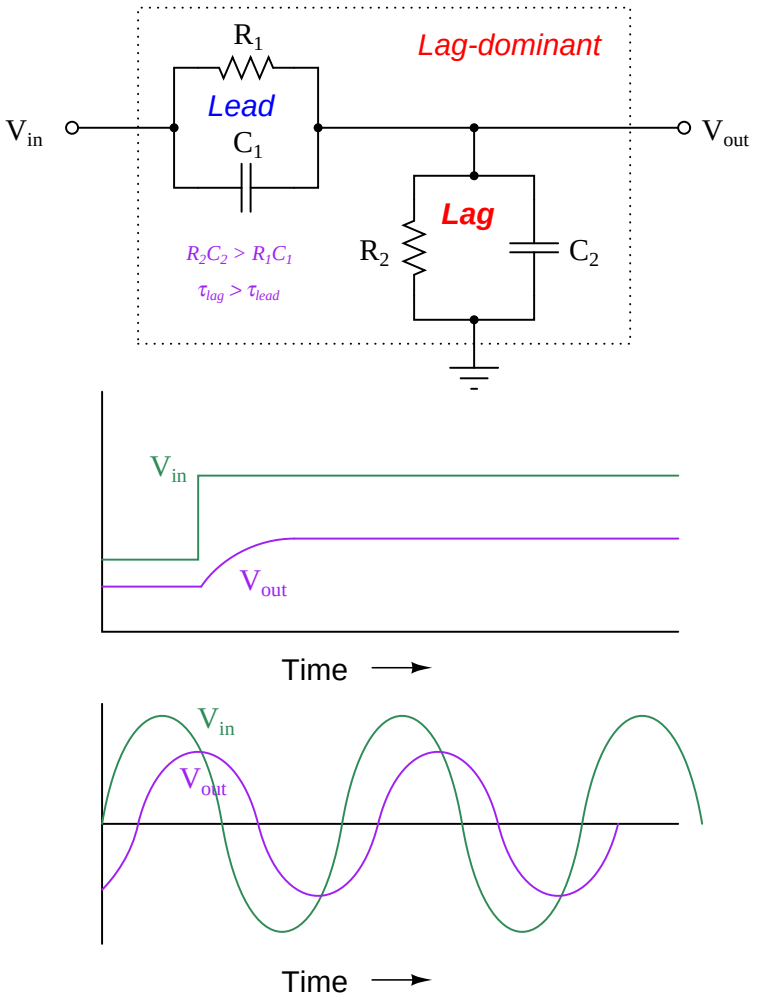
Each parallel RC subcircuit represents a time constant (τ), one for lead and one for lag. The overall behavior of the network is determined by the relative magnitudes of these two time constants. Which ever time constant is larger, determines the overall characteristic of the network.

If the two time constant values are equal to each other ($\tau_{lead} = \tau_{lag}$), then the circuit performs no dynamic compensation at all, simply passing the input signal to the output with no change except for some attenuation:



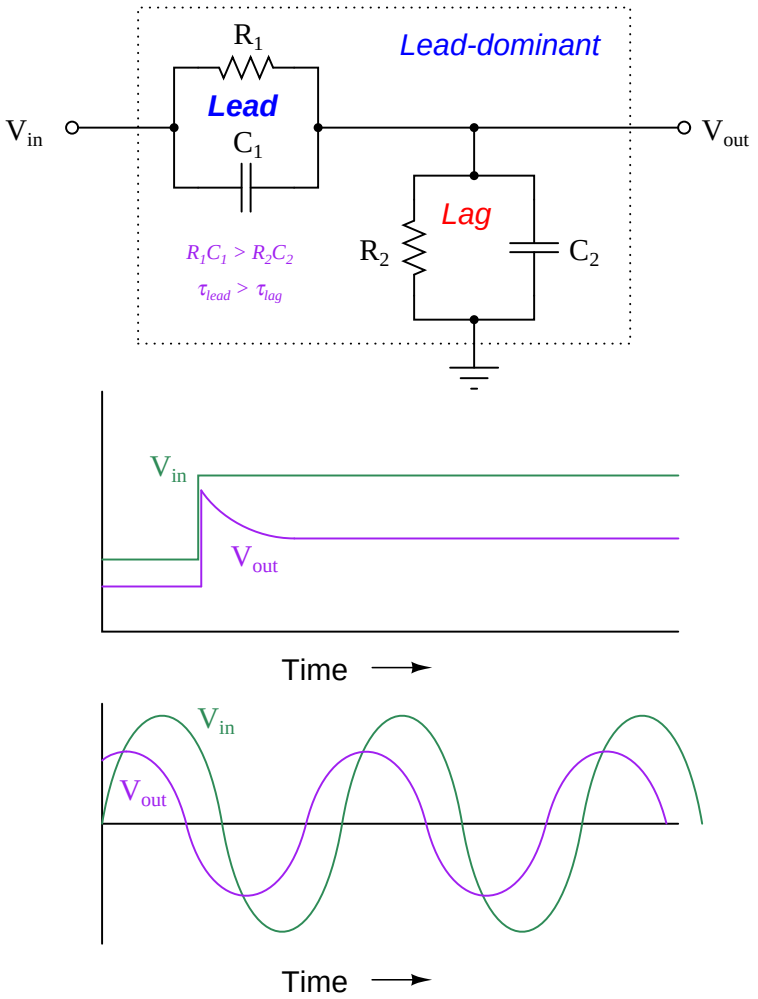
A square wave signal entering this network will exit the network as a square wave. If the input signal is sinusoidal, the output will also be sinusoidal and in-phase with the input.

If the lag time constant exceeds the lead time constant ($\tau_{lag} > \tau_{lead}$), then the overall behavior of the circuit will be to introduce a first-order lag to the voltage signal:



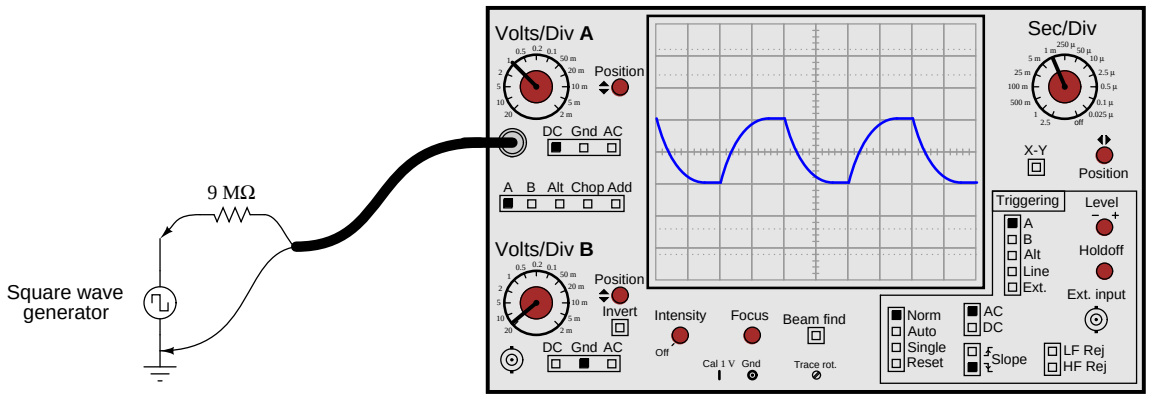
A square wave signal entering the network will exit the network as a sawtooth-shaped wave. A sinusoidal input will emerge sinusoidal, but with a lagging phase shift. This, in fact, is where the *lag* function gets its name: from the negative phase shift it imparts to a sinusoidal input.

Conversely, if the lead time constant exceeds the lag time constant ($\tau_{lead} > \tau_{lag}$), then the overall behavior of the circuit will be to introduce a first-order lead to the voltage signal (a step-change voltage input will cause the output to "spike" and then settle to a steady-state value):

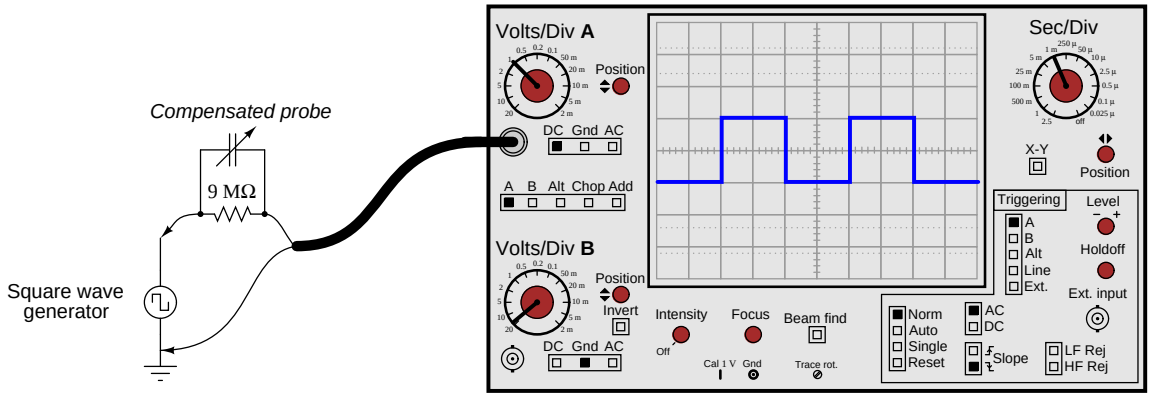


A square wave signal entering the network will exit the network with sharp transients on each leading edge. A sinusoidal input will emerge sinusoidal, but with a leading phase shift. Not surprisingly, this is where the *lead* function gets its name: from the positive phase shift it imparts to a sinusoidal input.

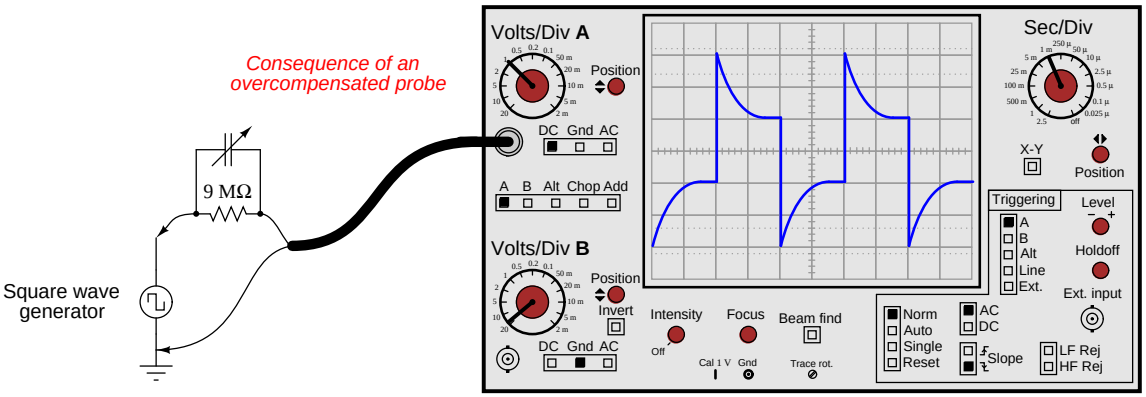
This exact form of lead/lag circuit finds application in a context far removed from process control: compensation for coaxial cable capacitance in a $\times 10$ oscilloscope probe. If a 9 MΩ resistor is connected in series with a standard oscilloscope input (having an input impedance of 1 MΩ) to create a 10:1 voltage division ratio, problems will result from the cable capacitance connecting the probe to the oscilloscope input. What should display as a square-wave input instead looks "rounded" by the effect of capacitance in the coaxial cable and at the oscilloscope input:



A simple solution to this problem is to build the 10:1 probe with a variable capacitor connected in parallel across the 9 MΩ resistor. The combination of the 9 MΩ resistor and this capacitor creates a lead network to cancel out the effects of the lag caused by the cable capacitance and 1 MΩ oscilloscope impedance in parallel. When the capacitor is properly adjusted, the oscilloscope will accurately show the shape of any waveform at the probe tip, including square waves:



If the compensation capacitor is adjusted to an excessive value, however, the probe will *overcompensate* for lag (too much lead), resulting in a "spiked" waveform on the oscilloscope display with a perfect square-wave input. While undesirable in the context of oscilloscope probes, this is precisely the effect we want in a *lead* function:



One of the design challenges for analog lead/lag networks was how to built them in such a way that they would not attenuate in the steady-state condition, since any reduction in signal strength would interfere with the proper proportioning of the feedforward signal, effectively altering the gain of the feedforward action and worse yet introducing an offset to the signal if the signal had a live-zero base such as 3-15 PSI or 1-5 volts. This is not a problem in digital lead/lag algorithms, where the lead/lag function is implemented using equations evaluated by a microprocessor.

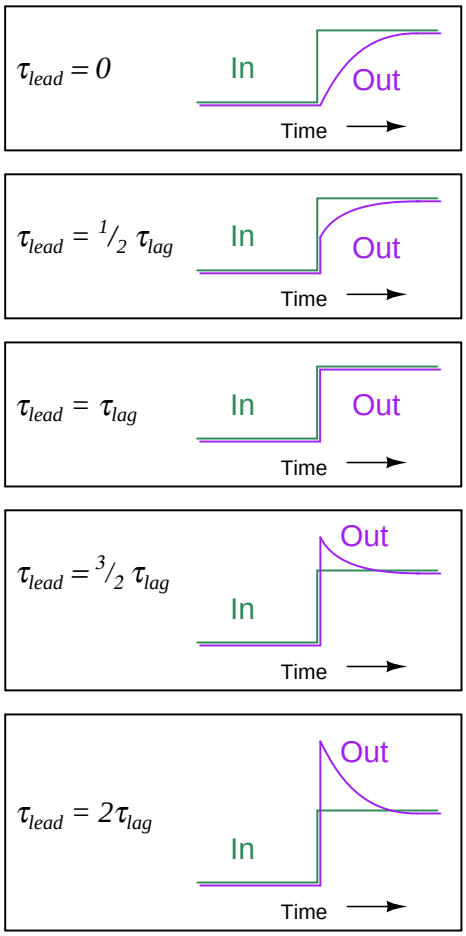
For example, a typical time-domain equation describing a digital lead/lag function block's output response (y) to an input step-change from zero (0) to magnitude x over time (t) is as follows:

$$y = x\left(1 + \frac{\tau_{lead} - \tau_{lag}}{\tau_{lag}\, e^{\frac{t}{\tau_{lag}}}}\right)$$

As you can see, if the two time constants are set equal to each other ($\tau_{lead} = \tau_{lag}$), the second term inside the parentheses will have a value of zero at all times, simplifying the equation and making y equal to x at all times. If the lead time constant exceeds the lag time constant ($\tau_{lead} > \tau_{lag}$), then the fraction will begin with a positive value and decay to zero over time, giving us the "spike" response we expect from a lead function. Conversely, if the lag time constant exceeds the lead ($\tau_{lag} > \tau_{lead}$), the fraction begin with a negative value at time = 0 (the beginning of the step-change) and decay to zero over time, giving us the "sawtooth" response we expect from a lag function.

From both an examination of the analog lead/lag networks and from this equation we can tell that the proper configuration of a lead/lag function requires *two* time constant values be set. The rate of decay for the lead/lag function (i.e. how quickly it settles to a steady-state condition after a step-change input) is primarily determined by the lag time constant (τ_{lag}), while the gain of the function (i.e. how severely the output will either "spike" or "retard" following a step-change input) is determined by the difference between the two time constants ($\frac{\tau_{lead}-\tau_{lag}}{\tau_{lag}}$). To summarize with several examples:

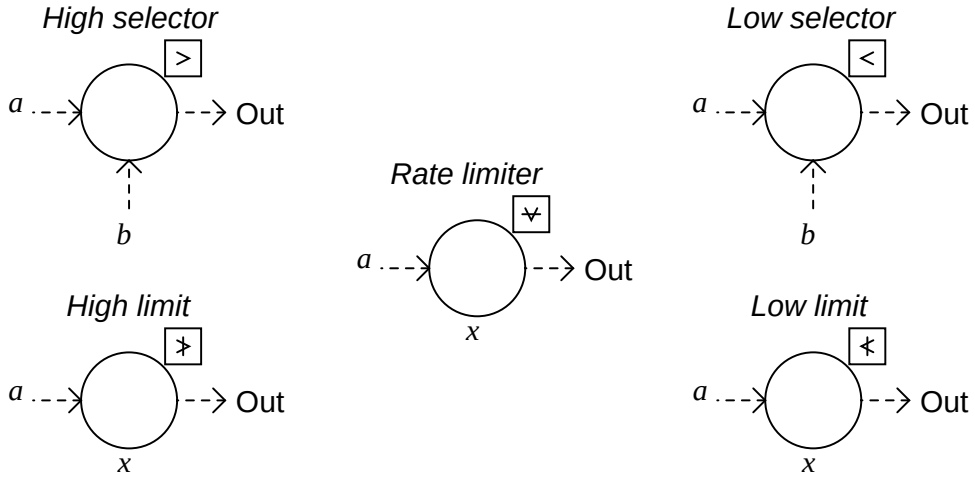
- If $\tau_{lead} = \tau_{lag}$, it will simply pass the input signal through to the output (no lead or lag action at all)
- If $\tau_{lead} = 0$, it will provide a lag function with a gain of unity and a time constant of τ_{lag}
- If $\tau_{lead} = 2(\tau_{lag})$, it will provide a lead function with a gain of unity and a time constant of τ_{lag}



28.7 Limit, Selector, and Override controls

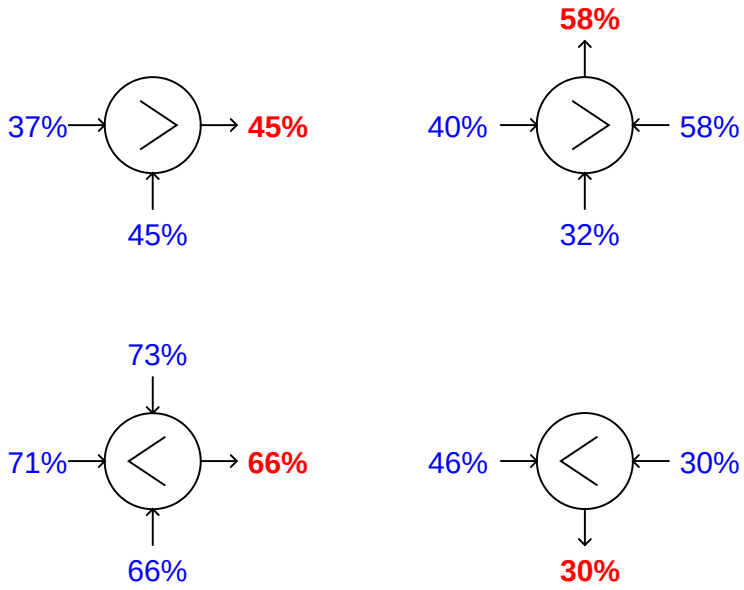
Another category of control strategies involves the use of signal relays or function blocks with the ability to switch between different signal values, or re-direct signals to new pathways. Such functions are useful when we need a control system to choose between multiple signals of differing value in order to make the best control decisions.

The "building blocks" of such control strategies are special relays (or function blocks in a digital control system) shown here:



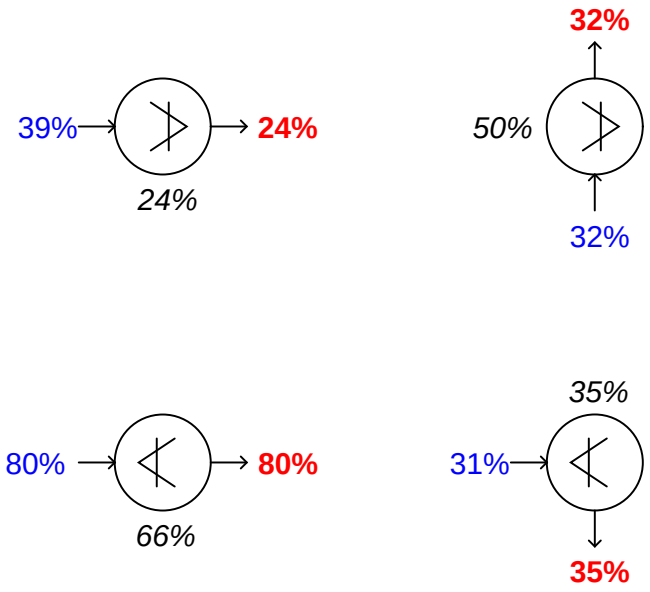
High-select functions output whichever input signal has the *greatest* value. *Low-select* functions do just the opposite: output whichever input signal has the *least* value. "Greater-than" and "Less than" symbols mark these two selector functions, respectively, and each type may be equipped to receive more than two input signals.

Sometimes you will see these relays represented in P&IDs simply by an inequality sign in the middle of the large bubble, rather than off to the side in a square. You should bear in mind that the location of the input lines has no relationship at all to the direction of the inequality symbol – e.g., it is not as though a high-select relay looks for the input on the left side to be greater than the input on the right. Note the examples shown below, complete with sample signal values:



High-limit and *low-limit* functions are similar to high- and low-select functions, but they only receive one input each, and the limit value is a parameter programmed into the function rather than received from another source. The purpose of these functions is to place a set limit on how high or how low a signal value is allowed to go before being passed on to another portion of the control system. If the signal value lies within the limit imposed by the function, the input signal value is simply passed on to the output with no modification.

Like the select functions, limit functions may appear in diagrams with nothing more than the limit symbol inside the bubble, rather than being drawn in a box off to the side:

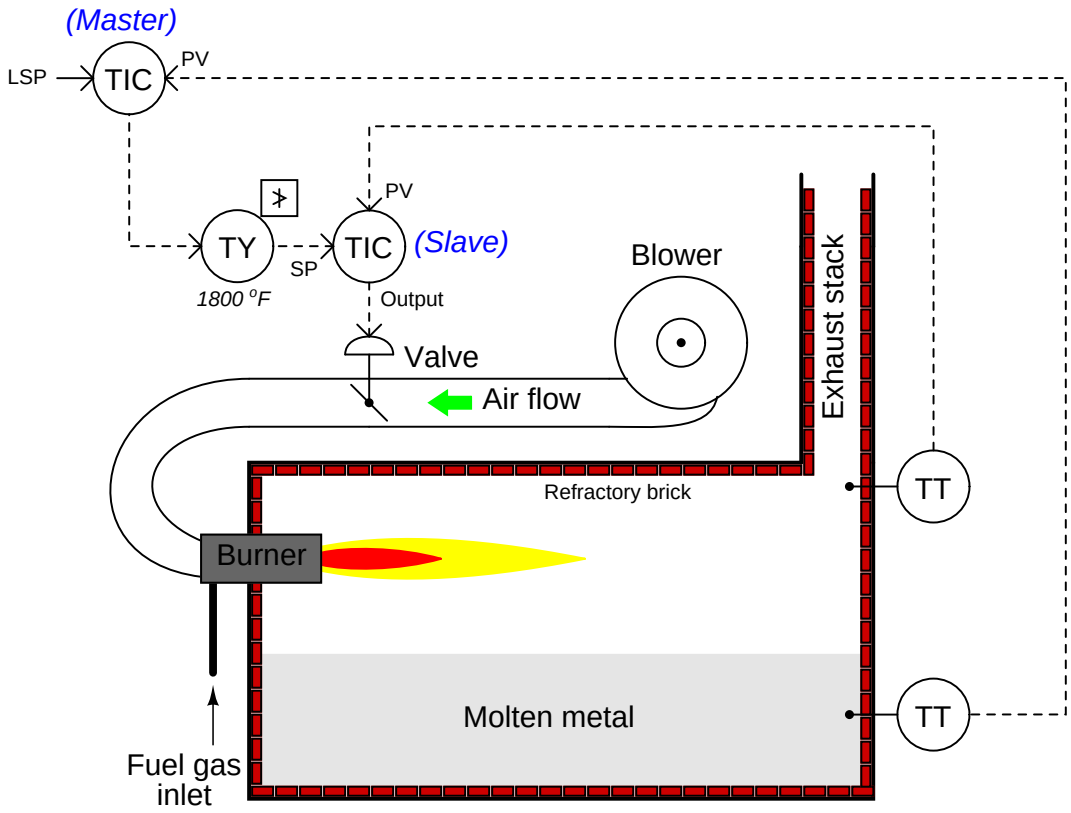


Rate limit functions place a maximum rate-of-change limit on the input signal, such that the output signal will follow the input signal precisely until and unless the input signal's rate-of-change over time ($\frac{dx}{dt}$) exceeds the pre-configured limit value. In that case, the relay still produces a ramping output value, but the rate of that ramp remains fixed at the limit $\frac{dx}{dt}$ value no matter how fast the input keeps changing. After the output value "catches up" with the input value, the function once again will output a value precisely matching the input unless the input begins to rise or fall at too fast a rate again.

28.7.1 Limit controls

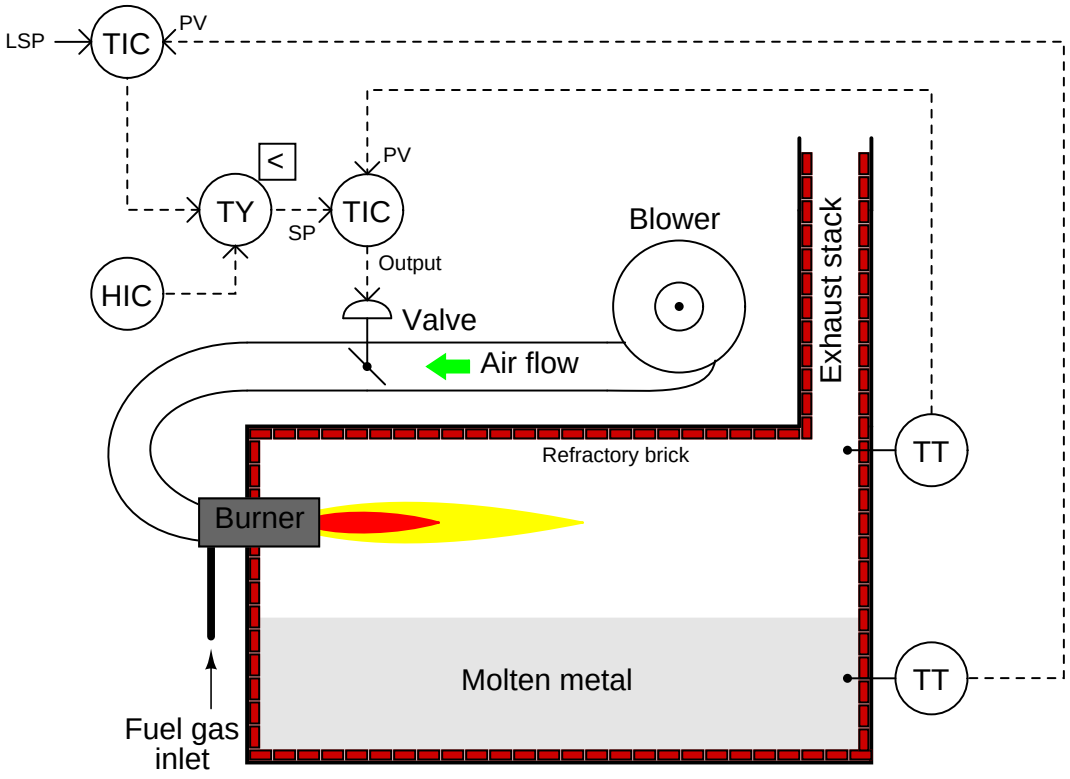
A common application for select and limit functions is in *cascade* control strategies, where the output of one controller becomes the setpoint for another. It is entirely possible for the primary (master) controller to call for a setpoint that is unreasonable or unsafe for the secondary (slave) to attain. If this possibility exists, it is wise to place a limit function between the two controllers to limit the cascaded setpoint signal.

In the following example, a cascade control system regulates the temperature of molten metal in a furnace, the output of the master (metal temperature) controller becoming the setpoint of the slave (air temperature) controller. A high limit function limits the maximum value this cascaded setpoint can attain, thereby protecting the refractory brick of the furnace from being exposed to excessive air temperatures:



It should be noted that although the different functions are drawn as separate bubbles in the P&ID, it is possible for multiple functions to exist within one physical control device. In this example, it is possible to find a controller able to perform the functions of both PID control blocks (master and slave) and the high limit function as well. It is also possible to use a distributed technology such as FOUNDATION Fieldbus to place all control functions inside field instruments, so only three field instruments exist in the loop: the air temperature transmitter, the metal temperature transmitter, and the control valve (with a Fieldbus positioner).

This same control strategy could have been implemented using a low select function block rather than a high limit:



Here, the low-select function selects whichever signal value is lesser: the setpoint value sent by the master temperature controller, or the maximum air temperature limit value sent by the hand indicating controller (HIC – sometimes referred to as a *manual loading station*).

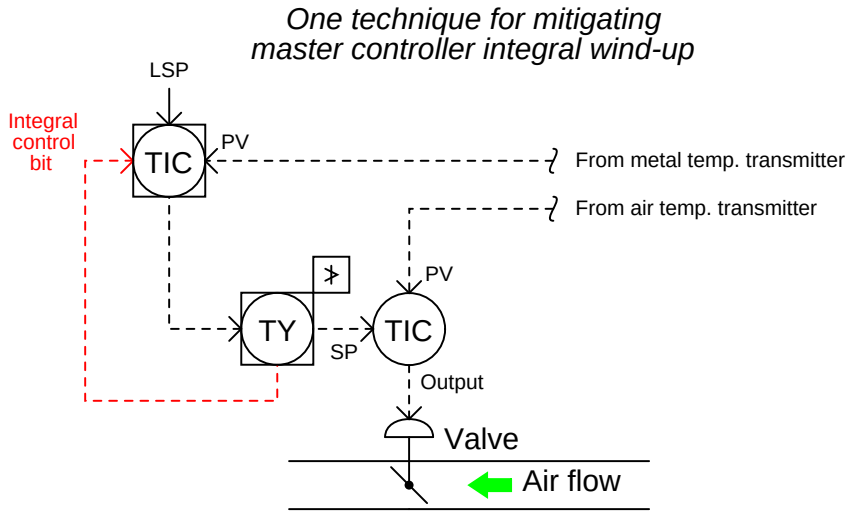
An advantage of this latter approach over the former might be ease of limit value changes. With a pre-configured limit value residing in a high-limit function, it might be that only qualified maintenance people have access to changing that value. If the decision of the operations department is to have the air temperature limit value easily adjusted by anyone, the latter control strategy's use of a manual loading station would be better suited[19].

Another detail to note in this system is the possibility of *integral windup* in the master controller in the event that the high setpoint limit takes effect. Once the high-limit (or low-select) function secures the slave controller's remote setpoint at a fixed value, the master controller's output is no

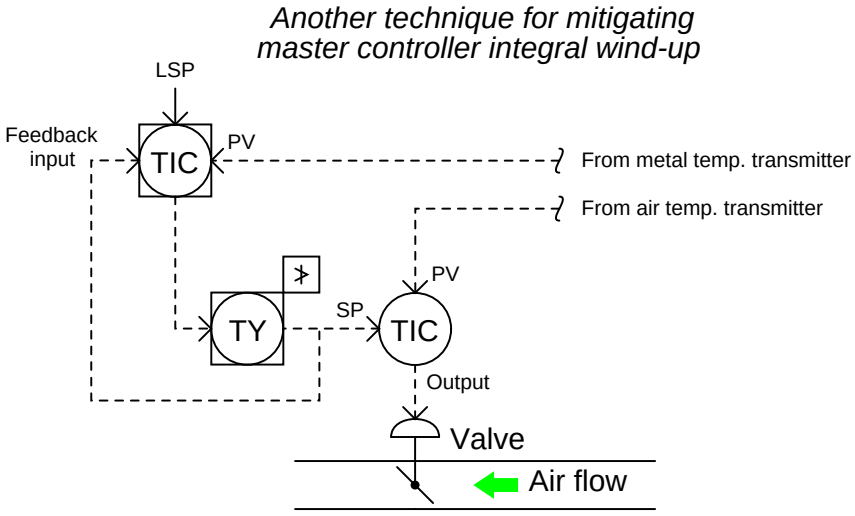
[19] I generally suggest keeping such limit values inaccessible to low-level operations personnel. This is especially true in cases such as this where the presence of a high temperature setpoint limit is intended for the longevity of the equipment. There is a strong tendency in manufacturing environments to "push the limits" of production beyond values considered safe or expedient by the engineers who designed the equipment. Limits are there for a reason, and should not be altered except by people with full understanding of and full responsibility over the consequences!

longer controlling anything: it has become decoupled from the process. If, when in this state of affairs, the metal temperature is still below setpoint, the master controller's integral action will "wind up" the output value over time with absolutely no effect, since the slave controller is no longer following its output signal. If and when the metal temperature reaches setpoint, the master controller's output will likely be saturated at 100% due to the time it spent winding up. This will cause the metal temperature to overshoot setpoint, as a positive error will be required for the master controller's integral action to wind back down from saturation.

A relatively easy solution to this problem is to configure the master controller to stop integral action when the high limit relay engages. This is easiest to do if the master PID and high limit functions both reside in the same physical controller. Many digital limit function blocks generate a bit representing the state of that block (whether it is passing the input signal to the output or limiting the signal at the pre-configured value), and some PID function blocks have a boolean input used to disable integral action. If this is the case with the function blocks comprising the high-limit control strategy, it may be implemented like this:

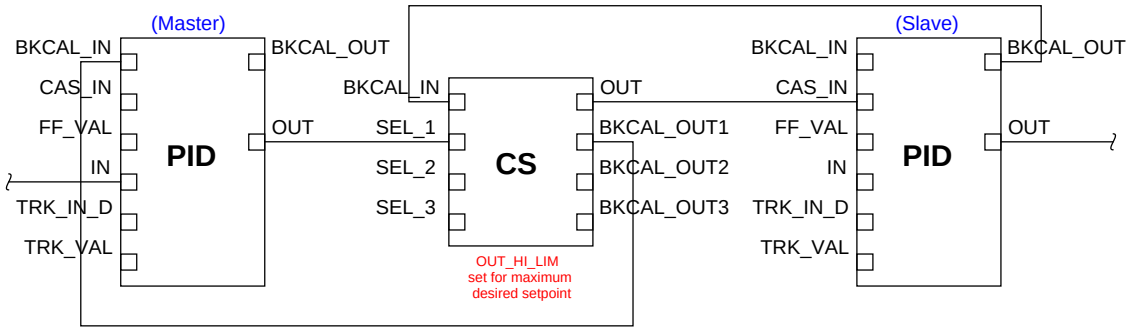


Another method used to prevent integral windup is to make use of the *feedback* input available on some PID function blocks. This is an input used to calculate the integral term of the PID equation. In the days of pneumatic PID controllers, this option used to be called *external reset*. Normally connected to the output of the PID block, if connected to the output of the high-limit function it will let the controller know whether or not any attempt to wind up the output is having an effect. If the output has been de-selected by the high-limit block, integral windup will cease:



Limit control strategies implemented in FOUNDATION Fieldbus instruments use the same principle, except that the concept of a "feedback" signal sending information backwards up the function block chain is an aggressively-applied design philosophy throughout the FOUNDATION Fieldbus standard. Nearly every function block in the Fieldbus suite provides a "back calculation" output, and nearly every function block accepts a "back calculation" input from a downstream block. The "Control Selector" (CS) function block specified in the FOUNDATION Fieldbus standard provides the limiting function we need between the master and slave controllers. The BKCAL_OUT signal of this selector block connects to the master controller's BKCAL_IN input, making the master controller aware of its selection status. If ever the Control Selector function block de-selects the master controller's output, the controller will immediately know to halt integral action:

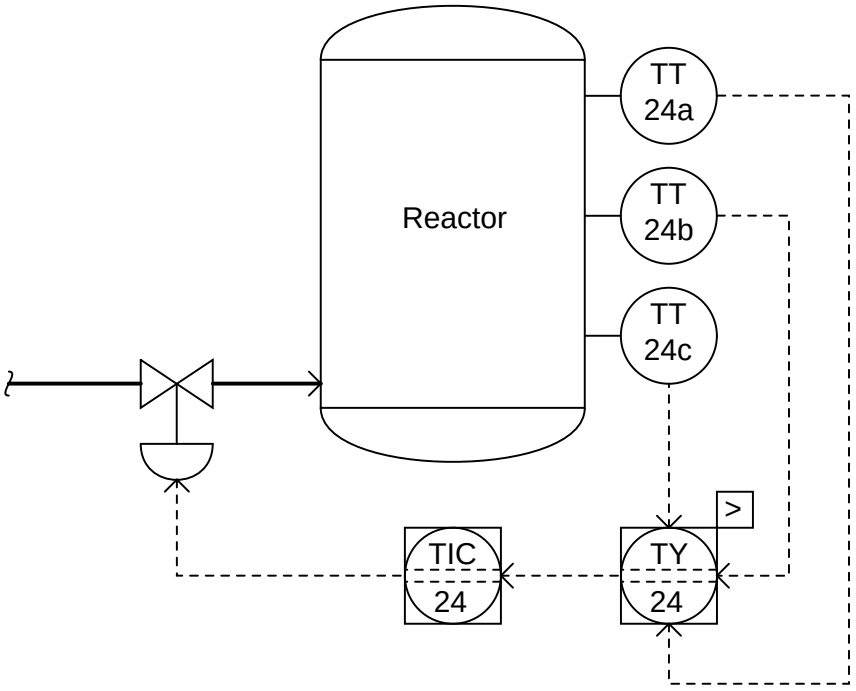
Mitigating master controller integral wind-up in a FOUNDATION Fieldbus high-limit control stragegy



28.7.2 Selector controls

In the broadest sense, a "selector" control strategy is one where one signal gets selected from multiple signals in a system to perform a measurement control function. In the context of this book and this chapter, I will use the term "selector" to categorize the automatic selection of a measurement or setpoint signal. Selection of a controller output signal will be explored in the next subsection.

Perhaps one of the simplest examples of a selector control strategy is where we must select a process variable signal from multiple transmitters. For example, consider this chemical reactor, where the control system must throttle the flow of coolant to keep the *hottest* measured temperature at setpoint, since the reaction happens to be exothermic (heat-releasing)[20]:

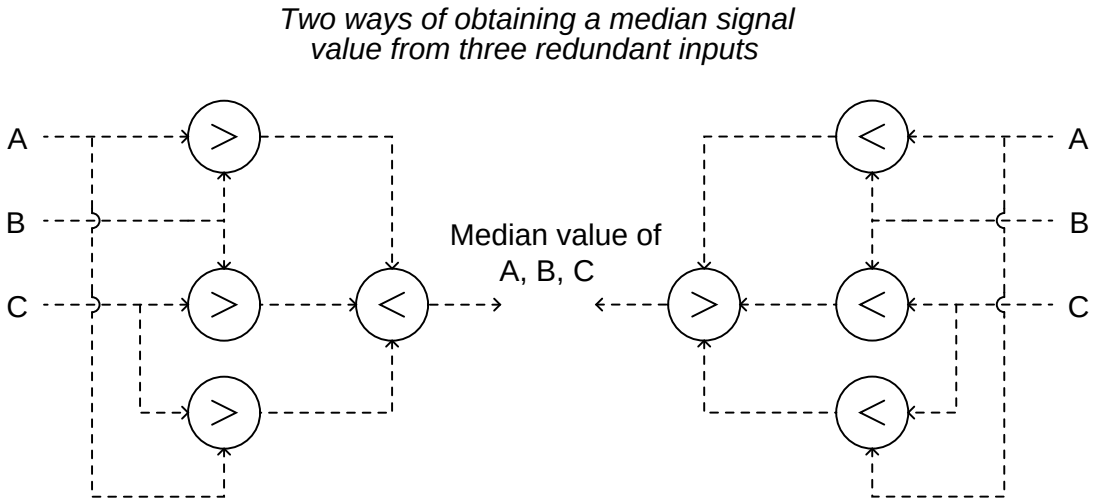


The high-select relay (TY-24) sends only the highest temperature signal from the three transmitters to the controller. The other two temperature transmitter signals are simply ignored.

Another use of selector relays (or function blocks) is for the determination of a *median* process measurement. This sort of strategy is often used on triple-redundant measurement systems, where three transmitters are installed to measure the exact same process variable, providing a valid measurement even in the event of transmitter failure.

[20] Only the coolant flow control instruments and piping are shown in this diagram, for simplicity. In a real P&ID, there would be many more pipes, valves, and other apparatus shown surrounding this process vessel.

The median select function may be implemented one of two ways using high- and low-select function blocks:

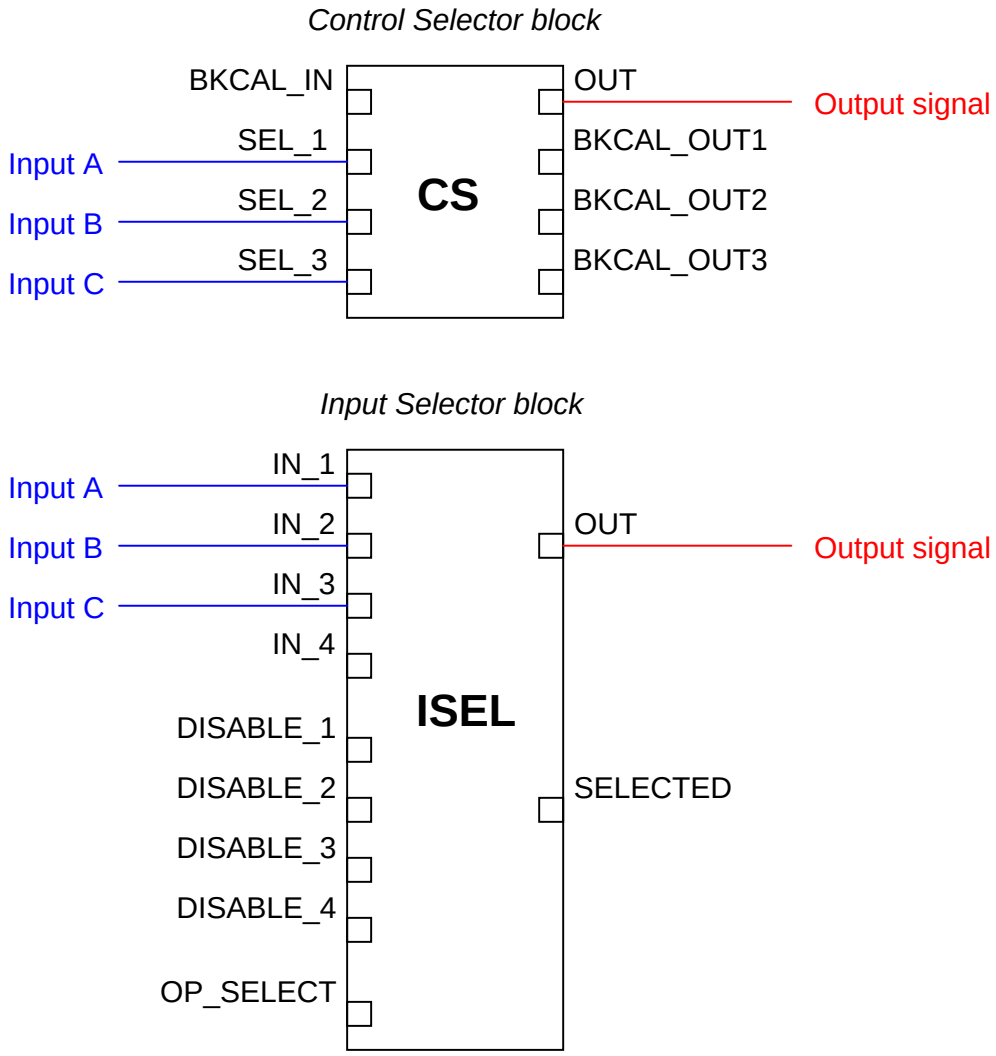


The left-hand selector strategy selects the highest value from each pair of signals (A and B, B and C, A and C), then selects the lowest value of those three primary selections. The right-hand strategy is exactly opposite – first selecting the lowest value from each input pair, then selecting the highest of those values – but it still accomplishes the same function. Either strategy outputs the *middle* value of the three input signals[21].

Although either of these methods of obtaining a median measurement requires four signal selector functions, it is quite common to find function blocks available in control systems ready to perform the median select function all in a single block. The median-select function is so common to redundant sensor control systems that many control system manufacturers provide it as a standard function unto itself.

[21] In order to understand how this works, I advise you try a "thought experiment" for each function block network whereby you arbitrarily assign three different numerical values for A, B, and C, then see for yourself which of those three values becomes the output value.

This is certainly true in the FOUNDATION Fieldbus standard, where two standardized function blocks are capable of this function, the CS (Control Selector) and the ISEL (Input Selector) blocks:



Of these two Fieldbus function blocks, the latter (ISEL) is expressly designed for selecting transmitter signals, whereas the former (CS) is best suited for selecting controller outputs with its "back calculation" facilities designed to modify the response of all de-selected controllers. Using the terminology of this book section, the ISEL function block is best suited for *selector* strategies, while the CS function block is ideal for *override* strategies (discussed in the next section).

If receiving three "good" inputs, the ISEL function block will output the middle (median) value

of the three. If one of the inputs carries a "bad" status[22], the ISEL block outputs the averaged value of the remaining two (good) inputs. Note how this function block also possesses individual "disable" inputs, giving external boolean (on/off) signals the ability to disable any one of the transmitter inputs to this block. Thus, the ISEL function block may be configured to de-select a particular transmitter input based on some programmed condition other than internal diagnostics.

If receiving four "good" inputs, the ISEL function block normally outputs the average value of the two middle (median) signal values. If one of the four inputs becomes "bad" is disabled, the block behaves as a normal three-input median select.

A general design principle for redundant transmitters is that you *never* install exactly two transmitters to measure the same process variable. Instead, you should install three (minimum). The problem with having two transmitters is a lack of information for "voting" if the two transmitters happen to disagree. In a three-transmitter system, the function blocks may select the median signal value, or average the "best 2 out of 3." If there are just two transmitters installed, and they do not substantially agree with one another, it is anyone's guess which one should be trusted[23].

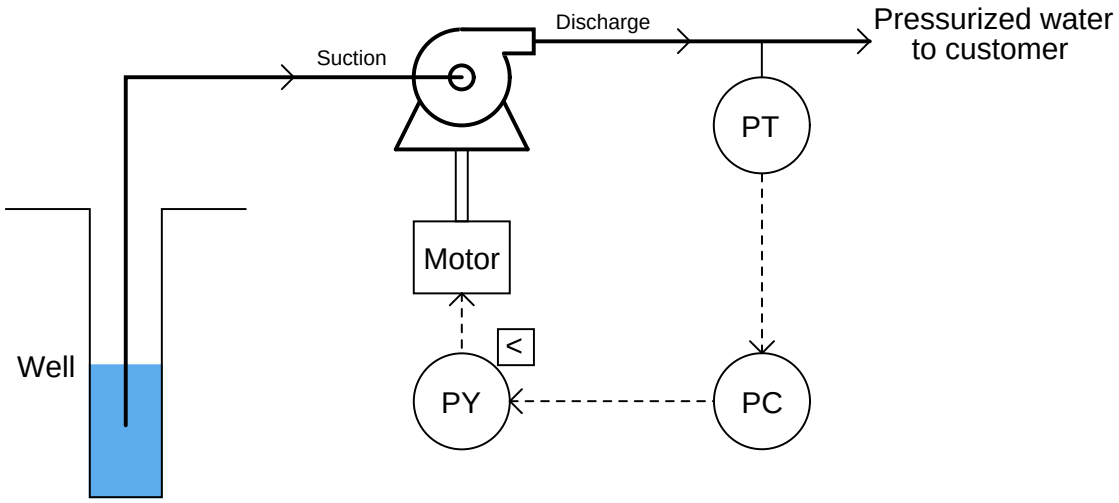
[22] In FOUNDATION Fieldbus, each and every signal path not only carries the signal value, but also a "status" flag declaring it to be "Good," "Bad," or "Uncertain." This status value gets propagated down the entire chain of connected function blocks, to alert dependent blocks of a possible signal integrity problem if one were to occur.

[23] This principle holds true even for systems with no function blocks "voting" between the redundant transmitters. Perhaps the installation consists of two transmitters with remote indications for a human operator to view. If the two displays substantially disagree, which one should the operator trust? A set of *three* indicators would be much better, providing the operator with enough information to make an intelligent decision on which display(s) to trust.

28.7.3 Override controls

An "override" control strategy involves a selection between two or more controller *output* signals, where only one controller at a time gets the opportunity to exert control over a process. All other "de-selected" controllers are thus *overridden* by the selected controller.

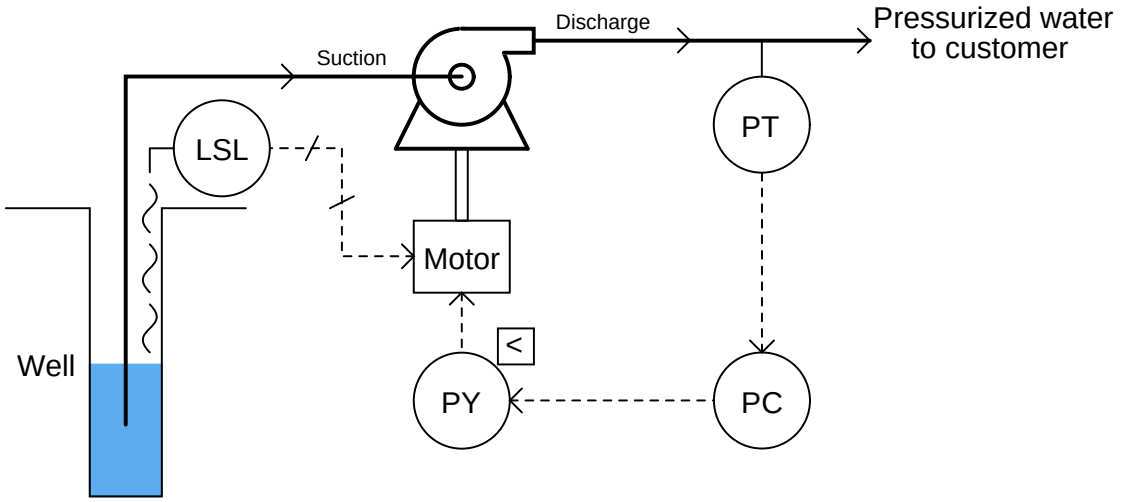
Consider this water pumping system, where a water pump is driven by a variable-speed electric motor to draw water from a well and provide constant water pressure to a customer:



Incidentally, this is an excellent application for a variable-speed motor as the final control element rather than a control valve. Reducing pump speed in low-flow conditions will save a lot of energy over time compared to the energy that would be wasted by a constant-speed pump and control valve.

A potential problem with this system is the pump running "dry" if the water level in the well gets too low, as might happen during summer months when rainfall is low and customer demand is high. If the pump runs for too long with no water passing through it, the seals will become damaged. This will necessitate a complete shut-down and costly rebuild of the pump, right at the time customers need it the most.

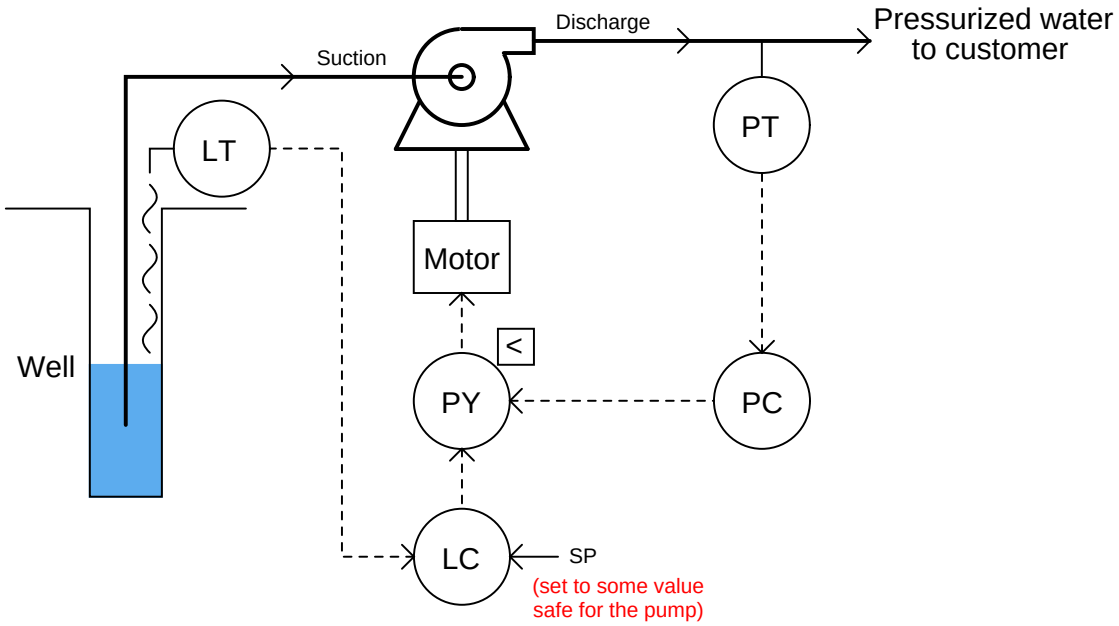
One solution to this problem would be to install a level switch in the well, sensing water level and shutting off the electric motor driving the pump if the water level ever gets too low:



This may be considered a kind of "override" strategy, because the low-level switch over-rides the pressure controller's command for the pump to turn. It is also a crude solution to the problem, for while it protects the pump from damage, it does so at the cost of completely shutting off water to customers. One way to describe this control strategy would be to call it a *hard override* system, suggesting the uncompromising action it will take to protect the pump.

A better solution to the dilemma would be to have the pump merely slow down as the well water level approaches a low-level condition. This way at least the pump could be kept running (and some amount of pressure maintained), decreasing demand on the well while maintaining curtailed service to customers and still protecting the pump from dry-running. This would be termed a *Soft override* system.

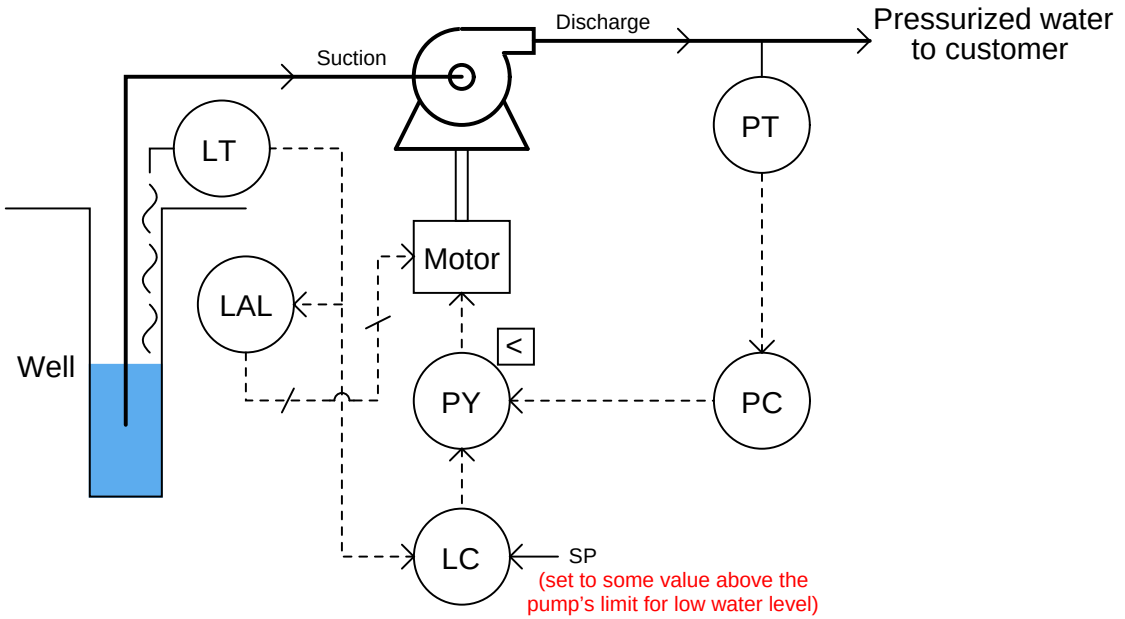
We may create just such a control strategy by replacing the well water level switch with a level *transmitter*, connecting the level transmitter to a level controller, and using a low-select relay or function block to select the lowest-valued output between the pressure and level controllers. The level controller's setpoint will be set at some low level above the acceptable limit for continuous pump operation:



If ever the well's water level goes below this setpoint, the level controller will command the pump to slow down, even if the pressure controller is calling for a higher speed. The level controller will have *overridden* the pressure controller, prioritizing pump longevity over customer demand.

Bear in mind that the concept of a low-level switch completely shutting off the pump is not an entirely bad idea. In fact, it might be prudent to integrate such a "hard" shutdown control in the override control system, just in case something goes wrong with the level controller (e.g. an improperly adjusted setpoint or poor tuning) or the low-select function.

With two layers of safety control for the pump, this system provides both a "soft constraint" providing moderated action and a "hard constraint" providing aggressive action to protect the pump from dry operation:



In order that these two levels of pump protection work in the proper order, the level controller's (LC) setpoint needs to be set to a higher value than the low level alarm's (LAL) trip point.

A very important consideration for any override control strategy is how to manage integral windup. Any time a controller with any integral (reset) action at all is de-selected by the selector function, the integral term of the controller will have the tendency to wind up (or wind down) over time. With the output of that controller de-coupled from the final control element, it can have no effect on the process variable. Thus, integral control action – the purpose of which being to constantly drive the output signal in the direction necessary to achieve zero error between process variable and setpoint – will work in vain to eliminate an error it cannot influence. If and when control is handed back to that controller, the integral action will have to spend time "winding" the other way to un-do what it did while it was de-selected.

Thus, override controls demand some form of integral windup limits that engage when a controller is de-selected. Methods of accomplishing this function are discussed in an earlier section on limit controls (section 28.7.1, beginning on page 1632).

References

Austin, George T., *Shreve's Chemical Process Industries*, McGraw-Hill Book Company, New York, NY, 1984.

"Foundation^{TM} Fieldbus Blocks", document 00809-0100-4783, Rev BA, Rosemount, Inc., Chanhassen, MN, 2000.

"Function Blocks Instruction Manual", document FBLOC-FFME, Smar Equipamentos Ind. Ltda., Sertãozinho, Brazil, 2005.

Lavigne, John R., *An Introduction To Paper Industry Instrumentation*, Miller Freeman Publications, Inc., San Francisco, CA, 1972.

Lavigne, John R., *Instrumentation Applications for the Pulp and Paper Industry*, The Foxboro Company, Foxboro, MA, 1979.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Control Volume II*, Third Edition, CRC Press, Boca Raton, FL, 1999.

Mollenkamp, Robert A., *Introduction to Automatic Process Control*, Instrument Society of America, Research Triangle Park, NC, 1984.

Palm, William J., *Control Systems Engineering*, John Wiley & Sons, Inc., New York, NY, 1986.

Shinskey, Francis G., *Energy Conservation through Control*, Academic Press, New York, NY, 1978.

Shinskey, Francis G., *Process-Control Systems – Application / Design / Adjustment*, Second Edition, McGraw-Hill Book Company, New York, NY, 1979.

Chapter 29

Process safety and instrumentation

This chapter discusses instrumentation issues related to industrial process safety. Instrumentation safety may be broadly divided into two categories: how instruments themselves may pose a safety hazard (electrical signals possibly igniting hazardous atmospheres), and how instruments and control systems may be configured to detect unsafe process conditions and automatically shut an unsafe process down.

In either case, the intent of this chapter is to help define and teach how to mitigate hazards encountered in certain instrumented processes. I purposely use the word "mitigate" rather than "eliminate" because the complete elimination of all risk is an impossibility. Despite our best efforts and intentions, no one can absolutely eliminate all dangers from industrial processes[1]. What we can do, though, is *significantly* reduce those risks to the point they begin to approach the low level of "background" risks we all face in daily life, and that is no small achievement.

29.1 Classified areas and electrical safety measures

Any physical location in an industrial facility harboring the potential of explosion due to the presence of flammable process matter suspended in the air is called a *hazardous* or *classified* location. In this context, the label "hazardous" specifically refers to the hazard of explosion, not of other health or safety hazards[2].

[1] For that matter, it is impossible to eliminate all danger from *life in general*. Every thing you do (or don't do) involves some level of risk. The question really should be, "how much risk is there in a given action, and how much risk am I willing to tolerate?" To illustrate, there does exist a non-zero probability that something you will read in this book is so shocking it will cause you to have a heart attack. However, the odds of you walking away from this book down and never reading it again over concern of epiphany-induced cardiac arrest are just as slim.

[2] Chemical corrosiveness, biohazardous substances, poisonous materials, and radiation are all examples of other types of industrial hazards not covered by the label "hazardous" in this context. This is not to understate the danger of these other hazards, but merely to focus our attention on the specific hazard of explosions and how to build instrument systems that will not trigger explosions due to electrical spark.

29.1.1 Classified area taxonomy

In the United States, the National Electrical Code (NEC) published by the National Fire Protection Association (NFPA) defines different categories of "classified" industrial areas and prescribes safe electrical system design practices for those areas. Article 500 of the NEC categorizes classified areas into a system of *Classes* and *Divisions*. Articles 505 and 506[3] of the NEC provide alternative categorizations for classified areas based on *Zones* that is more closely aligned with European safety standards.

The Class and Division taxonomy defines classified areas in terms of hazard type and hazard probability. Each "Class" contains (or may contain) different types of potentially explosive substances: Class I is for gases or vapors, Class II is for combustible dusts, and Class III is for flammable fibers. The three-fold class designation is roughly scaled on the size of the flammable particles, with Class I being the smallest (gas or vapor molecules) and Class III being the largest (fibers of solid matter). Each "Division" ranks a classified area according to the likelihood of explosive gases, dusts, or fibers being present. Division 1 areas are those where explosive concentrations can or do exist under normal operating conditions. Division 2 areas are those where explosive concentrations only exist infrequently or under abnormal conditions[4].

The "Zone" method of area classifications defined in Article 505 of the National Electrical Code applies to Class I (explosive gas or vapor) applications, but the three-fold Zone ranks (0, 1, and 2) are analogous to Divisions in their rating of explosive concentration probabilities. Zone 0 defines areas where explosive concentrations are continually present or normally present for long periods of time. Zone 1 defines areas where those concentrations may be present under normal operating conditions, but not as frequently as Zone 0. Zone 2 defines areas where explosive concentrations are unlikely under normal operating conditions, and when present do not exist for substantial periods of time. This three-fold Zone taxonomy may be thought of as expansion on the two-fold Division system, where Zones 0 and 1 are sub-categories of Division 1 areas, and Zone 2 is nearly equivalent to a Division 2 area[5]. A similar three-zone taxonomy for Class II and Class III applications is defined in Article 506 of the National Electrical Code, the zone ranks for these dust and fiber hazards numbered 20, 21, and 22 (and having analogous meanings to zones 0, 1, and 2 for Class I applications).

[3] Article 506 is a new addition to the NEC as of 2008. Prior to that, the only "zone"-based categories were those specified in Article 505.

[4] The final authority on Class and Division definitions is the National Electrical Code itself. The definitions presented here, especially with regard to Divisions, may not be precise enough for many applications. Article 500 of the NEC is quite specific for each Class and Division combination, and should be referred to for detailed information in any particular application.

[5] Once again, the final authority on this is the National Electrical Code, in this case Article 505. My descriptions of Zones and Divisions are for general information only, and may not be specific or detailed enough for many applications.

Within Class I and Class II (but not Class III), the National Electrical Code further sub-divides hazards according to explosive properties called *Groups*. Each group is defined either according to a substance type, or according to specific ignition criteria. Ignition criteria listed in the National Electrical Code (Article 500) include the *maximum experimental safe gap* (MESG) and the *minimum ignition current ratio* (MICR). The MESG is based on a test where two hollow hemispheres separated by a small gap enclose both an explosive air/fuel mixture and an ignition source. Tests are performed with this apparatus to determine the maximum gap width between the hemispheres that will not permit the excursion of flame from an explosion within the hemispheres triggered by the ignition source. The MICR is the ratio of electrical ignition current for an explosive air/fuel mixture compared to an optimum mixture of methane and air. The smaller of either these two values, the more dangerous the explosive substance is.

Class I substances are grouped according to their respective MESG and MICR values, with typical gas types given for each group:

Group	**Typical substance**	**Safe gap**	**Ignition current**
A	Acetylene		
B	Hydrogen	MESG $\leq$ 0.45 mm	MICR $\leq$ 0.40
C	Ethylene	0.45 mm $<$ MESG $\leq$ 0.75 mm	0.40 $<$ MICR $\leq$ 0.80
D	Propane	0.75 mm $<$ MESG	0.80 $<$ MICR

Class II substances are grouped according to material type:

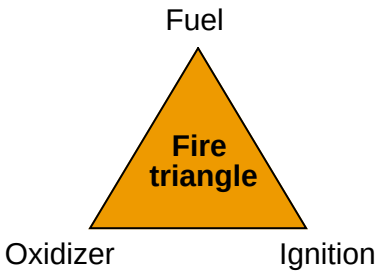
Group	**Substances**
E	Metal dusts
F	Carbon-based dusts
G	Other dusts (wood, grain, flour, plastic, etc.)

Just to make things confusing, the Class/Zone system described in NEC Article 505 uses a completely different lettering order to describe gas and vapor groups (at the time of this writing there is no grouping of dust or fiber types for the zone system described in Article 506 of the NEC):

Group	**Typical substance(s)**	**Safe gap**	**Ignition current**
IIC	Acetylene, Hydrogen	MESG $\leq$ 0.50 mm	MICR $\leq$ 0.45
IIB	Ethylene	0.50 mm $<$ MESG $\leq$ 0.90 mm	0.45 $<$ MICR $\leq$ 0.80
IIA	Acetone, Propane	0.90 mm $<$ MESG	0.80 $<$ MICR

29.1.2 Explosive limits

In order to have combustion (an explosion being a particularly aggressive form of combustion), three basic criteria must be satisfied: sufficient *fuel*, sufficient *oxidizer*, and sufficient *energy* for ignition. These three necessities are referred to as the *fire triangle*:



Remove any one (or more) of these elements from a location, according to the fire triangle, and you will successfully prevent the possibility of fire or explosion.

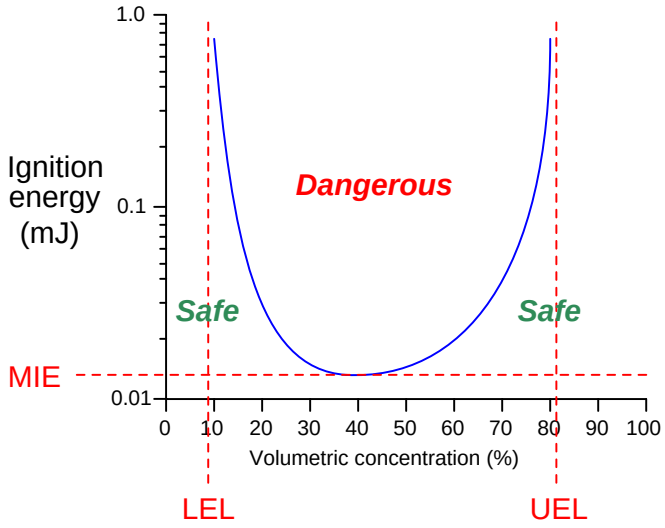
The fire triangle serves as a qualitative guide for *preventing* fires and explosions, but it does not give sufficient information to tell us if the necessary conditions exist to *support* a fire or explosion. For that, we need more quantitative data on the fuel, the oxidizer, and the ignition source. In order for a fire or explosion to occur, we need to have an adequate mixture of fuel and oxidizer in the correct proportions, and a source of ignition energy exceeding a certain minimum threshold.

Suppose we had a laboratory test chamber filled with a mixture of acetone vapor (70% by volume) and air at room temperature, with an electrical spark gap providing convenient ignition. No matter how energetic the spark, this mixture would not explode, because there is too *rich* a mixture of acetone (i.e. too much acetone mixed with not enough air). Every time the spark gap discharges, its energy would surely cause some acetone molecules to combust with available oxygen molecules. However, since the air is so dilute in this rich acetone mixture, those scarce oxygen molecules are depleted fast enough that the flame temperature quickly falls off and is no longer hot enough to trigger the remaining oxygen molecules to combust with the plentiful acetone molecules.

The same problem occurs if the acetone/air mixture is too *lean* (not enough acetone and too much air). This is what would happen if we diluted the acetone vapors to a volumetric concentration of only 0.5% inside the test chamber: any spark at the gap would indeed cause some acetone molecules to combust, but there would be too few available to support expansive combustion across the rest of the chamber.

We could also have an acetone/air mixture in the chamber ideal for combustion (about 9.5% acetone by volume) and still not have an explosion if the spark's energy were insufficient. Most combustion reactions require a certain minimum level of *activation energy* to overcome the potential barrier before molecular bonding between fuel atoms and oxidizer atoms occurs. Stated differently, many combustion reactions are not *spontaneous* at room temperature and at atmospheric pressure – they need a bit of "help" to initiate.

All the necessary conditions for an explosion may be quantified and plotted as an *ignition curve* for any particular fuel and oxidizer combination. This next graph shows an ignition curve for an hypothetical fuel gas mixed with air:



Note how any point in the chart lying *above* the curve is "dangerous," while any point *below* the curve is "safe." The three critical values on this graph are the *Lower Explosive Limit* (LEL), the *Upper Explosive Limit* (UEL), and the *Minimum Ignition Energy* (MIE). These critical values differ for every type of fuel and oxidizer combination, change with ambient temperature and pressure, and may be rendered irrelevant in the presence of a catalyst (a chemical substance that works to promote a reaction without itself being consumed by the reaction). Most ignition curves are published with the assumed conditions of air as the oxidizer, at room temperature and at atmospheric pressure.

The greater the difference in LEL and UEL values, the greater "explosive potential" a fuel gas or vapor presents (all other factors being equal), because it means the fuel may explode over a wider range of mixture conditions. It is instructive to research the LEL and UEL values for many common substances, just to see how "explosive" they are relative to each other:

Substance	**LEL** (% volume)	**UEL** (% volume)
Acetylene	2.5%	100%
Acetone	2.5%	12.8%
Butane	1.5%	8.5%
Carbon disulfide	1.3%	50%
Carbon monoxide	12.5%	74%
Ether	1.9%	36%
Gasoline	1.4%	7.6%
Kerosene	0.7%	5%
Hydrazine	2.9%	98%
Hydrogen	4.0%	75%
Methane	4.4%	17%
Propane	2.1%	9.5%

Note how acetylene has a UEL value of 100%. This means it is possible for acetylene gas to explode *even when there is no oxidizer present*. Some other chemical substances exhibit this same property (ethylene oxide and n-propyl nitrate being two more examples), where the lack of an oxidizer does not prevent an explosion. Some other substances have UEL values very close to 100%, hydrazine being one example at 98%. In these substances we see an important exception to the "fire triangle" rule that the elimination of any one element prevents combustion. With these substances in high concentration, our only practical hope of avoiding explosion is to eliminate the possibility of an ignition source in its presence.

29.1.3 Protective measures

Different strategies exist to help prevent electrical devices from triggering fires or explosions in classified areas. These strategies may be broadly divided four ways:

- **Contain the explosion:** enclose the device inside a very strong box that contains any explosion generated by the device so as to not trigger a larger explosion outside the box. This strategy may be viewed as eliminating the "ignition" component of the fire triangle, from the perspective of the atmosphere outside the explosion-proof enclosure (ensuring the explosion inside the enclosure does not ignite a larger explosion outside).
- **Shield the device:** enclose the electrical device inside a suitable box or shelter, then purge that enclosure with clean air (or a pure gas) that prevents an explosive mixture from forming inside the enclosure. This strategy works by eliminating either the "fuel" component of the fire triangle (if purged by air), by eliminating the "oxidizer" component of the fire triangle (if purged by fuel gas), or by eliminating both (if purged by an inert gas).
- **Encapsulated design:** manufacture the device so that it is self-enclosing. In other words, build the device in such a way that any spark-producing elements are sealed air-tight within the device from any explosive atmosphere. This strategy works by eliminating the "ignition" component of the fire triangle (from the perspective of outside the device) or by eliminating both "fuel" and "oxidizer" components (from the perspective of inside the device).
- **Limit total circuit energy:** design the circuit such that there is insufficient energy to trigger an explosion, even in the event of an electrical fault. This strategy works by eliminating the "ignition" component of the fire triangle.

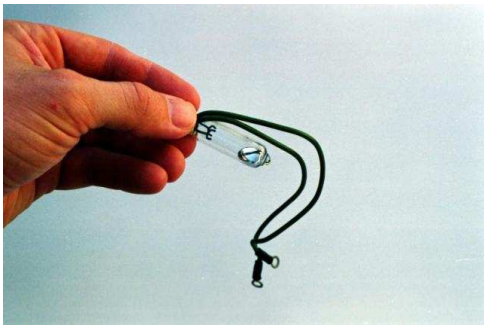
A common example of the first strategy is to use extremely rugged metal *explosion-proof* (NEMA 7) enclosures instead of the more common sheet-metal or fiberglass enclosures to house electrical equipment. Two photographs of explosion-proof electrical enclosures reveal their unusually rugged construction:

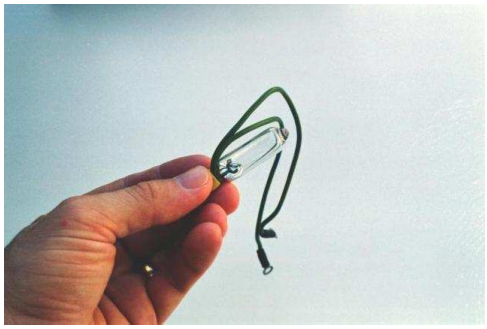
Note the abundance of bolts securing the covers of these enclosures! This is necessary in order to withstand the enormous forces generated by the pressure of an explosion developing inside the enclosure. Note also how most of the bolts have been removed from the door of the right-hand enclosure. This is an unsafe and very unfortunate occurrence at many industrial facilities, where technicians leave just a few bolts securing the cover of an explosion-proof enclosure because it is so time-consuming to remove all of them to gain access inside the enclosure for maintenance work. Such practices negate the safety of the explosion-proof enclosure, rendering it just as dangerous as a sheet metal enclosure in a classified area.

Explosion-proof enclosures are designed in such a way that high-pressure gases resulting from an explosion within the enclosure must pass through small gaps (either holes in vent devices, and/or the gap formed by a bulging door forced away from the enclosure box) en route to exiting the enclosure. As hot gases pass through these tight metal gaps, they are forced to cool to the point where they will not ignite explosive gases outside the enclosure, thus preventing the original explosion inside the enclosure from triggering a far more violent event.

A similar strategy involves the use of a non-flammable *purge gas* pressurizing an ordinary electrical enclosure such that explosive atmospheres are prevented from entering the enclosure. Ordinary compressed air may be used as the purge gas, so long as provisions are made to ensure the air compressor supplying the compressed air is in a non-classified area where explosive gases will never be drawn into the compressed air system.

Devices may be encapsulated in such a way that explosive atmospheres cannot penetrate the device to reach anything generating sufficient spark or heat. *Hermetically sealed* devices are an example of this protective strategy, where the structure of the device has been made completely fluid-tight by fusion joints of its casing. Mercury tilt-switches are good examples of such electrical devices, where a small quantity of liquid mercury is hermetically sealed inside a glass tube. No outside gases, vapors, dusts, or fibers can ever reach the spark generated when the mercury comes into contact (or breaks contact with) the electrodes:





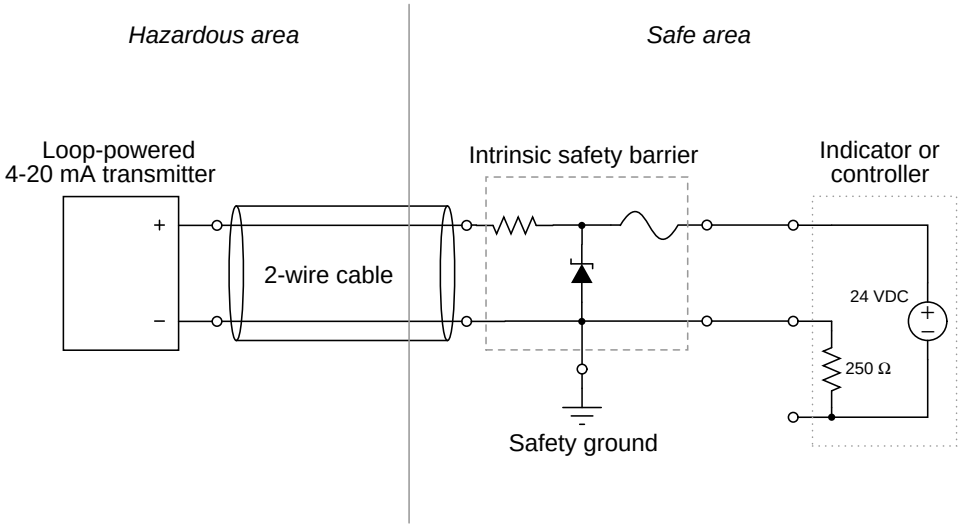
The ultimate method for ensuring instrument circuit safety in classified areas is to intentionally limit the amount of energy available within a circuit such that it *cannot* generate enough heat or spark to ignite an explosive atmosphere, even in the event of an electrical fault within the circuit. Article 504 of the National Electrical Code specifies standards for this method. Any system meeting these requirements is called an *intrinsically safe* or *I.S.* system. The word "intrinsic" implies that the safety is a natural property of the circuit, since it lacks even the ability to produce an explosion-triggering spark[6].

One way to underscore the meaning of intrinsic safety is to contrast it against a different concept that has the appearance of similarity. Article 500 of the National Electrical Code defines *nonincendive equipment* as devices incapable of igniting a hazardous atmosphere *under normal operating conditions*. However, the standard for nonincendive devices or circuits does not guarantee what will happen under *abnormal* conditions, such as an open- or short-circuit in the wiring. So, a "nonincendive" circuit may very well pose an explosion hazard, whereas an "intrinsically safe" circuit will not because the intrinsically safe circuit simply does not possess enough energy to trigger an explosion under any condition. As a result, nonincendive circuits are not approved in Class I or Class II Division 1 locations whereas intrinsically safe circuits are approved for all hazardous locations.

Most modern 4 to 20 mA analog signal instruments may be used as part of intrinsically safe circuits so long as they are connected to control equipment through suitable *safety barrier* interfaces.

[6] To illustrate this concept in a different context, consider my own personal history of automobiles. For many years I drove an ugly and inexpensive truck which I joked had "intrinsic theft protection:" it was so ugly, no one would ever want to steal it. Due to this "intrinsic" property of my vehicle, I had no need to invest in an alarm system or any other protective measure to deter theft. Similarly, the components of an intrinsically safe system need not be located in explosion-proof or purged enclosures because the intrinsic energy limitation of the system is protection enough.

A simple intrinsic safety barrier circuit made from passive components is shown in the following diagram[7]:



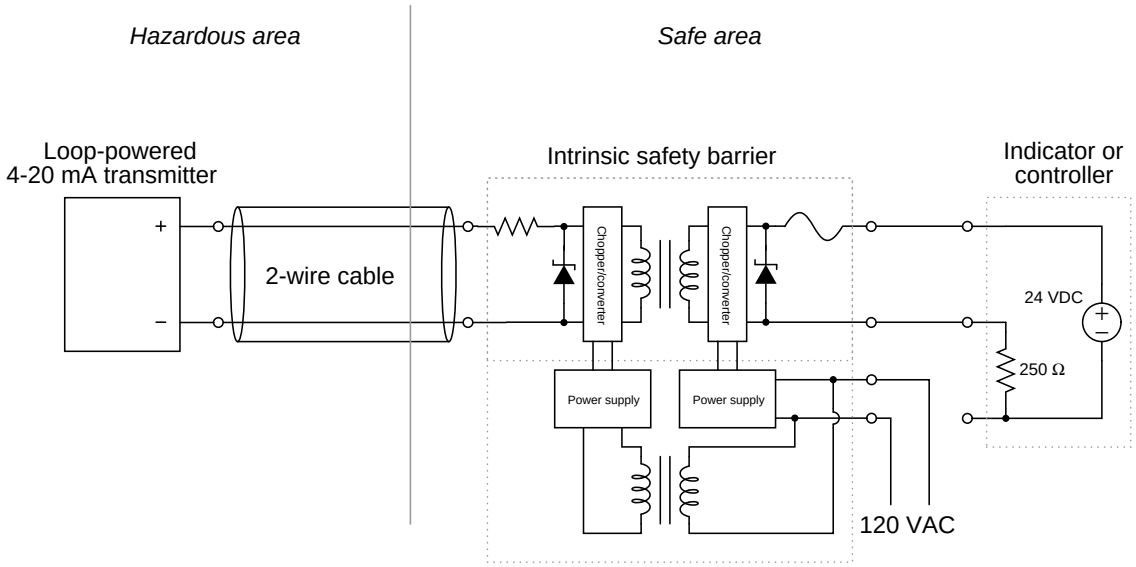
In normal operation, the 4-20 mA field instrument possesses insufficient terminal voltage and insufficient loop current to pose any threat of hazardous atmosphere ignition. The series resistance of the barrier circuit is low enough that the 4-20 mA signal will be unaffected by its presence. As far as the receiving instrument (indicator or controller) is "concerned," the safety barrier might as well not exist.

If a short-circuit develops in the field instrument, the series resistance of the barrier circuit will limit fault current to a value low enough not to pose a threat in the hazardous area. If something fails in the receiving instrument to cause a much greater power supply voltage to develop at its terminals, the zener diode inside the barrier will break down and provide a shunt path for fault current that bypasses the field instrument (and may possibly blow the fuse in the barrier). Thus, the intrinsic safety barrier circuit provides protection against overcurrent *and* overvoltage faults, so that neither type of fault will result in enough electrical energy available at the field device to ignite an explosive atmosphere.

Note that a barrier device such as this *must* be present in the 4-20 mA analog circuit in order for the circuit to be intrinsically safe. The "intrinsic" safety rating of the circuit depends on this barrier, not on the integrity of the field device or of the receiving device. Without this barrier in place, the instrument circuit is not intrinsically safe, even though the *normal* operating conditions of the field device and receiving device are well within the parameters of safety for classified areas. It is the barrier and the barrier alone which guarantees those voltage and current levels will remain within safe limits in the event of *abnormal* circuit conditions such as a field wiring short or a faulty loop power supply.

[7] Real passive barriers often used redundant zener diodes connected in parallel to ensure protection against excessive voltage even in the event of a zener diode failing open.

More sophisticated *active* barrier devices are manufactured which provide electrical isolation from ground in the instrument wiring, thus eliminating the need for a safety ground connection at the barrier device.



In the example shown here, transformers[8] are used to electrically isolate the analog current signal so that there is no path for DC fault current between the field instrument and the receiving instrument, ground or no ground.

29.2 Concepts of probability and reliability

While the term "probability" may evoke images of imprecision, probability is in fact an exact mathematical science. Reliability, which is the expression of how likely something is *not* to fail when needed, is based on the mathematics of probability. Therefore, a rudimentary understanding of probability mathematics is necessary to grasp what reliability means in a quantitative sense, and how system reliability may be improved through judicious application of probability principles.

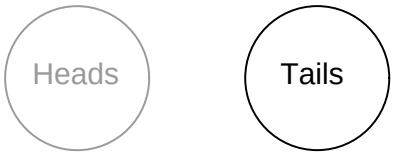
[8] Of course, transformers cannot be used to pass DC signals of any kind, which is why chopper/converter circuits are used before and after the signal transformer to convert each DC current signal into a form of chopped (AC) signal that *can* be fed through the transformer. This way, the *information* carried by each 4-20 mA DC current signal passes through the barrier, but electrical fault current cannot.

29.2.1 Mathematical probability

Probability may be defined as a ratio of specific outcomes to total (possible) outcomes. If you were to flip a coin, there are really only two possibilities[9] for how that coin may land: face-up ("heads") or face-down ("tails"). The probability of a coin falling "tails" is thus one-half ($\frac{1}{2}$), since "tails" is but one specific outcome out of two total possibilities. Calculating the probability (P) is a matter of setting up a ratio of outcomes:

$$P(\text{“tails”}) = \frac{\text{“tails”}}{\text{“heads”} + \text{“tails”}} = \frac{1}{2} = 0.5$$

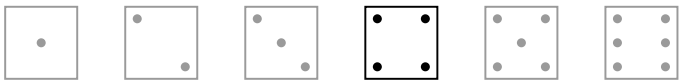
This may be shown graphically by displaying all possible outcomes for the coin's landing ("heads" or "tails"), with the one specific outcome we're interested in ("tails") highlighted for emphasis:



The probability of the coin landing "heads" is of course exactly the same, because "heads" is also *one* specific outcome out of *two* total possibilities.

If we were to roll a six-sided die, the probability of that die landing on any particular side (let's say the "four" side) is one out of six, because we're looking at one specific outcome out of six total possibilities:

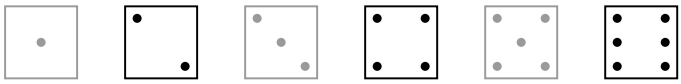
$$P(\text{“four”}) = \frac{\text{“four”}}{\text{“one”} + \text{“two”} + \text{“three”} + \text{“four”} + \text{“five”} + \text{“six”}} = \frac{1}{6} = 0.1\overline{66}$$



[9] To be honest, the coin could also land on its edge, which is a third possibility. However, that third possibility is so remote as to be negligible in the presence of the other two.

If we were to roll the same six-sided die, the probability of that die landing on an even-numbered side (2, 4, or 6) is three out of six, because we're looking at three specific outcomes out of six total possibilities:

$$P(\text{even}) = \frac{\text{“two” + “four” + “six”}}{\text{“one” + “two” + “three” + “four” + “five” + “six”}} = \frac{3}{6} = 0.5$$



As a ratio of specific outcomes to total possible outcomes, the probability of any event will always be a number ranging in value from 0 to 1, inclusive. This value may be expressed as a fraction ($\frac{1}{2}$), as a decimal (0.5), or as a verbal statement (e.g. “three out of six”). A probability value of zero (0) means a specific event is impossible, while a probability of one (1) means a specific event is guaranteed to occur.

Probability values realistically apply only to large samples. A coin tossed ten times may very well fail to land “heads” exactly five times and land “tails” exactly five times. For that matter, it may fail to land on each side exactly 500,000 times out of a million tosses. However, so long as the coin and the coin-tossing method are *fair* (i.e. not biased in any way), the experimental results will approach[10] the ideal probability value as the number of trials approaches infinity. Ideal probability values become less and less certain as the number of trials decreases, and become completely useless for singular (non-repeatable) events.

A familiar application of probability values is the forecasting of meteorological events such as rainfall. When a weather forecast service provides a rainfall prediction of 65% for a particular day, it means that out of a large number of days sampled in the past having similar measured conditions (cloud cover, barometric pressure, temperature and dew point, etc.), 65% of those days experienced rainfall. This past history gives us some idea of how likely rainfall will be for any present situation, based on similarity of measured conditions.

Like all probability values, forecasts of rainfall are more meaningful with greater samples. If we wish to know how many days with measured conditions similar to those of the forecast day will experience rainfall over the *next ten years* (3650 days total), the forecast probability value of 65% will be quite accurate. However, if we wish to know whether or not rain will fall on any particular (single) day having those same conditions, the value of 65% tells us very little. So it is with all measurements of probability: precise for large samples, ambiguous for small samples, and virtually meaningless for singular conditions[11].

In the field of instrumentation – and more specifically the field of *safety* instrumented systems – probability is useful for the mitigation of hazards based on equipment failures where the probability of failure for specific pieces of equipment is known from mass production of that equipment and years of data gathered describing the reliability of the equipment. If we have data showing the probabilities

[10] In his excellent book, *Reliability Theory and Practice*, Igor Bazovsky describes the relationship between true probability (P) calculated from ideal values and estimated probability ($\hat{P}$) calculated from experimental trials as a limit function: $P = \lim_{N \to \infty} \hat{P}$, where N is the number of trials.

[11] Most adults can recall instances where a weather forecast proved to be completely false: a prediction for rainfall resulting in a completely dry day, or visa-versa. In such cases, one is tempted to blame the weather service for poor forecasting, but in reality it has more to do with the nature of probability, specifically the meaninglessness of probability calculations in predicting singular events.

of failure for different pieces of equipment, we may use this data to calculate the probability of failure for the system as a whole. Furthermore, we may apply certain mathematical laws of probability to calculate system reliability for different equipment configurations, and therefore minimize the probability of system failure by optimizing those configurations.

Just like weather predictions, predictions of system reliability (or conversely, of system failure) become more accurate as the sample size grows larger. Given an accurate probabilistic model of system reliability, a system (or a set of systems) with enough individual components, and a sufficiently long time-frame, an organization may accurately predict the number of system failures and the cost of those failures (or alternatively, the cost of minimizing those failures through preventive maintenance). However, no probabilistic model will accurately predict which component in a large system will fail tomorrow, much less precisely 1000 days from now.

The ultimate purpose, then, in probability calculations for process systems and automation is to optimize the safety and availability of large systems over many years of time. Calculations of reliability, while useful to the technician in understanding the nature of system failures and how to minimize them, are actually more valuable (more meaningful) at the enterprise level. At the time of this writing (2009), there is already a strong trend in large-scale industrial control systems to provide more meaningful information to business managers in addition to the basic regulatory functions intrinsic to instrument loops, such that the control system actually functions as an optimizing engine for the enterprise as a whole[12], and not just for individual loops. I can easily foresee a day when control systems additionally calculate their own reliability based on manufacturer's test data (demonstrated Mean Time Between Failures and the like), maintenance records, and process history, offering forecasts of impending failure in the same way weather services offer forecasts of future rainfall.

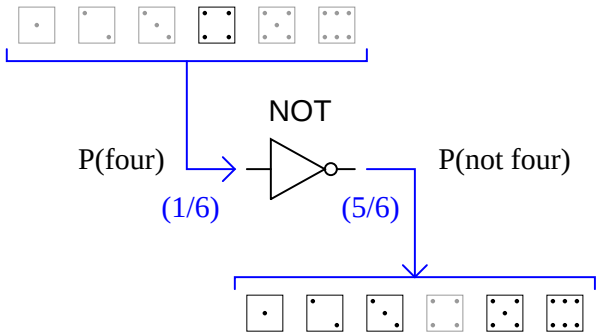
29.2.2 Laws of probability

Probability mathematics bears an interesting similarity to Boolean algebra in that probability values (like Boolean values) range between zero (0) and one (1). The difference, of course, is that while Boolean variables may *only* have values equal to zero or one, probability variables range continuously between those limits. Given this similarity, we may apply standard Boolean operations such as `NOT`, `AND`, and `OR` to probabilities. These Boolean operations lead us to our first "laws" of probability for combination events.

[12] As an example of this shift from basic loop control to enterprise optimization, consider the case of a highly automated lumber mill where logs are cut into lumber not only according to minimum waste, but also according to the real-time market value of different board types and stored inventory. Talking with an engineer about this system, we joked that the control system would purposely slice every log into toothpicks in an effort to maximize profit if the market value of toothpicks suddenly spiked!

The logical "NOT" function

For instance, if we know the probability of rolling a "four" on a six-sided die is $\frac{1}{6}$, then we may safely say the probability of *not* rolling a "four" is $\frac{5}{6}$, the complement of $\frac{1}{6}$. The common "inverter" logic symbol is shown here representing the complementation function, turning a probability of rolling a "four" into the probability of *not* rolling a "four":



Symbolically, we may express this as a sum of probabilities equal to one:

$$P(\text{total}) = P(\text{“one”}) + P(\text{“two”}) + P(\text{“three”}) + P(\text{“four”}) + P(\text{“five”}) + P(\text{“six”}) = 1$$

$$P(\text{total}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$$

$$P(\text{total}) = P(\text{“four”}) + P(\text{not “four”}) = \frac{1}{6} + \frac{5}{6} = 1$$

$$P(\text{“four”}) = 1 - P(\text{not “four”}) = 1 - \frac{5}{6} = \frac{1}{6}$$

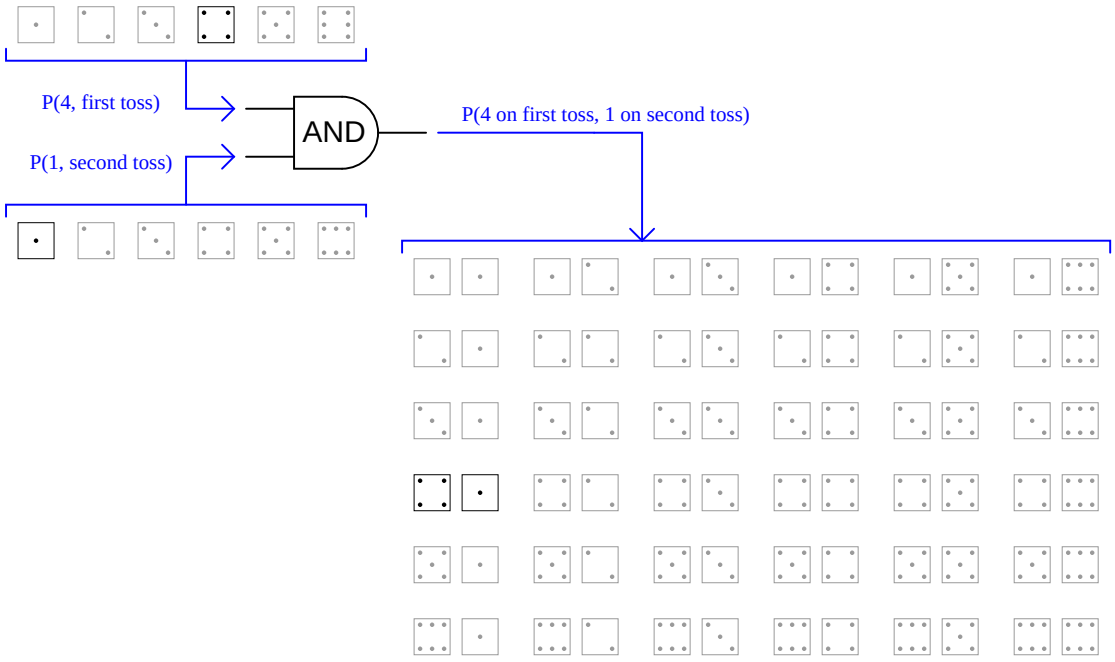
We may state this as a general "law" of complementation for any event (A):

$$P(A) = 1 - P(\overline{A})$$

The complement of a probability value finds frequent use in reliability engineering. If we know the probability value for the failure of a component (i.e. how likely it is to fail), then we know the *reliability* value (i.e. how likely it is to function properly) will be the complement of its failure probability. To illustrate, consider a device with a failure probability of $\frac{1}{100,000}$. Such a device could be said to have a reliability (R) value of $\frac{99,999}{100,000}$, or 99.999%, since $1 - \frac{1}{100,000} = \frac{99,999}{100,000}$.

The logical "AND" function

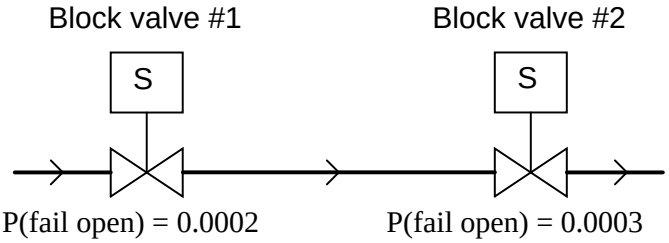
The AND function regards probabilities of two or more intersecting events (i.e. where the outcome of interest only happens if two or more events happen together, or in a specific sequence). Another example using a die is the probability of rolling a "four" on the first toss, then rolling a "one" on the second toss. It should be intuitively obvious that the probability of rolling this specific combination of values will be less (i.e. less likely) than rolling either of those values in a single toss, since two rolls gives us twice as many opportunities to land on the desired number. The shaded field of possibilities (36 in all) demonstrate the unlikelihood of this sequential combination of values compared to the unlikelihood of either value on either toss:



As you can see, there is but one outcome matching the specific criteria out of 36 total possible outcomes. This yields a probability value of one-in-thirty six ($\frac{1}{36}$) for the specified combination, which is the *product* of the individual probabilities. This, then, is our second law of probability:

$$P(\text{A } \textit{and } \text{B}) = P(\text{A}) \times P(\text{B})$$

A practical application of this would be the calculation of failure probability for a double-block valve assembly, designed to positively stop the flow of a dangerous process fluid. Double-block valves are used to provide increased assurance of shut-off, since the shutting of *either* block valve is sufficient in itself to stop fluid flow. The probability of failure for a double-block valve assembly – "failure" defined as not being able to stop fluid flow when needed – is the product of each valve's unreliability to close (i.e. probability of failing open):



With these two valves in service, the probability of neither valve successfully shutting off flow (i.e. *both* valve 1 *and* valve 2 failing on demand; remaining open when they should shut) is the product of their individual failure probabilities:

$$P(\text{assembly fail}) = P(\text{valve 1 fail open}) \times P(\text{valve 2 fail open})$$

$$P(\text{assembly fail}) = 0.0002 \times 0.0003$$

$$P(\text{assembly fail}) = 0.00000006 = 6 \times 10^{-8}$$

An extremely important assumption in performing such an `AND` calculation is that the probabilities of failure for each valve are not related. For instance, if the failure probabilities of both valve 1 and valve 2 were largely based on the possibility of a certain residue accumulating inside the valve mechanism (causing the mechanism to freeze in the open position), and *both* valves were equally susceptible to this residue accumulation, there would be virtually no advantage to having double block valves. If said residue were to accumulate in the piping, it would affect both valves practically the same. Thus, the failure of one valve due to this effect would virtually ensure the failure of the other valve as well. The probability of simultaneous or sequential events being the product of the individual events' probabilities is true if and only if the events in question are completely independent.

We may illustrate the same caveat with the sequential rolling of a die. Our previous calculation showed the probability of rolling a "four" on the first toss and a "one" on the second toss to be $\frac{1}{6} \times \frac{1}{6}$, or $\frac{1}{36}$. However, if the person throwing the die is extremely consistent in their throwing technique and the way they orient the die after each throw, such that rolling a "four" on one toss makes it very likely to roll a "one" on the next toss, the sequential events of a "four" followed by a "one" would be far more likely than if the two events were completely random and independent. The probability calculation of $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$ holds true only if all the throws' results are completely unrelated to each other.

Another, similar application of the Boolean `AND` function to probability is the calculation of system reliability (R) based on the individual reliability values of components necessary for the

system's function. If we know the reliability values for several crucial system components, and we also know those reliability values are based on independent (unrelated) failure modes, the overall system reliability will be the product (Boolean `AND`) of those component reliabilities. This mathematical expression is known as *Lusser's product law of reliabilities*:

$$R_{system} = R_1 \times R_2 \times R_3 \times \cdots \times R_n$$

As simple as this law is, it is surprisingly unintuitive. Lusser's Law tells us that any system depending on the performance of several crucial components will be *less* reliable than the least-reliable crucial component. This is akin to saying that a chain will be *weaker* than its weakest link!

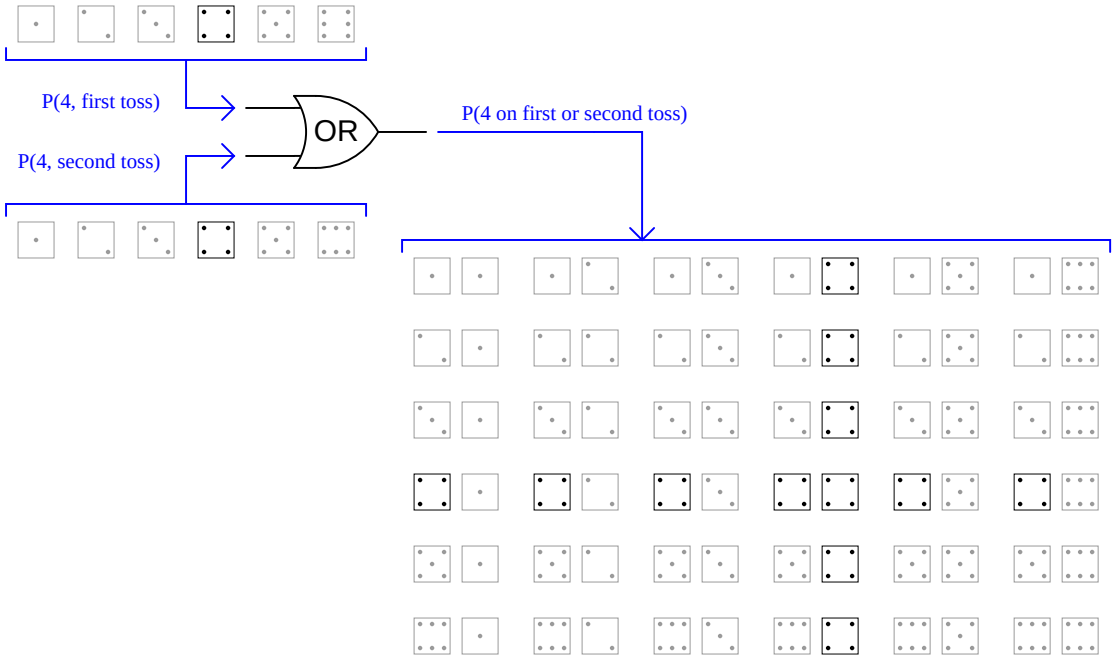
To give an illustrative example, suppose a complex system depended on the reliable operation of six key components in order to function, with the individual reliabilities of those six components being 91%, 92%, 96%, 95%, 93%, and 92%, respectively. Given individual component reliabilities all greater than 90%, one might be inclined to think the overall reliability would be quite good. However, following Lusser's Law we find the reliability of this system (as a whole) is only 65.3%.

In his excellent text *Reliability Theory and Practice*, author Igor Bazovsky recounts the German V1 missile project during World War Two, and how early assumptions of system reliability were grossly inaccurate[13]. Once these faulty assumptions of reliability were corrected, development of the V1 missile resulted in greatly increased reliability until a system reliability of 75% (three out of four) was achieved.

[13] According to Bazovsky (pp. 275-276), the first reliability principle adopted by the design team was that the system could be no more reliable than its least-reliable (weakest) component. While this is technically true, the mistake was to assume that the system would be *as reliable* as its weakest component (i.e. the "chain" would be as strong as its weakest link). This proved to be too optimistic, as the system would still fail due to the failure of "stronger" components even when the "weaker" components happened to survive. After noting the influence of "stronger" components' unreliabilities on overall system reliability, engineers somehow reached the bizarre conclusion that system reliability was equal to the mathematical *average* of the components' reliabilities. Not surprisingly, this proved even less accurate than the "weakest link" principle. Finally, the designers were assisted by the mathematician Erich Pieruschka, who helped formulate Lusser's Law.

The logical "OR" function

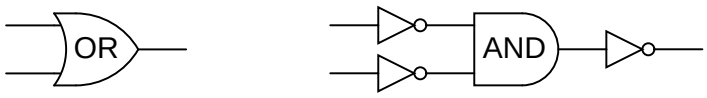
The OR function regards probabilities of two or more redundant events (i.e. where the outcome of interest happens if any one of the events happen). Another example using a die is the probability of rolling a "four" on either the first toss or on the second toss. It should be intuitively obvious that the probability of rolling a "four" on either toss will be more (i.e. more likely) than rolling a "four" on a single toss. The shaded field of possibilities (36 in all) demonstrate the likelihood of this either/or result compared to the likelihood of either value on either toss:



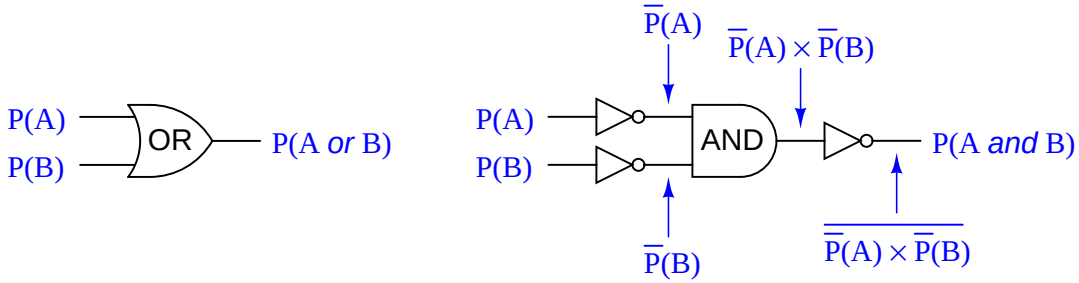
As you can see, there are eleven outcomes matching the specific criteria out of 36 total possible outcomes (the outcome with two "four" rolls counts as a single trial matching the stated criteria, just as all the other trials containing only one "four" roll count as single trials). This yields a probability value of eleven-in-thirty six ($\frac{11}{36}$) for the specified combination. This result may defy your intuition, if you assumed the OR function would be the simple *sum* of individual probabilities ($\frac{1}{6} + \frac{1}{6} = \frac{2}{6}$ or $\frac{1}{3}$), as opposed to the AND function's *product* of probabilities ($\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$). In truth, there is an application of the OR function where the probability is the simple sum, but that will come later in this presentation.

For now, a way to understand why we get a probability value of $\frac{11}{36}$ for our OR function with two $\frac{1}{6}$ input probabilities is to derive the OR function from other functions whose probability laws we already know with certainty. From Boolean algebra, DeMorgan's Theorem tells us an OR function is equivalent to an AND function with all inputs and outputs inverted ($A + B = \overline{\overline{A}\,\overline{B}}$):

(Equivalent logic functions)



We already know the complement (inversion) of a probability is the value of that probability subtracted from one ($\overline{P} = 1 - P$). This gives us a way to symbolically express the DeMorgan's Theorem definition of an OR function in terms of an AND function with three inversions:



Knowing that $\overline{P}(A) = 1 - P(A)$ and $\overline{P}(B) = 1 - P(B)$, we may substitute these inversions into the triple-inverted AND function to arrive at an expression for the OR function in simple terms of $P(A)$ and $P(B)$:

$$P(A \text{ or } B) = \overline{\overline{P}(A) \times \overline{P}(B)}$$

$$P(A \text{ or } B) = \overline{(1 - P(A))(1 - P(B))}$$

$$P(A \text{ or } B) = 1 - [(1 - P(A))(1 - P(B))]$$

Distributing terms on the right side of the equation:

$$P(A \text{ or } B) = 1 - [1 - P(B) - P(A) + P(A)P(B)]$$

$$P(A \text{ or } B) = P(B) + P(A) - P(A)P(B)$$

This, then, is our third law of probability:

$$P(\text{A } \textit{or} \text{ B}) = P(B) + P(A) - P(A) \times P(B)$$

Inserting our example probabilities of $\frac{1}{6}$ for both $P(A)$ and $P(B)$, we obtain the following probability for the OR function:

$$P(A \text{ or } B) = \frac{1}{6} + \frac{1}{6} - \left(\frac{1}{6}\right)\left(\frac{1}{6}\right)$$

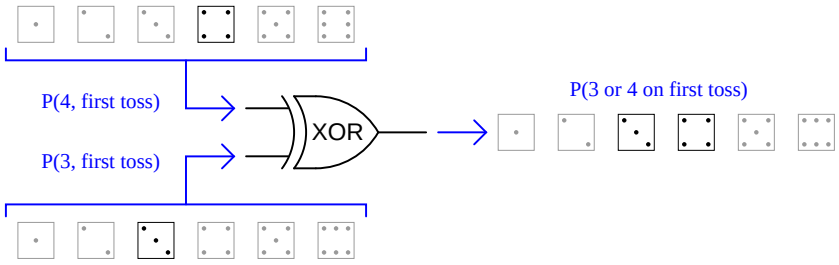
$$P(A \text{ or } B) = \frac{2}{6} - \left(\frac{1}{36}\right)$$

$$P(A \text{ or } B) = \frac{12}{36} - \frac{1}{36}$$

$$P(A \text{ or } B) = \frac{11}{36}$$

This confirms our previous conclusion of there being an $\frac{11}{36}$ probability of rolling a "four" on the first or second rolls of a die.

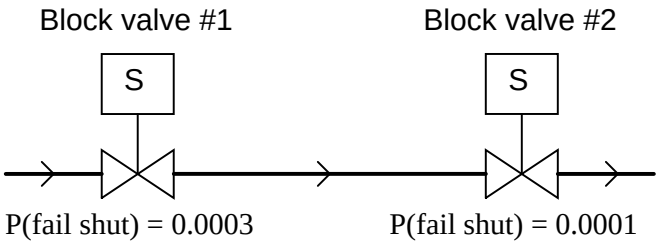
A similar application of the OR function is seen when we are dealing with *exclusive* events. For instance, we could calculate the probability of rolling either a "three" or a "four" in a single toss of a die. Unlike the previous example where we had two opportunities to roll a "four," and two sequential rolls of "four" counted as a single successful trial, here we know with certainty that the die cannot land on "three" *and* "four" in the same roll. Therefore, the exclusive OR probability (XOR) is much simpler to determine than a regular OR function:



This is the only type of scenario where the function probability is the simple sum of the input probabilities. In cases where the input probabilities are mutually exclusive (i.e. they *cannot* occur simultaneously or in a specific sequence), the probability of one *or* the other happening is the sum of the individual probabilities. This leads us to our fourth probability law:

$$P(\text{A } \textit{exclusively or} \text{ B}) = P(A) + P(B)$$

We may return to our example of a double-block valve assembly for a practical application of OR probability. When illustrating the AND probability function, we focused on the probability of both block valves failing to shut off when needed, since both valve 1 *and* valve 2 would have to fail open in order for the double-block assembly to fail in shutting off flow. Now, we will focus on the probability of *either* block valve failing to open when needed. While the AND scenario was an exploration of the system's unreliability (i.e. the probability it might fail to stop a dangerous condition), this scenario is an exploration of the system's *unavailability* (i.e. the probability it might fail to resume normal operation).



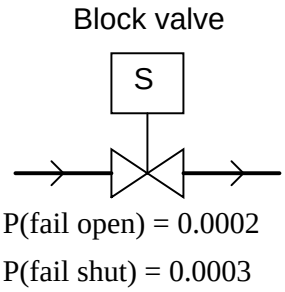
Each block valve is designed to be able to shut off flow independently, so that the flow of (potentially) dangerous process fluid will be halted if *either or both* valves shut off. The probability that process fluid flow may be impeded by the failure of either valve to open is thus a simple (non-exclusive) OR function:

$$P(\text{assembly fail}) = P(\text{valve 1 fail shut}) + P(\text{valve 2 fail shut}) - P(\text{valve 1 fail shut}) \times P(\text{valve 2 fail shut})$$

$$P(\text{assembly fail}) = 0.0003 + 0.0001 - (0.0003 \times 0.0001)$$

$$P(\text{assembly fail}) = 0.0003997 = 3.9997 \times 10^{-4}$$

A practical example of the exclusive-or (XOR) probability function may be found in the failure analysis of a single block valve. If we consider the probability this valve may fail in either condition (stuck open or stuck shut), and we have data on the probabilities of the valve failing open and failing shut, we may use the XOR function to model the system's general unreliability. We know that the exclusive-or function is the appropriate one to use here because the two "input" scenarios (failing open versus failing shut) *absolutely cannot* occur at the same time:



$$P(\text{valve fail}) = P(\text{valve fail open}) + P(\text{valve fail shut})$$

$$P(\text{valve fail}) = 0.0002 + 0.0003$$

$$P(\text{valve fail}) = 0.0005 = 5 \times 10^{-4}$$

Summary of probability laws

The complement (inversion) of a probability:

$$P(A) = 1 - P(\overline{A})$$

The probability of intersecting events (where both must happen either simultaneously or in specific sequence) for the result of interest to occur:

$$P(\text{A } \textit{and } \text{B}) = P(\text{A}) \times P(\text{B})$$

The probability of redundant events (where either or both may happen) for the result of interest to occur:

$$P(\text{A } \textit{or } \text{B}) = P(B) + P(A) - P(A) \times P(B)$$

The probability of exclusively redundant events (where either may happen, but not simultaneously or in specific sequence) for the result of interest to occur:

$$P(\text{A } \textit{exclusively or } \text{B } \textit{exclusively}) = P(A) + P(B)$$

29.2.3 Practical measures of reliability

In reliability engineering, it is important to be able to quantity the reliability (or conversely, the probability of failure) for common components, and for systems comprised of those components. As such, special terms and mathematical models have been developed to describe probability as it applies to component and system reliability.

Perhaps the first and most fundamental measure of (un)reliability is the *failure rate* of a component or system of components, symbolized by the Greek letter lambda (λ). The definition of "failure rate" for a group of components undergoing reliability tests is the instantaneous rate of failures per number of surviving components:

$$\lambda = \frac{\frac{dN_f}{dt}}{N_s} \qquad \text{or} \qquad \lambda = \frac{dN_f}{dt}\frac{1}{N_s}$$

Where,

λ = Failure rate

N_f = Number of components failed during testing period

N_s = Number of components surviving during testing period

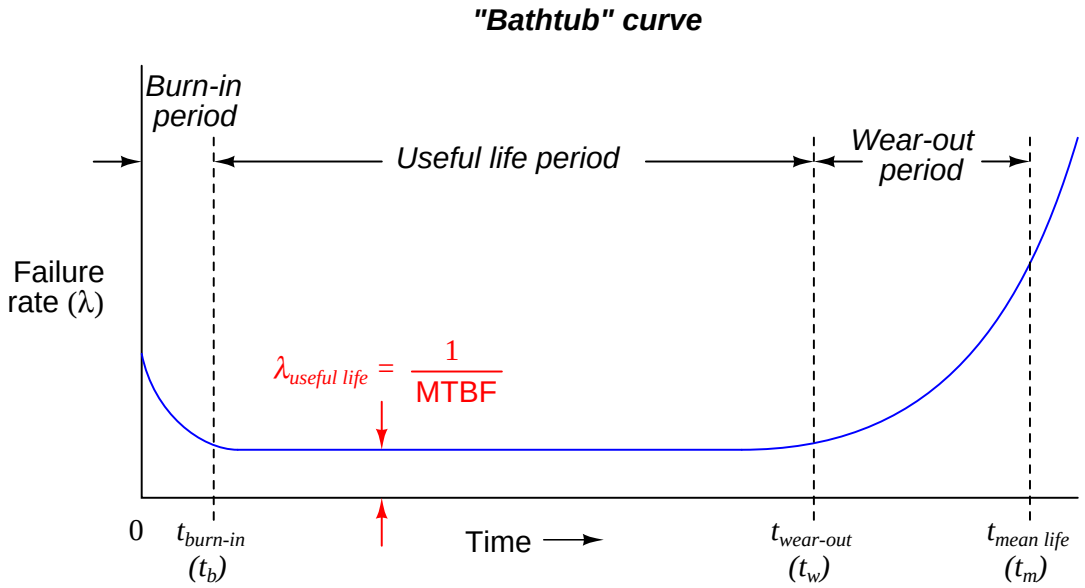
t = Time

The unit of measurement for failure rate (λ) is inverted time units (e.g. "per hour" or "per year"). An alternative expression for failure rate sometimes seen in reliability literature is the acronym *FIT* ("Failures In Time"), in units of 10^{-9} failures per hour. Using a unit with a built-in multiplier such as 10^{-9} makes it easier for human beings to manage the very small λ values normally associated with high-reliability industrial components and systems.

Failure rate may also be applied to discrete-switching (on/off) components and systems of discrete-switching components on the basis of the number of on/off cycles rather than clock time. In such cases, we define failure rate in terms of cycles (c) instead of in terms of minutes, hours, or any other measure of time (t):

$$\lambda = \frac{\frac{dN_f}{dc}}{N_s} \qquad \text{or} \qquad \lambda = \frac{dN_f}{dc}\frac{1}{N_s}$$

Failure rate may be constant, or it may be subject to change over time, depending on the type and age of a component (or system of components). A common graphical expression of failure rate is the so-called *bathtub curve* showing the typical failure rate profile over time from initial manufacture (brand-new) to wear-out:



This curve profiles the failure rate of a large sample of components (or a large sample of systems) as they age. Failure rate begins at a relatively high value starting at time zero due to defects in manufacture. Failure rate drops off rapidly during a period of time called the *burn-in period* where defective components experience an early death. After the burn-in period, failure rate remains relatively constant over the useful life of the components. Any failures occurring during the "useful life" period are due to random mishaps. Toward the end of the components' working lives when the components enter the *wear-out period*, failure rate begins to rise until all components eventually fail. The *mean (average) life* of a component (t_m) is the time required for one-half of the components surviving up until the wear-out time (t_w) to fail, the other half failing after the mean life time.

Several important features are evident in this "bathtub" curve. First, component reliability is greatest between the times of burn-in and wear-out. For this reason, many manufacturers of high-reliability components and systems perform their own burn-in testing prior to sale, so that the customers are purchasing products that have already passed the burn-in phase of their lives.

An important measure of reliability is MTBF, or *Mean Time Between Failure*. If the component or system in question is repairable, the expression *Mean Time To Failure* (MTTF) is often used instead[14]. As shown on the bathtub curve, MTBF is the reciprocal of failure rate during the useful life period. This is the period of time where failure rate is at a constant, low value, thus making MTBF a rather large value. Whereas failure rate (λ) is measured in reciprocal units of time (e.g. "per hour" or "per year"), MTBF is simply expressed in units of time (e.g. "hours" or "years").

[14]Since most high-quality industrial devices and systems are repairable for most faults, MTBF and MTTF are interchangeable terms.

Another important measure of reliability is the *mean life*. This is an expression of a component's (or system's) operating lifespan. At first this may sound synonymous with MTBF, but it is not. MTBF – and by extension the useful life failure rate, since MTBF is the reciprocal of failure rate – is an expression of susceptibility to random ("chance") failures. Both MTBF and λ_{useful} are quite independent of mean life[15]. In practice, values for MTBF often greatly exceed[16] values for mean life. When determining the length of time any component should be allowed to function in a high-reliability system, the mean life (or even better, the *wear-out* time) should be used as a guide, not the MTBF. This is not to suggest the MTBF is a useless figure – far from it. MTBF simply serves a different purpose, and that is to predict the rate of random failures *during* the useful life span of a large number of components or systems, whereas mean life predicts the service life period.

Reliability (R) is the probability that a component or system will perform as designed when needed. Like all probability figures, reliability may range in value from 0 to 1, inclusive. Given the tendency of manufactured devices to fail over time, reliability decreases with time. During the useful life of a component or system, reliability is related to failure rate by a simple exponential function:

$$R = e^{-\lambda t}$$

Where,

R = Reliability as a function of time (sometimes shown as $R(t)$

e = Euler's constant (≈ 2.71828)

λ = Failure rate (assumed to be a constant during the useful life period)

t = Time

Knowing that failure rate is the mathematical reciprocal of mean time between failures (MTBF), we may re-write this equation in terms of MTBF as a "time constant" (τ) for random failures during the useful life period:

$$R = e^{\frac{-t}{\tau}}$$

Thus, reliability exhibits the same asymptotic approach to zero over time that we would expect from a first-order decay process such as a cooling object (approaching ambient temperature) or a capacitor discharging to zero volts. A practical example of this equation in use would be the reliability calculation for a Rosemount model 1151 analog differential pressure transmitter (with a demonstrated MTBF value of 226 years as published by Rosemount) over a service life of 5 years following burn-in:

$$R = e^{\frac{-5}{226}}$$

$$R = 0.9781 = 97.81\%$$

[15] One could even imagine some theoretical component immune to wear-out, but still having finite values for failure rate and MTBF. Remember, λ_{useful} and MTBF refer to *chance* failures, not the normal failures associated with age and extended use.

[16] For example, the Rosemount model 3051C differential pressure transmitter has a suggested useful lifetime of 50 years (based on the expected service life of tantalum electrolytic capacitors used in its circuitry), while its demonstrated MTBF is 136 years.

Reliability, as previously defined, is the probability that a component or system will perform as designed when needed. Like all probability values, reliability is expressed a number ranging between 0 and 1, inclusive. A reliability value of zero (0) means the component or system is totally unreliable (i.e. it is guaranteed to fail). Conversely, a reliability value of one (1) means the component or system is completely reliable (i.e. guaranteed to properly perform when needed). The mathematical complement of reliability is referred to as *PFD*, an acronym standing for *Probability of Failure on Demand*. Like reliability, this is also a probability value ranging from 0 to 1, inclusive. A PFD value of zero (0) means there is no probability of failure (i.e. it is guaranteed to properly perform when needed), while a PFD value of one (1) means it is completely unreliable (i.e. guaranteed to fail). Thus:

$$R + \text{PFD} = 1$$

$$\text{PFD} = 1 - R$$

$$R = 1 - \text{PFD}$$

Obviously, a system designed for high reliability should exhibit a large R value (very nearly 1) and a small PFD value (very nearly 0). Just how large R needs to be (how small PFD needs to be) is a function of how critical the component or system is to the fulfillment of our human needs.

The degree to which a system must be reliable in order to fulfill our modern expectations is often surprisingly high. Suppose someone were to tell you the reliability of electric power service to a neighborhood in which you were considering purchasing a home in was 99 percent (0.99). This sounds rather good, doesn't it? However, when you actually calculate how many hours of "blackout" you would experience in a typical year given this degree of reliability, the results are seen to be rather poor (at least to modern American standards of expectation). If the reliability value for electric power in this neighborhood is 0.99, then the *unreliability* is 0.01:

$$\left(\frac{365 \text{ days}}{1 \text{ year}}\right)\left(\frac{24 \text{ hours}}{1 \text{ day}}\right)(0.01) = 87.6 \text{ hours}$$

99% doesn't look so good now, does it? Let's suppose an industrial manufacturing facility requires steady electric power service all day and every day for its continuous operation. This facility has back-up diesel generators to supply power during utility outages, but they are budgeted only for 5 hours of back-up generator operation per year. How reliable would the power service need to be in order to fulfill this facility's operational requirements? The answer may be calculated simply by determining the unreliability (PFD) of power based on 5 hours of "blackout" per year's time:

$$\text{PFD} = \frac{5 \text{ hours}}{\text{Hours in a year}} = \frac{5}{8760} = 0.00057$$

$$R = 1 - \text{PFD} = 1 - 0.00057 = 0.99943$$

Thus, the utility electric power service to this manufacturing facility must be 99.943% reliable in order to fulfill the expectations of no more than 5 hours (average) back-up generator usage per year.

A common order-of-magnitude expression of desired reliability is the number of "9" digits in the reliability value. A reliability value of 99.9% would be expressed as "three nine's" and a reliability value of 99.99% as "four nine's."

29.3 High-reliability systems

As discussed at the beginning of this chapter, instrumentation safety may be broadly divided into two categories: the safety hazards posed by malfunctioning instruments, and special instrument systems designed to reduce safety hazards of industrial processes. This section regards the first category.

All methods of reliability improvement incur some extra cost on the operation, whether it be capital expense (initial purchase/installation cost) or continuing expense (labor or consumables). The choice to improve system reliability is therefore very much an economic one. One of the human challenges associated with reliability improvement is continually justifying this cost over time. Ironically, the more successful a reliability improvement program has been, the less important that program seems. The manager of an operation suffering from reliability problems does not need to be convinced of the economic benefit of reliability improvement as much as the manager of a trouble-free facility. Furthermore, the people most aware of the benefits of reliability improvement are usually those tasked with reliability-improving duties (such as preventive maintenance), while the people least aware of the same benefits are usually those managing budgets. If ever a disagreement erupts between the two camps, pleas for continued financial support of reliability improvement programs may be seen as nothing more self-interest, further escalating tensions[17].

A variety of methods exist to improve the reliability of systems. The following subsections investigate several of them.

[17] Preventive maintenance is not the only example of such a dynamic. Modern society is filled with monetarily expensive programs and institutions existing for the ultimate purpose of avoiding *greater* costs, monetary and otherwise. Public education, health care, and national militaries are just a few that come to my mind. Not only is it a challenge to continue justifying the expense of a well-functioning cost-avoidance program, but it is also a challenge to detect and remove unnecessary expenses (waste) within that program. To extend the preventive maintenance example, an appeal by maintenance personnel to continue (or further) the maintenance budget may happen to be legitimate, but a certain degree of self-interest will always be present in the argument. Just because preventive maintenance is actually necessary to avoid greater expense due to failure, does not mean *all* preventive maintenance demands are economically justified! Proper funding of any such program depends on the financiers being fair in their judgment *and* the executors being honest in their requests. So long as both parties are human, this territory will remain contentious.

29.3.1 Design and selection for reliability

Many workable designs may exist for electronic and mechanical systems alike, but not all are equal in terms of reliability. A major factor in machine reliability, for example, is *balance*. A well-balanced machine will operate with little vibration, whereas an ill-balanced machine will tend to shake itself (and other devices mechanically coupled to it) apart over time[18].

Electronic circuit reliability is strongly influenced by design as well as by component choice. An historical example of reliability-driven design is found in the Foxboro SPEC 200 analog control system. The reliability of the SPEC 200 control system is legendary, owing to several factors. According to Foxboro technical literature, several design guidelines were developed following application experience with Foxboro electronic field instruments (most notably the "E" and "H" model lines), among them the following:

- All critical switches should spend most of their time in the *closed* state
- Avoid the use of carbon composition resistors – use wirewound or film-type resistors instead
- Avoid the use of plastic-cased semiconductors – use glass-cased or hermetically sealed instead
- Avoid the use of electrolytic capacitors wherever possible – use polycarbonate or tantalum instead

In addition to high-quality component characteristics and excellent design practices, components were "burned in" prior to circuit board assembly, thus avoiding many "early failures" due to components burning in during actual service.

[18] Sustained vibrations can do really strange things to equipment. It is not uncommon to see threaded fasteners undone slowly over time by vibrations, as well as cracks forming in what appear to be extremely strong supporting elements such as beams, pipes, etc. Vibration is almost never good for mechanical (or electrical!) equipment, so it should be eliminated wherever reliability is a concern.

29.3.2 Preventive maintenance

The term *preventive maintenance* refers to the maintenance (repair or replacement) of components prior to their inevitable failure in a system. In order to intelligently schedule the replacement of critical system components, some knowledge of those components' useful lifetimes is necessary. On the standard "bathtub curve," this corresponds with the *wear-out time* or $t_{wear-out}$.

In many industrial operations, preventive maintenance schedules (if they exist at all) are based on past history of component lifetimes, and the operational expenses incurred due to failure of those components. Preventive maintenance represents an up-front cost, paid in exchange for the avoidance of larger costs later in time.

A common example of preventive maintenance and its cost savings is the periodic replacement of lubricating oil and oil filters for automobile engines. Automobile manufacturers provide specifications for the replacement of oil and filters based on testing of their engines, and assumptions made regarding the driving habits of their customers. Some manufacturers even provide dual maintenance schedules, one for "normal" driving and another for "heavy" or "performance" driving to account for accelerated wear. As trivial as an oil change might seem to the average driver, regular maintenance to an automobile's lubrication system is absolutely critical not only to long service life, but also to optimum performance. Certainly, the consequences of not performing this preventive maintenance task on an automobile's engine will be costly[19].

Another example of preventive maintenance for increased system reliability is the regular replacement of light bulbs in traffic signal arrays. For rather obvious reasons, the proper function of traffic signal lights is critical for smooth traffic flow and public safety. It would not be a satisfactory state of affairs to replace traffic signal light bulbs only when they failed, as is common with the replacement of most light bulbs. In order to achieve high reliability, these bulbs must be replaced in advance of their expected wear-out times[20]. The cost of performing this maintenance is undeniable, but then so is the (greater) cost of congested traffic and accidents caused by burned-out traffic light bulbs.

An example of preventive maintenance in industrial instrumentation is the installation and service of *dryer* mechanisms for compressed air, used to power pneumatic instruments and valve actuators. Compressed air is a very useful medium for transferring (and storing) mechanical energy, but problems will develop within pneumatic instruments if water is allowed to collect within air distribution systems. Corrosion, blockages, and hydraulic "locking" are all potential consequences of "wet" instrument air. Consequently, instrument compressed air systems are usually installed separate from utility compressed air systems (used for operating general-purpose pneumatic tools and equipment actuators), using different types of pipe (plastic, copper, or stainless steel rather than black iron or galvanized iron) to avoid corrosion and using *air dryer* mechanisms near the compressor to absorb and expel moisture. These air dryers typically use a beaded *dessicant* material to absorb

[19] On an anecdotal note, a friend of mine once blew up his car's engine, having never performed an oil or filter change on it since the day he purchased it. His car died with only about 70,000 miles on it – a mere fraction of its normal service life with regular maintenance. Given the type of car it was, he could have easily expected 200,000 miles of service between engine rebuilds had he performed the recommended maintenance on it.

[20] Another friend of mine used to work as a traffic signal technician in a major American city. Since the light bulbs they replaced still had some service life remaining, they decided to donate the bulbs to a charity organization where the used bulbs would be freely given to low-income citizens. Incidentally, this same friend also instructed me on the proper method of inserting a new bulb into a socket: twisting the bulb just enough to maintain some spring tension on the base, rather than twisting the bulb until it will not turn further (as most people do). Maintaining some natural spring tension on the metal leaf within the socket helps further the socket's useful life as well!

water vapor from the compressed air, and then this dessicant material is periodically purged of its retained water. After some time of operation, though, the dessicant must be physically removed and replaced with fresh dessicant.

29.3.3 Component de-rating

Some[21] control system components exhibit an inverse relationship between service load (how "hard" the component is used) and service life (how long it will last). In such cases, a way to increase service life is to *de-rate* that component: operate it at a load reduced from its design rating.

For example, a variable-frequency motor drive (VFD) takes AC power at a fixed frequency and voltage and converts it into AC power of varying frequency and voltage to drive an induction motor at different speeds and torques. These electronic devices dissipate some heat owing mostly to the imperfect (slightly resistive) "on" states of power transistors. Temperature is a well-known wear factor for semiconductor devices, with greater temperatures leading to reduced service lives. A VFD operating at high temperature, therefore, will fail sooner than a VFD operating at low temperature, all other factors being equal. One way to reduce the operating temperature of a VFD is to over-size it for the application. If the motor to be driven requires 2 horsepower of electrical power at full load, and increased reliability is demanded of the drive, then perhaps a 5 horsepower VFD (programmed with reduced trip settings appropriate to the smaller motor) could be chosen to drive the motor.

In addition to extending service life, de-rating also has the ability to amplify the mean time between failure (MTBF) of load-sensitive components. Recall that MTBF is the reciprocal of failure rate during the low area of the "bathtub curve," representing failures due to random causes. This is distinct from wear-out, which is an increase in failure rate due to irreversible wear and aging. The main reason a component will exhibit a greater MTBF value as a consequence of de-rating is that the component will be better able to absorb transient overloads, which is a typical cause of failure during the operational life of a component.

Consider the example of a pressure sensor in a process known to exhibit transient pressure surges. A sensor chosen such that the typical process operating pressure spans most of its range will have little overpressure capacity. Perhaps just a few over-pressure events will cause this sensor to fail well before its rated service life. A de-rated pressure sensor (with a pressure-sensing range covering much greater pressures than what are normally encountered in this process), by comparison, will have more pressure capacity to withstand random surges, and therefore exhibit less probability of random failure.

The costs associated with component de-rating include initial investment (usually greater, owing to the greater capacity and more robust construction compared to a "normally" rated component) and reduced sensitivity. The latter factor is an important one to consider if the component is expected to provide high accuracy as well as high reliability. In the example of the de-rated pressure sensor, accuracy will likely suffer because the full pressure range of the sensor is not being used for normal process pressure measurements. If the instrument is digital, resolution will certainly suffer as a result of de-rating the instrument's measurement range. Alternative methods of reliability improvement (including more frequent preventive maintenance) may be a better solution than de-rating in such cases.

[21] Many components do not exhibit any relationship between load and lifespan. An electronic PID controller, for example, will last just as long controlling an "easy" self-regulating process as it will controlling a "difficult" unstable ("runaway") process. The same might not be said for the other components of those loops, however! If the control valve in the self-regulating process rarely changes position, but the control valve in the runaway process continually moves in an effort to stabilize it at setpoint, the less active control valve will most likely enjoy a longer service life.

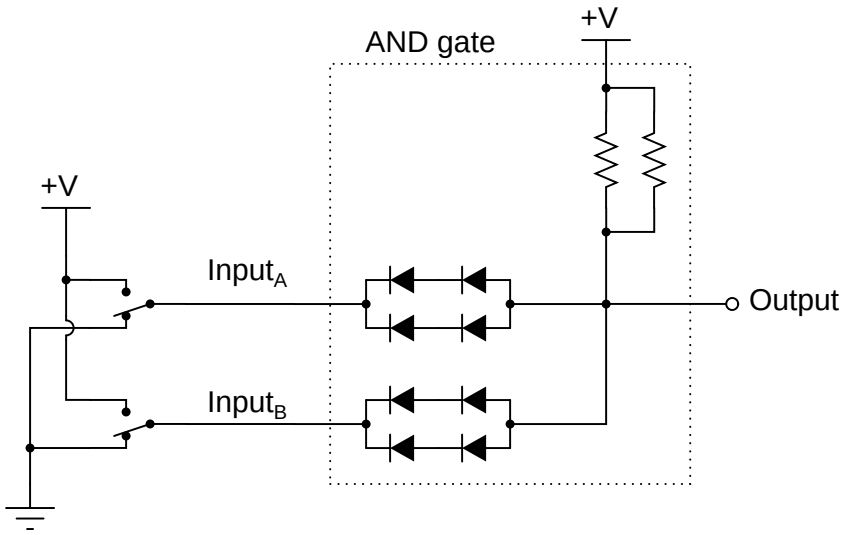
29.3.4 Redundant components

The MTBF of any system dependent upon certain critical components may be extended by duplicating those components in parallel fashion, such that the failure of only one does not compromise the system as a whole. This is called *redundancy*. A common example of component redundancy in instrumentation and control systems is the redundancy offered by distributed control systems (DCSs), where processors, network cables, and even I/O (input/output) channels may be equipped with "hot standby" duplicates ready to assume functionality in the event that the primary component fails.

Redundancy tends to extend the MTBF of a system without necessarily extending its service life. A DCS, for example, equipped with redundant microprocessor control modules in its rack, will exhibit a greater MTBF because a random microprocessor fault will be covered by the presence of the spare ("hot standby") microprocessor module. However, given the fact that both microprocessors are continually powered, and therefore tend to "wear" at the same rate, their operating lives will not be additive. In other words, two microprocessors will not function twice as long before wear-out than one microprocessor.

The extension of MTBF resulting from redundancy holds true only if the random failures are truly independent events – that is, not associated by a common cause. To use the example of a DCS rack with redundant microprocessor control modules again, the susceptibility of that rack to a random microprocessor fault will be reduced by the presence of redundant microprocessors *only* if the faults in question are unrelated to each other, affecting the two microprocessors separately. There may exist common-cause fault mechanisms capable of disabling *both* microprocessor modules as easily as it could disable one, in which case the redundancy adds no value at all. Examples of such common-cause faults include power surges (because a surge strong enough to kill one module will likely kill the other at the same time) and a computer virus infection (because a virus able to attack one will be able to attack the other just as easily, and at the same time).

A simple example of component redundancy in a larger system is the design of this passive "AND" logic gate[22], using eight diodes and four resistors:



A non-redundant version of this same passive gate circuit would require only two diodes (one on each input line) and a single pull-up resistor. The extra diodes and extra resistors are in place to provide redundancy in the event of component failures. The series-parallel configuration of the input diodes provides single-fault tolerance for either a shorted diode or an open diode. The parallel configuration of the resistors provides single-fault tolerance for open failures and zero-fault tolerance for shorted failures. Diodes in such applications exhibit a moderate tendency to fail shorted as opposed to open, while resistors exhibit a strong tendency to fail open as opposed to shorted[23]. The decision to provide shorted-fault tolerance for the diodes but not for the resistors was driven by not only the resistors' tendency to fail open, but also by the general reliability of certain resistor types (most notably wirewound) above and beyond that of diodes.

A common example of redundancy in industrial instrumentation is the use of multiple transmitters to sense the same process variable, the notion being that the critical process variable will still be monitored even in the event of a transmitter failure. Thus, installing redundant transmitters should increase the MTBF of the system's sensing ability.

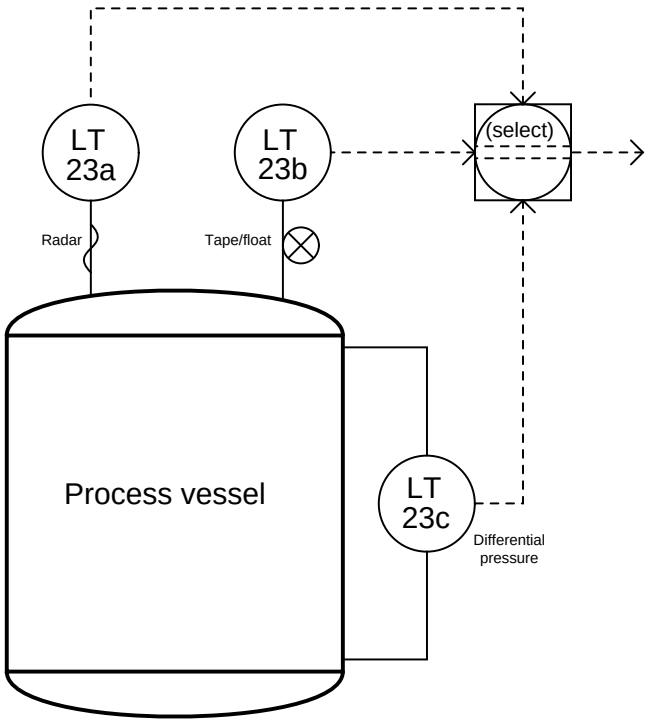
However, the problem of common-cause failures must be considered for this approach to be valid. If three liquid level transmitters are installed to measure the exact same liquid level, their combined

[22] This is an AND gate because any "low" input forces a "low output." Conversely, only when all inputs are "high" can the output rise to a "high" logic state as well.

[23] These statistics come from data found in the 1997 *Failure Mode / Mechanism Distributions* report of the Reliability Analysis Center. Shorted-to-Open failure distributions of switching diodes were reported as 40.9% to 9.1% (nearly 4.5 times more likely to fail shorted than to fail open!). Rectifier diodes exhibited a shorted failure incidence almost twice that of failing open (21.2% versus 11.6%). Interestingly, microwave and small-signal (analog application) diodes were just the opposite: more open faults than shorted faults. Resistors of all types displayed a strong tendency to fail open as opposed to shorted: fixed film resistors failed 37.5% open to 5.0% shorted; fixed wirewound resistors 26.7% open to 3.1% shorted.

signals represent an increase in measurement system MTBF *only* for independent faults. A failure mechanism common to all three transmitters will leave the system just as vulnerable to random failure as a single transmitter. In order to achieve optimum MTBF in redundant sensor arrays, the sensors must be immune to common faults.

In this example, three different types of level transmitter monitor the level of liquid inside a vessel, their signals processed by a *selector* function programmed inside a DCS:



Here, level transmitter 23a is a guided-wave radar (GWR), level transmitter 23b is a tape-and-float, and level transmitter 23c is a differential pressure sensor. All three level transmitters sense liquid level using different technologies, each one with its own strengths and weaknesses. Better redundancy of measurement is obtained this way, since no single process condition or other random event is likely to fault more than one of the transmitters at any given time.

For instance, if the process liquid density happened to suddenly change, it would affect the measurement accuracy of the differential pressure transmitter (LT-23c), but not the radar transmitter nor the tape-and-float transmitter. If the process vapor density were to suddenly change, it might affect the radar transmitter (since vapor density generally affects dielectric constant, and dielectric constant affects the propagation velocity of radio waves, which in turn will affect the time taken for the radar pulse to strike the liquid surface and return), but this will not affect the float transmitter's accuracy nor will it affect the differential pressure transmitter's accuracy. Surface turbulence of the liquid inside the vessel may severely affect the float transmitter's ability to accurately sense liquid level, but it will have little effect on the differential pressure transmitter's reading nor the radar transmitter's measurement (assuming the radar transmitter is shrouded in a *stilling well*.

If the selector function takes either the median (middle) measurement or an average of the best 2-out-of-3, none of these random process occurrences will greatly affect the selected measurement of liquid level inside the vessel. True redundancy is achieved here, since the three level transmitters

are not only less likely to (all) fail simultaneously than for any single transmitter to fail, but also because the level is being sensed in three completely different ways.

A crucial requirement for redundancy to be effective is that all redundant components must have precisely the same process function. In the case of redundant DCS components such as processors, I/O cards, and network cables, each of these redundant components must do nothing more than serve as "backup" spares for their primary counterparts. If a particular DCS node were equipped with two processors – one as the primary and another as a secondary (backup) – but yet the backup processor were tasked with some detail specific to it and not to the primary processor (or visa-versa), the two processors would *not* be truly redundant to each other. If one processor were to fail, the other would not perform *exactly* the same function, and so the system's operation would be affected (even if only in a small way) by the processor failure.

Likewise, redundant sensors must perform the exact same process measurement function in order to be truly redundant. A process equipped with triplicate measurement transmitters such as the previous example were a vessel's liquid level was being measured by a guided-wave radar, tape-and-float, and differential pressure based level transmitters, would enjoy the protection of redundancy if and only if all three transmitters sensed the exact same liquid level over the exact same calibrated range. This often represents a challenge, in finding suitable locations on the process vessel for three different instruments to sense the exact same process variable. Quite often, the pipe fittings penetrating the vessel (often called *nozzles*) are not conveniently located to accept multiple instruments at the points necessary to ensure consistency of measurement between them. This is often the case when an existing process vessel is retrofitted with redundant process transmitters. New construction is usually less of a problem, since the necessary nozzles and other accessories may be placed in their proper positions during the design stage[24].

If fluid flow conditions inside a process vessel are excessively turbulent, multiple sensors installed to measure the same variable will sometimes report significant differences. Multiple temperature transmitters located in close proximity to each other on a distillation column, for example, may report significant differences of temperature if their respective sensing elements (thermocouples, RTDs) contact the process liquid or vapor at points where the flow patterns vary. Multiple liquid level sensors, even of the same technology, may report differences in liquid level if the liquid inside the vessel swirls or "funnels" as it enters and exits the vessel.

Not only will substantial measurement differences between redundant transmitters compromise their ability to function as "backup" devices in the event of a failure, such differences may actually "fool" a redundant system into thinking one or more of the transmitters has already failed, thereby causing the deviating measurement to be ignored. To use the triplicate level-sensing array as an example again, suppose the radar-based level transmitter happened to register two inches greater level than the other two transmitters due to the effects[25] of liquid swirl inside the vessel. If the selector function is programmed to ignore such deviating measurements, the system degrades to a duplicate-redundant instead of triplicate-redundant array. In the event of a dangerously low liquid level, for example, only the radar-based and float-based level transmitters will be ready to signal

[24] Of course, this assumes good communication and proper planning between all parties involved. It is not uncommon for piping engineers and instrument engineers to mis-communicate during the crucial stages of process vessel design, so that the vessel turns out not to be configured as needed for redundant instruments.

[25] If a swirling fluid inside the vessel encounters a stationary baffle, it will tend to "pile up" on one side of that baffle, causing the liquid level to actually be greater in that region of the vessel than anywhere else inside the vessel. Any transmitter placed within this region will register a greater level, regardless of the measurement technology used.

this dangerous process condition to the control system, because the pressure-based level transmitter is registering too high.

29.3.5 Proof tests and self-diagnostics

A reliability enhancing technique related to preventive maintenance of critical instruments and functions, but generally not as expensive as component replacement, is periodic *testing* of component and system function. Regular "proof testing" of critical components enhances the MTBF of a system through two different means:

- Early detection of developing problems
- Regular "exercise" of components

First, proof testing may reveal weaknesses developing in components, indicating the need for replacement in the near future. This is sometimes referred to as *predictive maintenance*, because the results of the testing serve to predict impending failure.

The second way proof testing increases system reliability is by realizing the beneficial effects of regular function. The performance of many component and system types tends to degrade after prolonged periods of inactivity[26]. This tendency is most prevalent in mechanical systems, but holds true for some electrical components and systems as well. Solenoid valves, for instance, may become "stuck" in place if not cycled for long periods of time. Bearings may corrode and seize in place if left immobile. Both primary- and secondary-cell batteries are well known for their tendency to fail after prolonged periods of non-use. Regular cycling of such components actually *enhances* their reliability, decreasing the probability of a "stagnation" related failure well before the rated useful life has elapsed.

An important part of any proof-testing program is to ensure a ready stock of spare components is kept on hand in the event proof-testing reveals a failed component. Proof testing is of little value if the failed component cannot be immediately repaired or replaced, and so these warehoused components should be configured (or be easily configurable) with the exact parameters necessary for immediate installation. A common tendency in business is to focus attention on the engineering and installation of process and control systems, but neglect to invest in the support materials and infrastructure to keep those systems in excellent condition. High-reliability systems have special needs, and this is one of them.

[26] The father of a certain friend of mine has operated a used automobile business for many years. One of the tasks given to this friend when he was a young man, growing up helping his father in his business, was to regularly drive some of the cars on the lot which had not been driven for some time. If an automobile is left un-operated for many weeks, there is a marked tendency for batteries to fail and tires to lose their air pressure, among other things. The salespeople at this used car business jokingly referred to this as *lot rot*, and the only preventive measure was to routinely drive the cars so they would not "rot" in stagnation. Machines, like people, suffer if subjected to a lack of physical activity.

Methods of proof testing

The most direct method of testing a critical system is to stimulate it to its range limits and observe its reaction. For a process transmitter, this sort of test usually takes the form of a full-range calibration check. For a controller, proof testing would consist of driving all input signals through their respective ranges in all combinations to check for the appropriate output response(s). For a final control element (such as a control valve), this requires full stroking of the element, coupled with physical leakage tests (or other assessments) to ensure the element is having the intended effect on the process.

An obvious challenge to proof testing is how to perform such comprehensive tests without disrupting the process in which it functions. Proof-testing an out-of-service instrument is a simple matter, but proof-testing an instrument installed in a working system is something else entirely. How can transmitters, controllers, and final control elements be manipulated through their entire operating ranges without actually disturbing (best case) or halting (worst case) the process? Even if all tests may be performed at the required intervals during shut-down periods, the tests are not as realistic as they could be with the process operating at typical pressures and temperatures. Proof-testing components during actual "run" conditions is the most realistic way to assess their readiness.

One way to proof-test critical instruments with minimal impact to the continued operation of a process is to perform the tests on only some components, not all. For instance, it is a relatively simple matter to take a transmitter out of service in an operating process to check its response to stimuli: simply place the controller in manual mode and let a human operator control the process manually while an instrument technician tests the transmitter. While this strategy admittedly is not comprehensive, at least proof-testing some of the instruments is better than proof-testing none of them.

Another method of proof-testing is to "test to shutdown:" choose a time when operations personnel plan on shutting the process down anyway, then use that time as an opportunity to proof-test one or more critical component(s) necessary for the system to run. This method enjoys the greatest degree of realism, while avoiding the inconvenience and expense of an unnecessary process interruption.

Yet another method to perform proof tests on critical instrumentation is to accelerate the speed of the testing stimuli so that the final control elements will not react fully enough to actually disrupt the process, but yet will adequately assess the responsiveness of all (or most) of the components in question. The nuclear power industry sometimes uses this proof-test technique, by applying high-speed pulse signals to safety shutdown sensors in order to test the proper operation of shutdown logic, without actually shutting the reactor down. The test consists of injecting short-duration pulse signals at the sensor level, then monitoring the output of the shutdown logic to ensure consequent pulse signals are sent to the shutdown device(s). Various chemical and petroleum industries apply a similar proof-testing technique to safety valves called *partial stroke testing*, whereby the valve is stroked only part of its travel: enough to ensure the valve is capable of adequate motion without closing (or opening, depending on the valve function) enough to actually disrupt the process.

Redundant systems offer unique benefits and challenges to component proof-testing. The benefit of a redundant system in this regard is that any one redundant component may be removed from service for testing without any special action by operations personnel. Unlike a "simplex" system where removal of an instrument requires a human operator to manually take over control during the

duration of the test, the "backup" components of a redundant system should do this automatically, theoretically making the test much easier to conduct. However, the challenge of doing this is the fact that the portion of the system responsible for ensuring seamless transition in the event of a failure is in fact a component liable to failure itself. The only way to test this component is to actually disable one (or more, in highly redundant configurations) of the redundant components to see whether or not the remaining component(s) perform their redundant roles. So, proof-testing a redundant system harbors no danger if all components of the system are good, but risks process disruption if there happens to be an undetected fault.

Let us return to our triplicate level transmitter system once again to explore these concepts. Suppose we wished to perform a proof-test of the pressure-based level transmitter. Being one of three transmitters measuring liquid level in this vessel, we should be able to remove it from service with no preparation (other than notifying operations personnel of the test, and of the potential consequences) since the selector function should automatically de-select the disabled transmitter and continue measuring the process via the remaining two transmitters. If the proof-testing is successful, it proves not only that the transmitter works, but also that the selector function adequately performed its task in "backing up" the tested transmitter while it was removed. However, if the selector function happened to be failed when we disable the one level transmitter for proof-testing, the selected process level signal could register a faulty value instead of switching to the two remaining transmitters' signals. This might disrupt the process, especially if the selected level signal went to a control loop or to an automatic shutdown system. We could, of course, proceed with the utmost caution by having operations personnel place the control system in "manual" mode while we remove that one transmitter from service, just in case the redundancy does not function as designed. Doing so, however, fails to fully test the system's redundancy, since by placing the system in manual mode before the test we do not allow the redundant logic to fully function as it would be expected to in the event of an actual instrument failure.

Regular proof-testing is an essential activity to realize optimum reliability for any critical system. However, in all proof-testing we are faced with a choice: either test the components to their fullest degree, in their normal operating modes, and risk (or perhaps guarantee) a process disruption; or perform a test that is less than comprehensive, but with less (or no) risk of process disruption. In the vast majority of cases, the latter option is chosen simply due to the costs associated with process disruption. Our challenge as instrumentation professionals is to formulate proof tests that are as comprehensive as possible while being the least disruptive to the process we are trying to regulate.

Instrument self-diagnostics

One of the great advantages of digital electronic technology in industrial instrumentation is the inclusion of *self-diagnostic* ability in field instruments. A "smart" instrument containing its own microprocessor may be programmed to detect certain conditions known to indicate sensor failure or other problems, then signal the control system that something is wrong. Though self-diagnostics can never be perfectly effective in that there will inevitably be cases of undetected faults and even false positives (declarations of a fault where none exists), the current state of affairs is considerably better than the days of purely analog technology where instruments possessed little or no self-diagnostic capability.

Digital field instruments have the ability to communicate self-diagnostic error messages to their host systems over the same "fieldbus" networks they use to communicate regular process data. FOUNDATION Fieldbus instruments in particular have extensive error-reporting capability, including a "status" variable associated with every process signal that propagates down through all function blocks responsible for control of the process. Detected faults are efficiently communicated throughout the information chain in the system when instruments have full digital communication ability.

"Smart" instruments with self-diagnostic ability but limited to analog (e.g. 4-20 mA DC) signaling may also convey error information, just not as readily or as comprehensively as a fully digital instrument. The NAMUR recommendations for 4-20 mA signaling (NE-43) provide a means to do this:

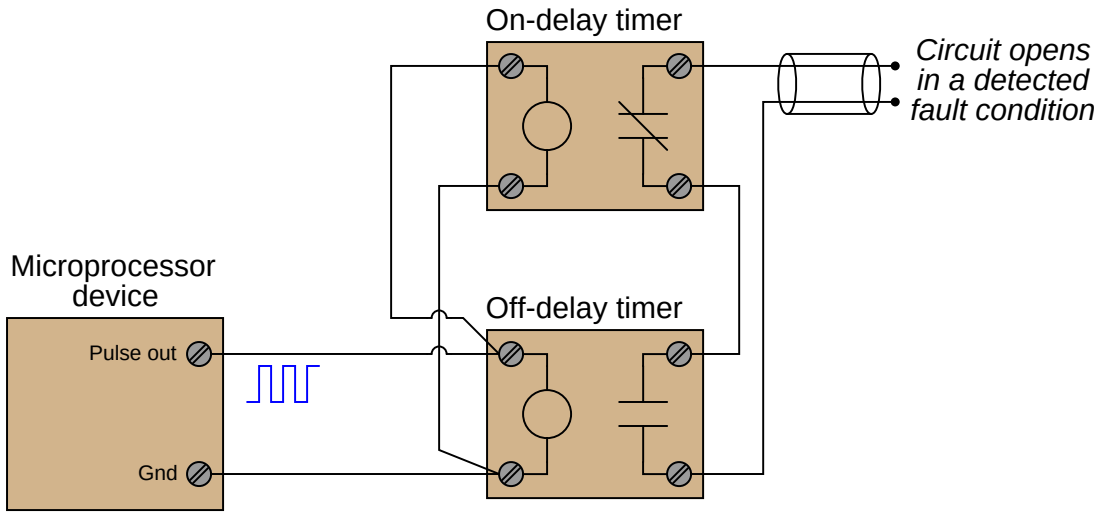
Signal level	**Fault condition**
Output $\leq$ 3.6 mA	Sensing transducer failed low
3.6 mA $<$ Output $<$ 3.8 mA	Sensing transducer failed (detected) low
3.8 mA $\leq$ Output $<$ 4.0 mA	Measurement under-range
21.0 $>$ Output $\geq$ 20.5 mA	Measurement over-range
Output $\geq$ 21.0 mA	Sensing transducer failed high

Proper interpretation of these special current ranges, of course, demands a receiver capable of accurate current measurement outside the standard 4-20 mA range. Many control systems with analog input capability are programmed to recognize the NAMUR error-indicating current levels.

A challenge for any self-diagnostic system is how to check for faults in the "brain" of the unit itself: the microprocessor. If a failure occurs within the microprocessor of a "smart" instrument – the very component responsible for performing logic functions related to self-diagnostic testing – how would it be able to detect a fault in logic? The question is somewhat philosophical, equivalent to determining whether or not a neurologist is able to diagnose his or her own neurological problems.

One simple method of detecting gross faults in a microprocessor system is known as a *watchdog timer*. The principle works like this: the microprocessor is programmed to output continuous a low-frequency pulse signal, with an external circuit "watching" that pulse signal for any interruptions or freezing. If the microprocessor fails in any significant way, the pulse signal will either skip pulses or "freeze" in either the high or low state, thus indicating a microprocessor failure to the "watchdog" circuit.

One may construct a watchdog timer circuit using a pair of solid-state timing relays connected to the pulse output channel of the microprocessor device:



Both the on-delay and off-delay timers receive the same pulse signal from the microprocessor, their inputs connected directly in parallel with the microprocessor's pulse output. The off-delay timer immediately actuates upon receiving a "high" signal, and begins to time when the pulse signal goes "low." The on-delay timer begins to time during a "high" signal, but immediately de-actuates whenever the pulse signal goes "low." So long as the time settings for the on-delay and off-delay timer relays are greater than the "high" and "low" durations of the watchdog pulse signal, respectively, neither relay contact will open as long as the pulse signal continues in its regular pattern.

When the microprocessor is behaving normally, outputting a regular watchdog pulse signal, the off-delay timer's contact will hold in a closed state because it keeps getting energized with each "high" signal and never has enough time to drop out during each "low" signal. Likewise, the on-delay timer's contact will remain in its normally closed state because it never has enough time to pick up during each "high" signal before being de-actuated with each "low" signal. Both timing relay contacts will be in a closed state when all is well.

However, if the microprocessor's pulse output signal happens to freeze in the "low" state (or skip a "high" pulse), the off-delay timer will de-actuate, opening its contact and signaling a fault. Conversely, if the microprocessor's pulse signal happens to freeze in the "high" state (or skip a "low" pulse), the on-delay timer will actuate, opening its contact and signaling a fault. Either timing relay opening its contact signals an interruption or cessation of the watchdog pulse signal, indicating a serious microprocessor fault.

29.4 Safety Instrumented Functions and Systems

A *Safety Instrumented Function*, or *SIF*, is one or more components designed to execute a specific safety-related task in the event of a specific dangerous condition. The over-temperature shutdown switch inside a clothes dryer or an electric water heater is a simple, domestic example of an SIF, shutting off the source of energy to the appliance in the event of a detected over-temperature condition. Safety Instrumented Functions are alternatively referred to as *Instrument Protective Functions*, or *IPF*s.

A *Safety Instrumented System*, or *SIS*, is a collection of SIFs designed to bring an industrial process to a safe condition in the event of any one of multiple dangerous detected conditions. Also known as *Emergency Shutdown* (ESD) or *Protective Instrument Systems* (PIS), these systems serve as an additional "layer" of protection against process equipment damage, adverse environmental impact, and/or human injury beyond the protection normally offered by a properly operating regulatory control system.

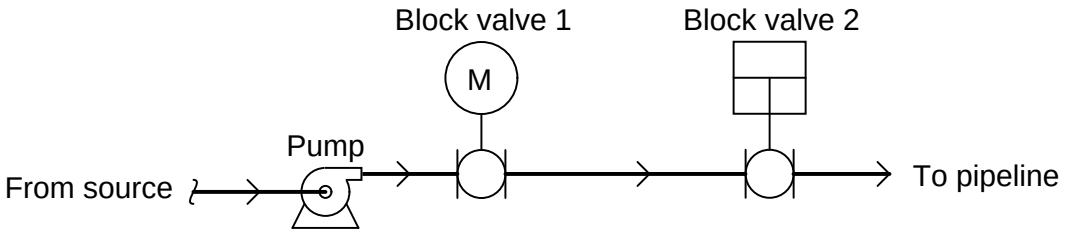
Some industries, such as chemical processing and nuclear power, have extensively employed safety instrumented systems for many decades. Likewise, automatic shutdown controls have been standard on steam boilers and combustion furnaces for years. The increasing capability of modern instrumentation, coupled with the realization of enormous costs (both social and fiscal) resulting from industrial disasters has pushed safety instrumentation to new levels of sophistication and new breadths of application. It is the purpose of this section to explore some common safety instrumented system concepts as well as some specific industrial applications.

One of the challenges inherent to safety instrumented system design is to balance the goal of maximum safety against the goal of maximum economy. If an industrial manufacturing facility is equipped with enough sensors and layered safety shutdown systems to virtually ensure no unsafe condition will ever prevail, that same facility will be plagued by "false alarm" and "spurious trip" events[27] where the safety systems malfunction in a manner detrimental to the profitable operation of the facility. In other words, a process system designed with an emphasis on automatic shut-down will probably shut down more frequently than it actually needs to. While the avoidance of unsafe process conditions is obviously a noble goal, it cannot come at the expense of economically practical operation or else there will be no reason for the facility to exist at all[28]. A safety system must provide *reliability* in its intended protective function, but not at the expense of minimizing the operational *availability* of the process itself.

[27] Many synonyms exist to describe the action of a safety system needlessly shutting down a process. The term "nuisance trip" is often (aptly) used to describe such events. Another (more charitable) label is "fail-to-safe," meaning the failure brings the process to a safe condition, as opposed to a dangerous condition.

[28] Of course, there do exist industrial facilities operating at a financial loss for the greater public benefit (e.g. certain waste processing operations), but these are the exception rather than the rule. It is obviously the point of a *business* to turn a profit, and so the vast majority of industries simply cannot sustain a philosophy of safety at *any* cost. One could argue that a "paranoid" safety system even at a waste processing plant is unsustainable, because too many "false trips" result in inefficient processing of the waste, posing a greater public health threat the longer it remains unprocessed.

To illustrate the tension between reliability and availability in a safety system, we may analyze a double-block shutoff valve[29] system for a petroleum pipeline:



The safety function of these block valves is, of course, to shut off flow from the petroleum source to the distribution pipeline in the event that the pipeline suffers a leak or rupture. Having two block valves in "series" adds an additional layer of safety, in that only one of the block valves need shut to fulfill the safety (reliability) function. Note the use of two different valve actuator technologies: one electric (motor) and the other a piston (either pneumatic or hydraulically actuated). This diversity of actuator technologies helps avoid common-cause failures, helping to ensure both valves will not simultaneously fail due to a single cause.

However, the typical operation of the pipeline demands both block valves be open in order for petroleum to flow through. The presence of redundant (dual) block valves, while increasing safety, decrease operational availability for the pipeline. If *either* of the two block valves happened to fail shut when they were called to open, the pipeline would be needlessly shut down.

A precise method of quantifying reliability and availability for redundant systems is to label the system according to how many redundant elements need to function properly in order to achieve the desired result. If the desired result for our double-block valve array is to shut down the pipeline in the event of a detected leak or rupture, we would say the system is *one out of two* (1oo2) redundant for safety reliability. In other words, only one out of the two redundant valves needs to function properly (shut off) in order to bring the pipeline to a safe condition. If the desired result is to open flow to the pipeline when it is known the pipeline is leak-free, we would say the system is *two out of two* (2oo2) redundant for operational availability. This means *both* of the two block valves need to function properly (open up) in order to allow petroleum to flow through the pipeline.

This numerical notation showing the number of essential elements versus number of total elements is often referred to as *MooN* ("*M* out of *N*") notation, or sometimes as *NooM* ("*N* out of *M*") notation[30].

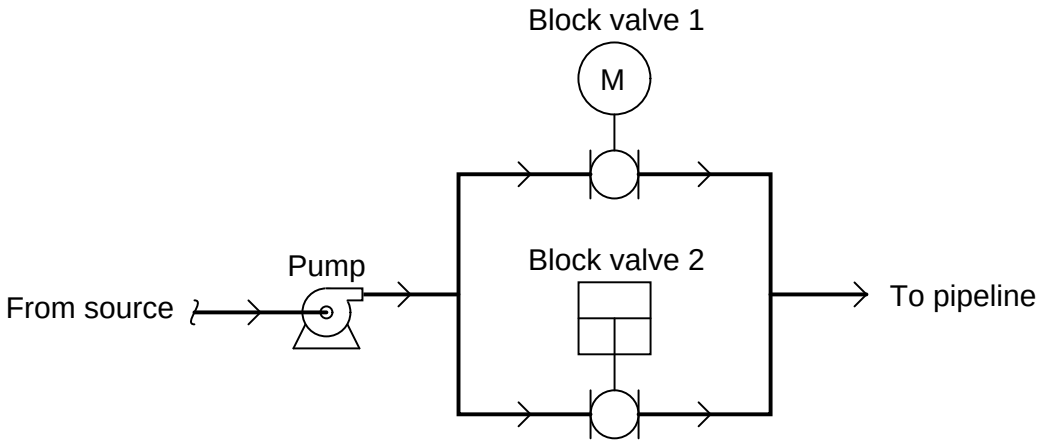
A complementary method of quantifying reliability and availability for redundant systems is to label in terms of how many element failures the system may sustain while still achieving the desired result. For this series set of double block valves, the safety (shutdown) function has a *fault tolerance* of one (1), since one of the valves may fail to shut when called upon but the other valve remains

[29] As drawn, these valves happen to be ball-design, the first actuated by an electric motor and the second actuated by a pneumatic piston. As is often the case with redundant instruments, an effort is made to diversify the technology applied to the redundant elements in order to minimize the probability of common-cause failures. If both block valves were electrically actuated, a failure of the electric power supply would disable both valves. If both block valves were pneumatically actuated, a failure of the compressed air supply would disable both valves. The use of one electric valve and one pneumatic valve grants greater independence of operation to the double-block valve system.

[30] For what it's worth, the ISA safety standard 84 defines this notation as "MooN," but I have seen sufficient examples of the contrary ("NooM") to question the authority of either label.

sufficient in itself to shut off the flow of petroleum to the pipeline. The operational availability of the system, however, has a fault tolerance of zero (0). Both block valves must open up when called upon in order to establish flow through the pipeline.

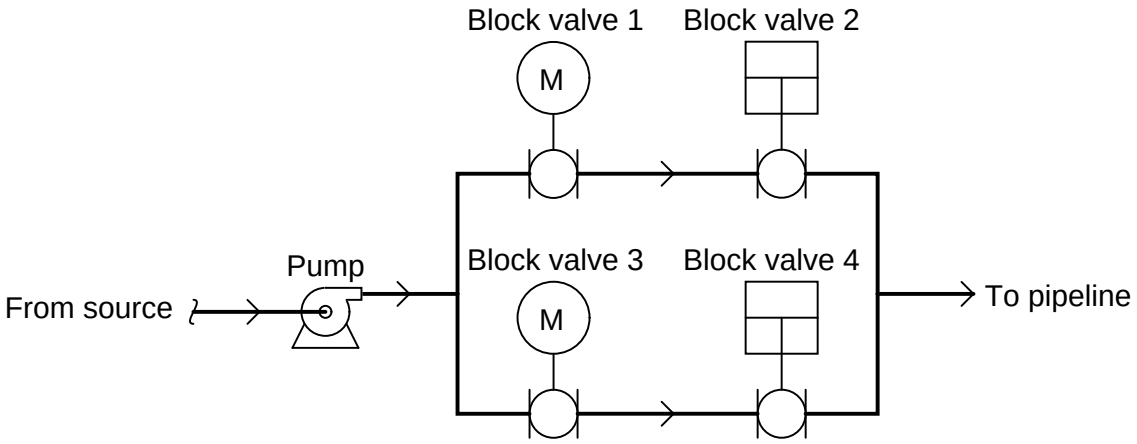
It should be clearly evident that a series set of block valves emphasizes safety (the ability to shut off flow through the pipeline) at the expense of availability (the ability to allow flow through the pipeline). We may now analyze a parallel block valve scheme to compare its redundant characteristics:



In this system, the safety (reliability) redundancy function is 2oo2, since *both* block valves would have to shut off in order to bring the pipeline to a safe condition in the event of a detected pipeline leak. However, operational availability would be 1oo2, since only one of the two valves would have to open up in order to establish flow through the pipeline. Thus, a parallel block valve array emphasizes availability (the ability to allow flow through the pipeline) at the expense of safety (the ability to shut off flow through the pipeline).

Another way to express the redundant behavior of the parallel block valve array is to say that the safety reliability function has a fault tolerance of zero (0), while the operational availability function has a fault tolerance of one (1).

One way to increase the fault tolerance of a redundant system is to increase the number of redundant components, forming arrays of greater complexity. Consider this quadruple block valve array, designed to serve the same function on a petroleum pipeline:



In order to fulfill its safety function of shutting off the flow of petroleum to the pipeline, both parallel pipe "branches" must be shut off. At first, this might seem to indicate a two-out-of-four (2oo4) redundancy, because all we would need is for one valve in each branch (two valves total) out of the four valves to shut off in order to shut off flow to the pipeline. We must remember, however, that we do not have the luxury of assuming idealized faults. If only two of the four valves function properly in shutting off, they just might happen to be two valves *in the same branch*, in which case two valves properly functioning is not enough to guarantee a safe pipeline condition. Thus, this redundant system actually exhibits *three*-out-of-four (3oo4) redundancy for safety (i.e. it has a safety fault tolerance of one), because we need three out of the four block valves to properly shut off in order to *guarantee* a safe pipeline condition.

Analyzing this quadruple block valve array for operational availability, we see that three out of the four valves need to function properly (open up) in order to guarantee flow to the pipeline. Once again, it may appear at first as though all we need are two of the four valves to open up in order to establish flow to the pipeline, but this will not be enough if those two valves are in different parallel branches. So, this system exhibits three-out-of-four (3oo4) redundancy with respect to operational availability (i.e. it has an operational fault tolerance of one).

29.4.1 SIS sensors

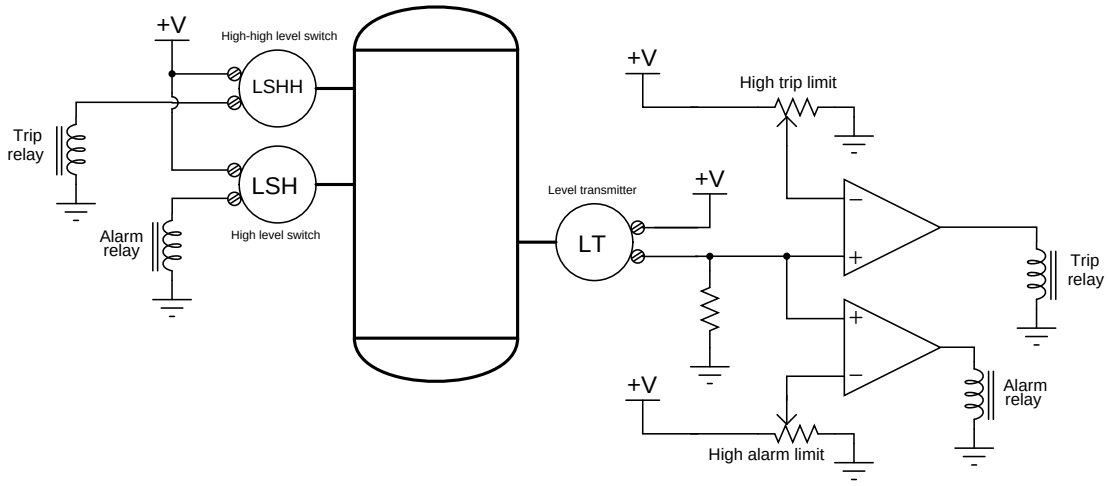
Perhaps the simplest form of sensor providing process information for a safety instrumented function is a *process switch*. Examples of process switches include temperature switches, pressure switches, level switches, and flow switches[31]. SIS sensors must be properly calibrated and configured to indicate the presence of a dangerous condition. They must be separate and distinct from the sensors used for regulatory control, in order to ensure a level of safety protection beyond that of the basic process control system.

Referring to the clothes dryer and domestic water heater over-temperature shutdown switches, these high-temperature shutdown sensors are distinctly separate from the regulatory (temperature-controlling) sensors used to maintain the appliance's temperature at setpoint. As such, they should only ever spring into action in the event of a high-temperature *failure* of the basic control system. That is, the over-temperature safety switch on a clothes dryer or a water heater should only ever reach its high-temperature limit if the normal temperature control system of the appliance fails to do its job of regulating temperature to normal levels.

A modern trend in safety instrumented systems is to use continuous process transmitters rather than discrete process switches to detect dangerous process conditions. Any process transmitter – analog or digital – may be used as a safety shutdown sensor if its signal is compared against a "trip" limit value by a comparator relay or function block. This comparator function provides an on-or-off (discrete) output based on the transmitter's signal value relative to the trip point.

[31] For a general introduction to process switches, refer to chapter 9 beginning on page 367.

A simplified example of a continuous transmitter used as a discrete alarm and trip device is shown here, where analog comparators generate discrete "trip" and "alarm" signals based on the measured value of liquid in a vessel. Note the necessity of *two* level switches on the other side of the vessel to perform the same dual alarm and trip functions:



Benefits to using a continuous transmitter instead of discrete switches include the ability to easily change the alarm or trip value, and better diagnostic capability. The latter point is not as obvious as the former, and deserves more explanation. A transmitter continuously measuring liquid level will produce an output signal that varies over time with the measured process variable. A "healthy" transmitter should therefore exhibit a continuously changing output signal, proportional to the degree of change in the process. Discrete process switches, in contrast to transmitters, provide no indication of "healthy" operation. The only time a process switch should ever change states is when its trip limit is reached, which in the case of a safety shutdown sensor indicates a dangerous (rare) condition. A process switch showing a "normal" process variable may indeed be functional and indicating properly, but it might also be failed and incapable of registering a dangerous condition should one arise – there is no way to tell by monitoring its un-changing status. The continuously varying output of a process transmitter therefore serves as an indicator[32] of proper function.

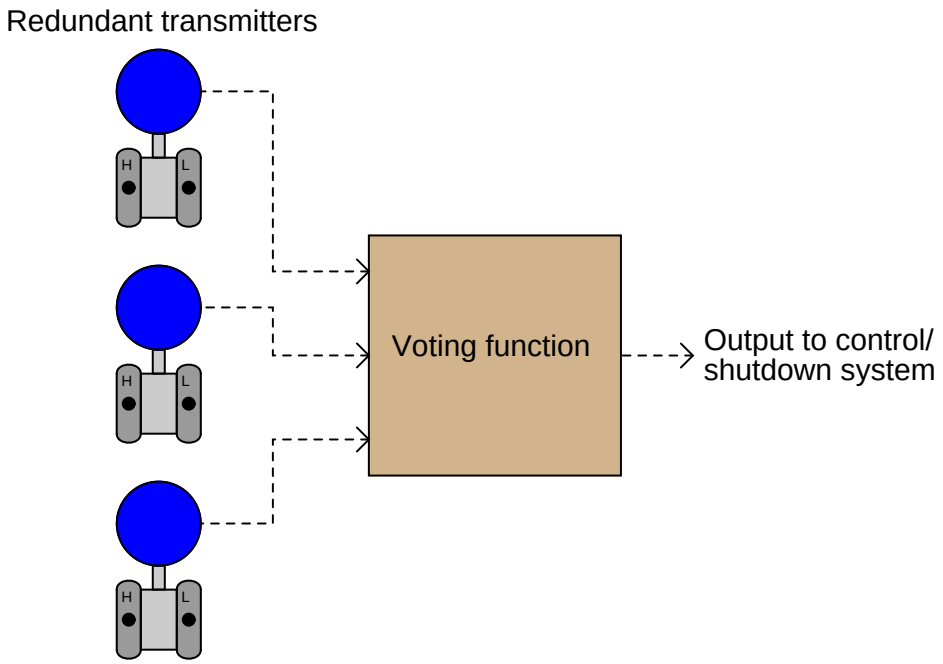
[32]Of course, the presence of some variation in a transmitter's output over time is no guarantee of proper operation. Some failures may cause a transmitter to output a randomly "walking" signal when in fact it is not registering the process at all. However, being able to measure the continuous output of a process transmitter provides the instrument technician with far more data than is available with a discrete process switch. A safety transmitter's output signal may be correlated against the output signal of another transmitter measuring the same process variable, perhaps even the transmitter used in the regulatory control loop. If two transmitters measuring the same process variable agree closely with one another over time, chances are extremely good are both functioning properly.

In applications where Safety Instrumented Function (SIF) reliability is paramount, redundant transmitters may be installed to yield additional reliability. The following photograph shows triple-redundant transmitters measuring liquid flow by sensing differential pressure dropped across an orifice plate:

A single orifice plate develops the pressure drop, with the three differential pressure transmitters "tubed" in parallel with each other, all the "high" side ports connected together through common[33] impulse tubing and all the "low" side ports connected together through common impulse tubing. These particular transmitters happen to be FOUNDATION Fieldbus rather than 4-20 mA analog electronic. The yellow instrument tray cable (ITC) used to connect each transmitter to a segment coupling device may be clearly seen in this photograph.

[33]It should be noted that the use of a single orifice plate and of common (parallel-connected) impulse lines represents a point of common-cause failure. A blockage at one or more of the orifice plate ports, or a closure of a manual block valve, would disable all three transmitters. As such, this might not be the best method of achieving high flow-measurement reliability.

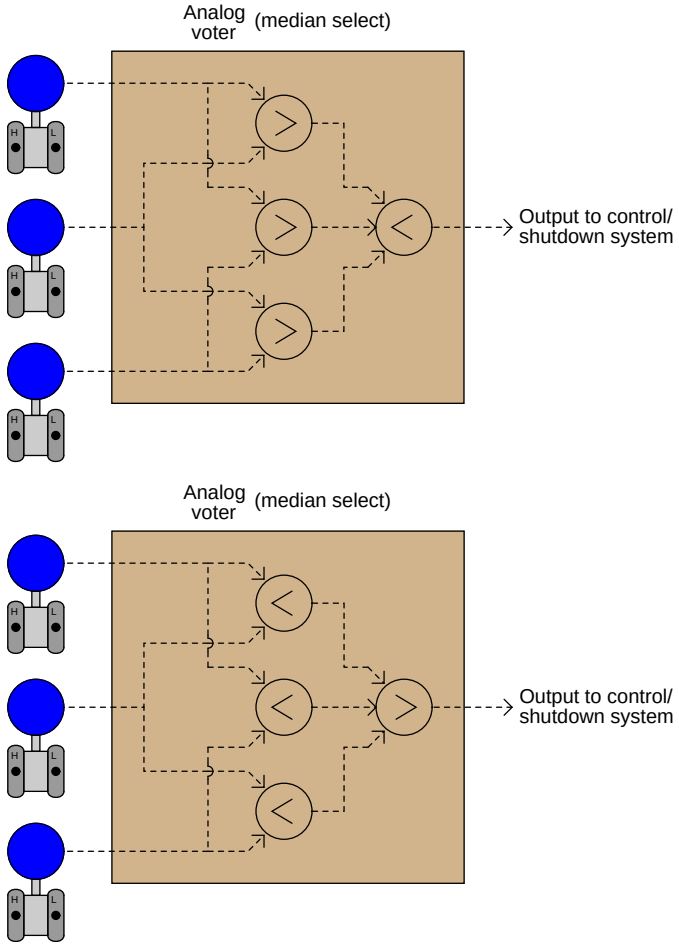
The "trick" to using redundant transmitters is to have the system self-determine what the actual process value is in the event one or more of the redundant transmitters disagree with each other. *Voting* is the name given to this important function, and it often takes the form of signal selector functions:



Multiple selection criteria are typically offered by "voting" modules, including *high*, *low*, *average*, and *median*. A "high" select voter would be suitable for applications where the dangerous condition is a large measured value, the voting module selecting the highest-valued transmitter signal in an effort to err on the side of safety. This would represent a 1oo3 safety redundancy (since only one transmitter out of the three would have to register beyond the high trip level in order to initiate the shutdown). A "low" select voter would, of course, be suitable for any application where the dangerous condition is a small measured value (once again providing a 1oo3 safety redundancy).

The "average" selection function merely calculates and outputs the mathematical average of all transmitter signals – a strategy prone to problems if one of the redundant transmitters happens to fail in the "safe" direction (thus skewing the average value away from the "dangerous" direction and thereby possibly causing the system to respond to an actual dangerous condition later than it should).

The *median select* criterion is very useful in safety systems because it effectively ignores any measurements deviating substantially from the others. Median selector functions may be constructed of high- and low-select function blocks in either of the following manners:



The best way to prove to yourself the median-selecting abilities of both function block networks is to perform a series of "thought experiments" where you declare three arbitrary transmitter signal values, then follow through the selection functions until you reach the output. For any three signal values you might choose, the result should always be the same: the *median* signal value is the one chosen by the voter.

Three transmitters filtered through a median select function effectively provide a 2oo3 safety redundancy, since just a single transmitter registering a value beyond the safety trip point would be ignored by the voting function. *Two* or more transmitters would have to register values past the trip point in order to initiate a shutdown.

29.4.2 SIS controllers (logic solvers)

Control hardware for safety instrumented functions should be separate from the control hardware used to regulate the process, if only for the simple reason that the SIF exists to bring the process to a safe state in the event of any unsafe condition arising, including dangerous failure of the basic regulatory controls. If a single piece of control hardware served the dual purposes of regulation *and* shutdown, a failure within that hardware resulting in loss of regulation (normal control) would not be protected because the safety function would be disabled by the same fault.

Safety controls are usually discrete with regard to their output signals. When a process needs to be shut down for safety reasons, the steps to implement the shutdown often take the form of opening and closing certain valves fully rather than partially. This sort of all-or-nothing control action is most easily implemented in the form of discrete signals triggering solenoid valves or electric motor actuators. A digital controller specially designed for and tasked with the execution of safety instrumented functions is usually called a *logic solver*, or sometimes a *safety PLC*, in recognition of this discrete-output nature.

A photograph of a "safety PLC" used as an SIS in an oil refinery processing unit is shown here, the controller being a Siemens "Quadlog" model:

Although logic solvers are usually designed with levels of self-diagnostics and internal redundancy beyond that of a normal PLC, logic solver programming is very much the same. Ladder Diagram (LD) and Function Block Diagram (FBD) are two popular graphical programming languages used in logic solvers. These two programming languages are limited by their very nature to sets of well-defined algorithms, unlike text-based languages which grant the human programmer much more freedom in determining what will be done and how. A strong rationale for programming logic solvers in a more limited language is safety: the more flexible and unbounded a programming language is, the more potential there will be for complicated "run-time" errors that may be very difficult to troubleshoot. The ISA safety standard number 84 classifies industrial programming languages as either *Fixed Programming Languages* (FPL), *Limited Variability Languages* (LVL), or *Full Variability Languages* (FVL). Ladder Diagram and Function Block Diagram programming are both considered to be "limited variability" languages, whereas Instruction List (and traditional

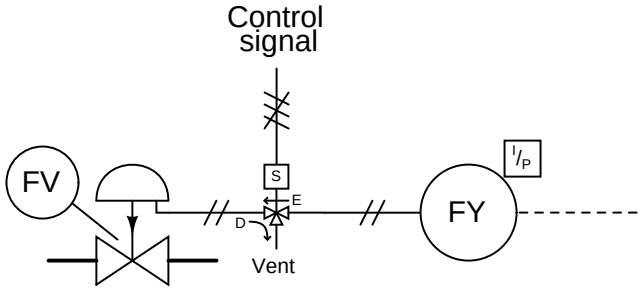
computer programming languages such as C/C++, FORTRAN, BASIC, etc.) are considered "full variability" languages with all the attendant potential for complex errors.

29.4.3 SIS final control elements

When a dangerous condition in a volatile process is sensed by process transmitters (or process switches), triggering a shutdown response from the logic solver, the final control elements must move with decisive and swift action. Such positive response may be obtained from a standard regulatory control valve (such as a globe-type throttling valve), but for more critical applications a rotary ball or plug valve may be more suitable. If the valve in question is used for safety shutdown purposes only and not regulation, it is often referred to as a *chopper* valve for its ability to "chop" (shut off quickly and securely) the process fluid flow. A more formal term for this is an *Emergency Isolation Valve*, or *EIV*.

Some process applications may tolerate the over-loading of both control and safety functions in a single valve, using the valve to regulate fluid flow during normal operation and fully stroke (either open or closed depending on the application) during a shutdown condition. A common method of achieving this dual functionality is to install a solenoid valve in-line with the actuating air pressure line, such that the valve's normal pneumatic signal may be interrupted at any moment, immediately driving the valve to a fail-safe position at the command of a discrete "trip" signal.

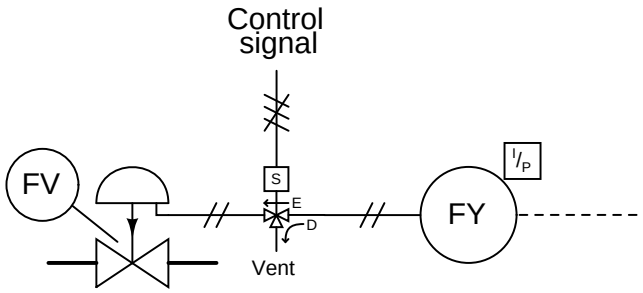
Such a "trip" solenoid (sometimes referred to as a *dump* solenoid, because it "dumps" all air pressure stored in the actuating mechanism) is shown here, connected to a fail-closed (air-to-open) control valve:



Compressed air passes through the solenoid valve from the I/P transducer to the valve's pneumatic diaphragm actuator when energized, the letter "E" and arrow showing this path in the diagram. When de-energized, the solenoid valve blocks air pressure coming from the I/P and vents all air pressure from the valve's actuating diaphragm as shown by the letter "D" and arrow. Venting all actuating air pressure from a fail-closed valve will cause the valve to fail closed, obviously.

If we wished to have the fail fail open on demand, we could use the exact same solenoid and instrument air plumbing, but swap the fail-closed control valve for a fail-open control valve. When energized (regular operation), the solenoid would pass variable air pressure from the I/P transducer to the valve actuator so it could serve its regulating purpose. When de-energized, the solenoid would force the valve to the fully-open position.

For applications where it is safer to lock the control valve in its last position than to have it fail either fully closed or fully open, we might elect to use a solenoid valve in a different manner:



Here, de-energization of the solenoid valve causes the I/P transducer's air pressure output to vent, while trapping and holding all air pressure inside the actuator at the trip time. Regardless of the valve's "natural" fail-safe state, this system forces the valve to lock position[34] until the solenoid is re-energized.

[34] This is assuming, of course, that there are no air leaks anywhere in the actuator, tubing, or solenoid which would cause the trapped pressure to decrease over time.

An example of a trip solenoid installed on a control valve appears in the following photograph. This valve also happens to have a *hand jack* wheel installed in the actuating mechanism, allowing a human operator to manually override the valve position by forcing it closed (or open) when the hand wheel is turned sufficiently:

Of all the components of a Safety Instrumented System (SIS), the final control elements (valves) are generally the least reliable, contributing most towards the system's probability of failure on demand (PFD). Sensors generally come in at second place in their contribution toward unreliability, and logic solvers a distant third place. Redundancy may be applied to control elements by creating valve networks where the failure of a single valve does not cause the system as a whole to fail. Unfortunately, this approach is extremely expensive, as valves have both high capital and high maintenance costs compared to SIS sensors and logic solvers.

A less expensive approach than redundancy to increasing safety valve reliability is to perform regular proof tests of their operation. This is commonly referred to in the industry as *partial stroke*

testing. Rather than proof-test each safety valve to its full travel, which would interrupt normal process operations, the valve is commanded to move only part of its full travel. If the valve responds well to this "partial stroke" test, there is a high probability that it is able to move all the way, thus fulfilling the basic requirements of a proof test without actually shutting the process down[35].

[35] Of course, if there is opportunity to fully stroke the safety valve to the point of process shutdown without undue interruption to production, this is the superior way of performing valve proof tests. Such "test-to-shutdown" proof testing may be scheduled at a time convenient to operations personnel, such as at the beginning of a planned process shutdown.

29.4.4 Safety Integrity Levels

A common way of ranking the reliability of a Safety Instrumented Function (SIF) is to use a simple numerical scale from one to four, with four being extremely reliable and one being only moderately reliable:

SIL number	Required Safety Availability (RSA)	Probability of Failure on Demand (PFD)
1	90% to 99%	0.1 to 0.01
2	99% to 99.9%	0.01 to 0.001
3	99.9% to 99.99%	0.001 to 0.0001
4	99.99% to 99.999%	0.0001 to 0.00001

The Required Safety Availability (RSA) value refers to the reliability of a Safety Instrumented Function in performing its duty. This is the probability that the SIF will perform as needed, when needed. Conversely, the Probability of Failure on Demand (PFD) is the mathematical complement of RSA (PFD = 1 - RSA), expressing the probability that the SIF will fail to perform as needed, when needed.

Conveniently, the SIL number matches the minimum number of "nines" in the Required Safety Availability (RSA) value. For instance, a safety instrumented function with a Probability of Failure on Demand (PFD) of 0.00073, will have an RSA value of 99.927%, which equates to a SIL 3 rating.

It is important to understand that SIL ratings apply only to whole Safety Instrumented Functions, and not to specific devices or even to entire systems or processes. An overpressure protection system on a chemical reactor process with a SIL rating of 2, for example, has a Probability of Failure on Demand between 0.01 and 0.001 of all critical components of that specific shutdown system, from the sensor(s) to the logic solver to the final control element(s) to the vessel itself including relief valves and other auxiliary equipment. If there arises a need to decrease the probability that the reactor vessel will become overpressured, engineers have a variety of options at their disposal for doing so. The safety instruments themselves might be upgraded, preventive maintenance schedules increased in frequency, or even process equipment changed to make an overpressure event less likely.

SIL ratings do not apply to an entire process. It is quite possible that the chemical reactor mentioned in the previous paragraph with an overpressure protection system SIL rating of 3 might have an over*temperature* protection system SIL rating of only 2, due to differences in how the two different safety systems function.

Adding to this confusion is the fact that many instrument manufacturers rate their products as approved for use in certain SIL-rated applications. It is easy to misunderstand these claims, thinking that a safety instrumented function will be rated at some SIL value simply because instruments rated for that SIL value are used to implement it. In reality, the SIL value of any safety function is a much more complex determination. It is possible, for instance, to purchase and install a pressure transmitter rated for use in SIL 2 applications, and have the safety function as a whole be less than 99% reliable (PFD greater than 0.01, or a SIL level no greater than 1).

As with so many other complex calculations in instrumentation engineering, there exist software packages with all the necessary formulae pre-programmed for engineers and technicians alike to use for calculating SIL ratings of safety instrumented functions. These software tools not only factor in the inherent reliability ratings of different system components, but also correct for preventive

maintenance schedules and proof testing intervals so the user may determine the proper maintenance attention required to achieve a given SIL rating.

29.4.5 SIS example: burner management systems

One "classic" example of an industrial automatic shutdown system is a *Burner Management System* (or *BMS*) designed to monitor the operation of a combustion burner and shut off the fuel supply in the event of a dangerous condition. Sometimes referred to as *flame safety systems*, these systems watch for such potentially dangerous conditions as *low fuel pressure*, *high fuel pressure*, and *loss of flame*. Other dangerous conditions related to the process being heated (such as *low water level* for a steam boiler) may be included as additional trip conditions.

The safety shutdown action of a burner management system is to halt the flow of fuel to the burner in the event of any hazardous detected condition. The final control element is therefore one or more shutoff valves (and sometimes a vent valve in addition) to positively stop fuel flow to the burner.

A typical ultraviolet flame sensor appears in this photograph:

This flame sensor is sensitive to ultraviolet light only, not to visible or infrared light. The reason for this specific sensitivity is to ensure the sensor will not be "fooled" by the visible or infrared glow of hot surfaces inside the firebox if ever the flame goes out unexpectedly. Since ultraviolet light is emitted *only* by an active gas-fueled flame, the sensor acts as a true flame detector, and not a heat detector.

One of the more popular models of fuel gas safety shutoff valve used in the United States for burner management systems is shown here, manufactured by Maxon:

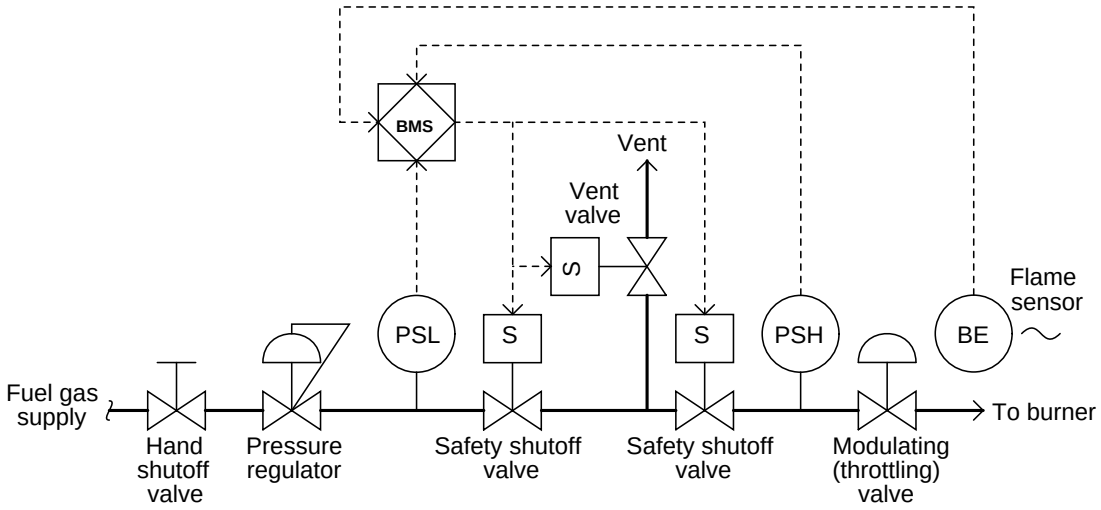
This particular model of shutoff valve has a viewing window on it where a metal tag linked to the valve mechanism marked "Open" (in red) or "Shut" (in black) positively indicates the valve's mechanical status. Like most safety shutoff valves on burner systems, this valve is electrically actuated, and will automatically close by spring tension in the event of a power loss.

Another safety shutoff valve, this one manufactured by ITT, is shown here:

Close inspection of the nameplate on this ITT safety valve reveals several important details. Like the Maxon safety valve, it is electrically actuated, with a "holding" current indicated as 0.14 amps at 120 volts AC. Inside the valve is an "auxiliary" switch designed to actuate when the valve has mechanically reached the full "open" position. An additional switch, labeled *valve seal overtravel interlock*, indicates when the valve has securely reached the full "shut" position. This "valve seal" switch generates a *proof of closure* signal used in burner management systems to verify a safe shutdown condition of the fuel line. Both switches are rated to carry 15 amps of current at 120

VAC, which is important when designing the electrical details of the system to ensure the switch will not be tasked with too much current.

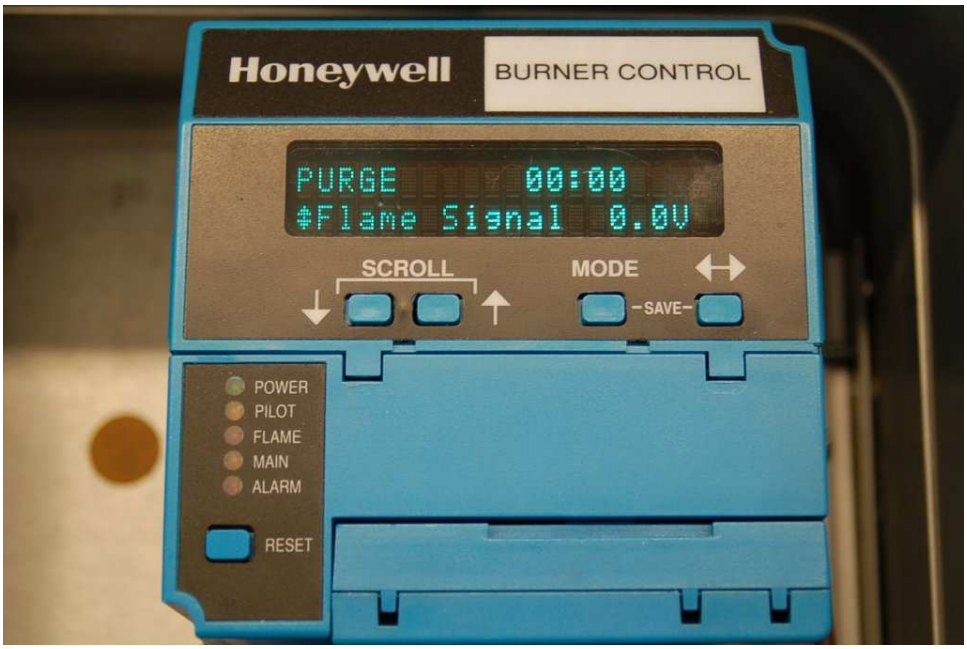
A simple P&ID for a gas-fired combustion burner system is shown here. The piping and valving shown is typical for a single burner. Multiple-burner systems are often equipped with individual shutoff valve manifolds and individual fuel pressure limit switches. Each burner, if multiple exist in the same furnace, *must* be equipped with its own flame sensor:



Note the use of double-block and bleed shutdown valves to positively isolate the fuel gas supply from the burner in the event of an emergency shutdown. The two block valves are specially designed for the purpose (such as the Maxon and ITT safety valves previously shown), while the bleed valve is often nothing more than an ordinary electric solenoid valve.

Most burner management systems are charged with a dual role: both to manage the safe shutdown of a burner in the event of a hazardous condition, *and* the safe start-up of a burner in normal conditions. Start-up of a large industrial burner system usually includes a lengthy *purge time* prior to ignition where the combustion air damper is left wide-open and the blower running for several minutes to positively purge the firebox of any residual fuel vapors. After the purge time, the burner management system will ignite the burner (or sometimes ignite a smaller burner called the *pilot*, which in turn will light the main burner). A burner management system handles all these pre-ignition and timing functions to ensure the burners will ignite safely and without incident.

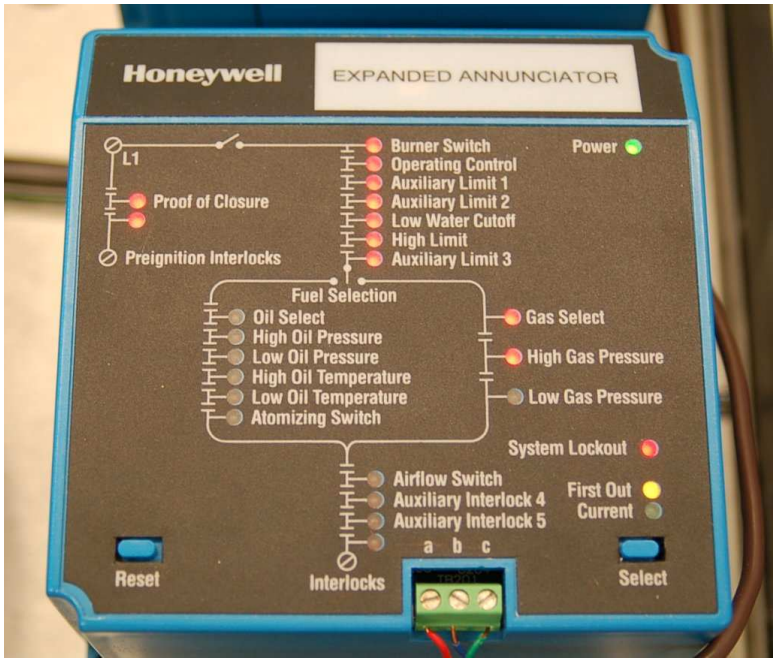
While many industrial burners are managed by electromechanical relay or analog electronic control systems, the modern trend is toward microprocessor-based digital electronic controls. One popular system is the Honeywell 7800 series burner control system, an example of which is shown in this photograph:



Microprocessor controls provide numerous advantages over relay-based and analog electronic burner management systems. Timing of purge cycles is far more accurate with microprocessor control, and the requisite purge time is more difficult to override[36]. Microprocessor-based burner controls usually have digital networking capability as well, allowing the connection of multiple controls to a single computer for remote monitoring.

[36] Yes, maintenance and operations personnel alike are often tempted to bypass the purge time of a burner management system out of impatience and a desire to resume production. I have personally witnessed this in action, performed by an electrician with a screwdriver and a "jumper" wire, overriding the timing function of a flame safety system during a troubleshooting exercise simply to get the job done faster. The electrician's rationale was that since the burner system was having problems lighting, and had been repeatedly purged in prior attempts, the purge cycle did not have to be full-length in subsequent attempts. I asked him if he would feel comfortable repeating those same words in court as part of the investigation of why the furnace exploded. He didn't think this was funny.

The Honeywell 7800 series additionally offers local "annunciator" modules to visually indicate the status of permissive (interlock) contacts, showing maintenance personnel which switches are closed and what state the burner control system is in:



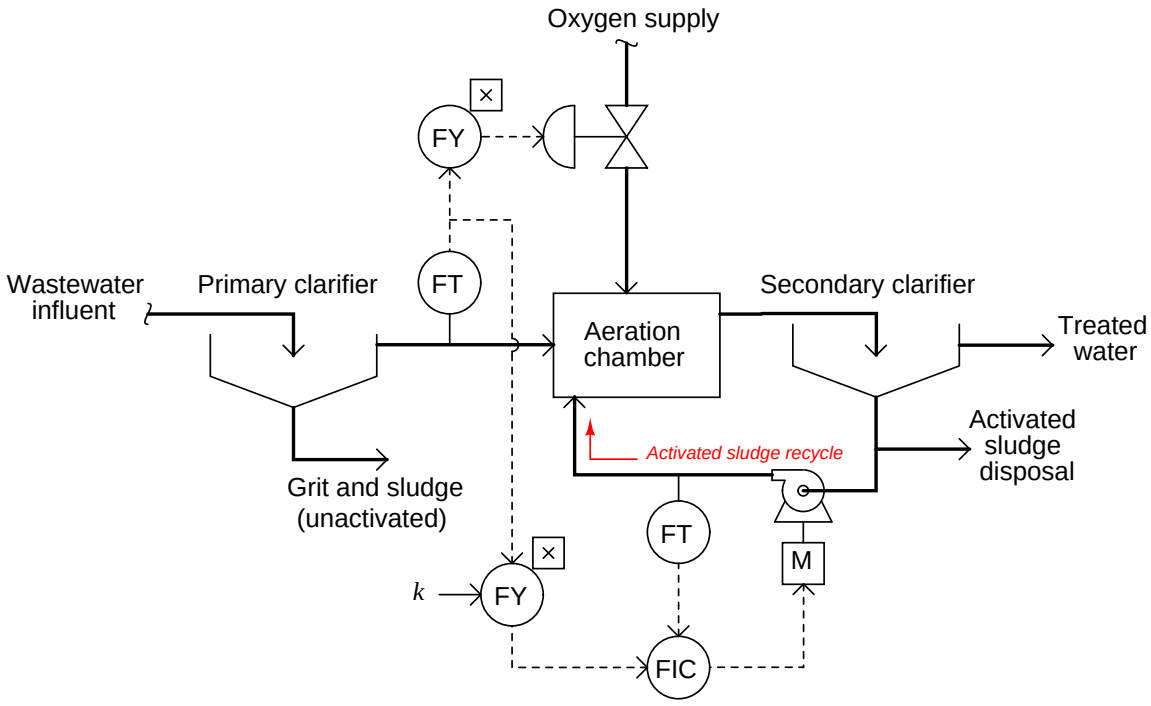
The entire "gas train" piping system for a dual-fuel boiler at a wastewater treatment facility appears in the following photograph. Note the use of double-block and bleed valves on both "trains" (one for utility-supplied natural gas and the other for "sludge gas" produced by the facility's anaerobic digesters), the block valves for each train happening to be of different manufacture. A Honeywell 7800 flame safety control system is located in the blue enclosure:

29.4.6 SIS example: water treatment oxygen purge system

One of the processes of municipal wastewater treatment is the aerobic digestion of organic matter by bacteria. This process emulates one of many waste-decomposition processes in nature, performed on an accelerated time frame for the needs of large wastewater volumes in cities. The process consists of supplying naturally occurring bacteria within the wastewater with enough oxygen to metabolize the organic waste matter, which to the bacteria is food. In some treatment facilities, this aeration is performed with ambient air. In other facilities, it is performed with nearly pure oxygen.

Aerobic decomposition is usually part of a larger process called *activated sludge*, whereby the effluent from the decomposition process is separated into solids (sludge) and liquid (supernatant), with a large fraction of the sludge recycled back to the aerobic chamber to sustain a healthy culture of bacteria and also ensure adequate retention time for decomposition to occur. Separating liquids from solids and recycling the solids ensures a short retention time for the liquid (allowing high processing rates) and a long retention time for the solids (ensuring thorough digestion of organic matter by the bacteria).

A simplified P&ID of an activated sludge water treatment system is shown here, showing how both the oxygen flow into the aeration chamber and the sludge recycle flow back to the aeration chamber are controlled as a function of influent wastewater flow:

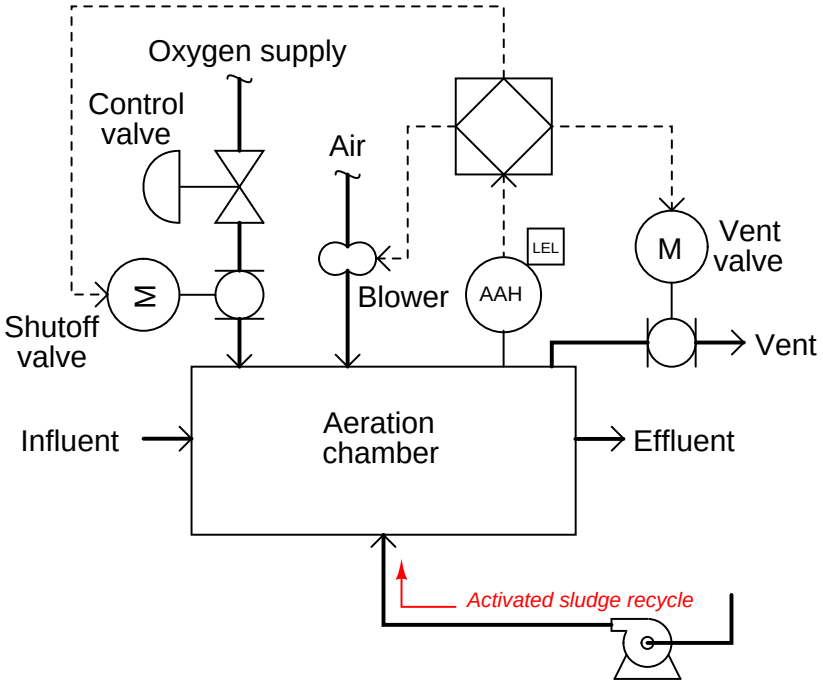


Aerobic decomposition performed with ambient air as the oxidizer is a very simple and safe process. Pure oxygen may be chosen instead of ambient air because it accelerates the metabolism of the bacteria, allowing more processing flow capacity in less physical space. For the same reason that pure oxygen accelerates bacterial metabolism, it also accelerates combustion of any flammable substances. This means if ever a flammable vapor or liquid were to enter the aeration chamber, there would be a risk of explosion.

Although flammable liquids are not a normal component of municipal wastewater, it is possible for flammable liquids to find their way to the wastewater treatment plant. One possibility is the event of a fuel carrier vehicle spilling its cargo, with gasoline or some other volatile fuel draining into a sewer system tunnel through holes in a grate. Such an occurrence is not normal, but certainly possible. Furthermore, it may occur without warning for the operations personnel to take preemptive action at the wastewater treatment plant.

To decrease this safety hazard, *Low Explosive Limit* (LEL) sensors installed on the aeration chamber detect and signal the presence of flammable gases or vapors inside the chamber. If any of the sensors register the presence of flammable substances, a safety shutdown system purges the chamber of pure oxygen by taking the following steps:

- Stop the flow of pure oxygen into the aeration chamber
- Open large vent valves to atmosphere
- Start air blowers to purge the chamber of residual pure oxygen



As with the P&ID, this diagram is a simplified representation of the real safety shutdown system. In a real system, multiple analytical high-alarm (LEL) sensors work to detect the presence of flammable gases or vapors, and the oxygen block valve arrangement would most likely be a double block and bleed rather than a single block valve.

The following photograph shows an LEL sensor mounted inside an insulated enclosure for protection from cold weather conditions at a wastewater treatment facility:

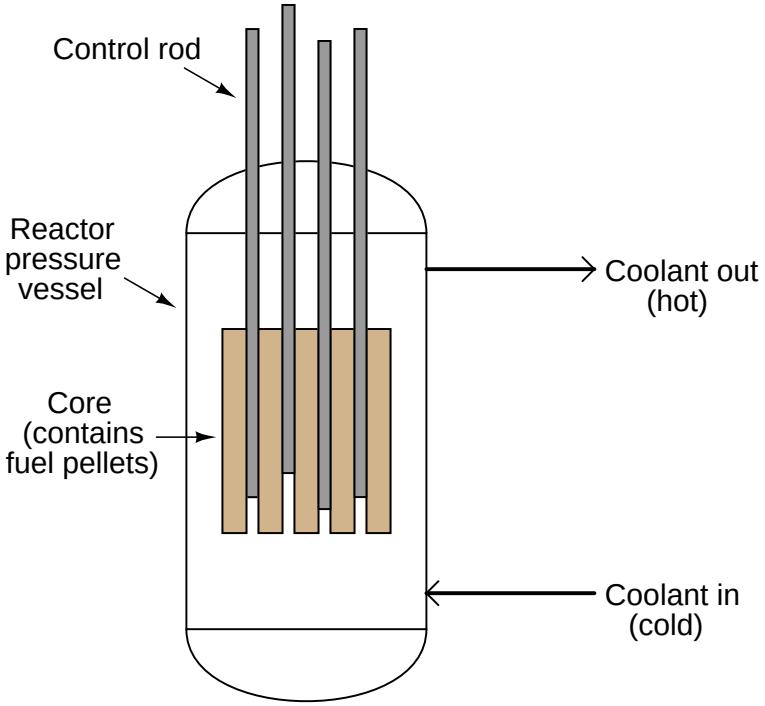
In this photograph, we see a purge air blower used to sweep the aeration chamber of pure oxygen during an emergency shutdown condition:

Since this is a centrifugal blower, providing no seal against air flow through it when stopped, an automatic purge valve located downstream (not to be confused with the manually-actuated vent valve seen in this photograph) is installed to block off the blower from the oxygen-filled chamber. This purge valve remains shut during normal operation, and opens only after the blower has started to initiate a purge.

29.4.7 SIS example: nuclear reactor scram controls

Nuclear fission is a process by which the nuclei of specific types of atoms (most notably uranium-235 and plutonium-239) undergo spontaneous disintegration upon the absorption of an extra neutron, with the release of significant thermal energy and additional neutrons. A quantity of fissile material subjected to a source of neutron particle radiation will begin to fission, releasing massive quantities of heat which may then be used to boil water into steam and drive steam turbine engines to generate electricity. The "chain reaction" of neutrons splitting fissile atoms, which then eject more neutrons to split more fissile atoms, is inherently exponential in nature, but may be regulated by natural and artificial control loops.

A simplified diagram of a pressurized[37] water reactor (PWR) appears here:



[37] Boiling-water reactors (BWR), the other major design type in the United States, output saturated steam at the top rather than heated water. Control rods enter a BWR from the bottom of the pressure vessel, rather than from the top as is standard for PWRs.

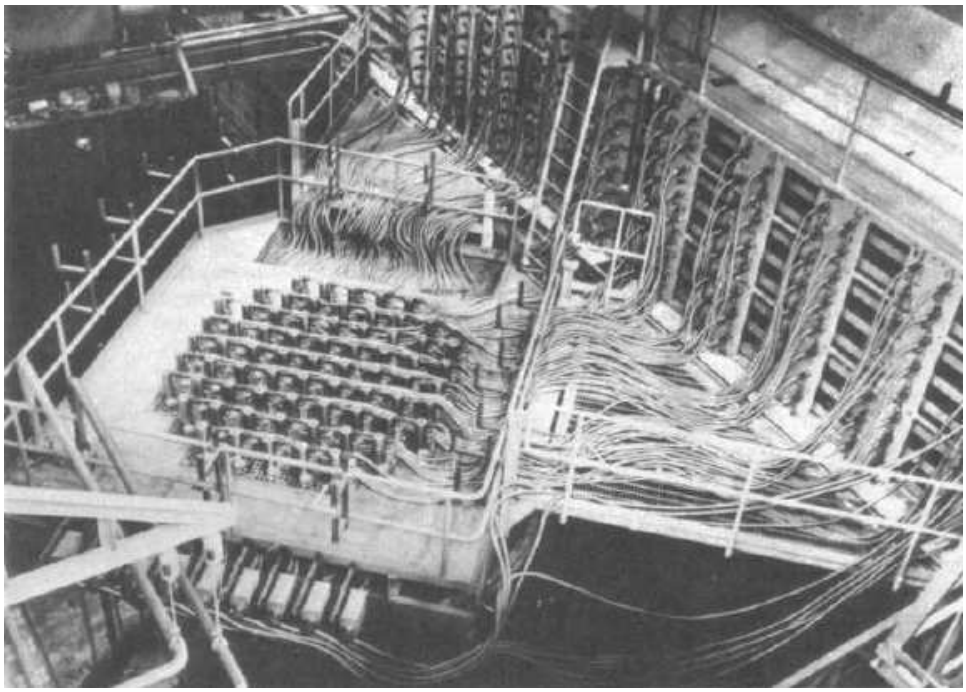
In the United States of America, nuclear reactors are designed to exhibit what is called a *negative temperature coefficient*, which means the chain reaction naturally slows as the temperature of the coolant increases. This physical tendency, engineered by the configuration of the reactor core and the design of the coolant system, adds a measure of self-stabilization to what would otherwise be an inherently unstable ("runaway") process.

Additional regulation ability comes from the insertion of special *control rods* into the reactor core, designed to absorb neutrons and prevent them from "splitting" more atoms. With enough control rods inserted into a reactor core, a chain reaction cannot self-sustain. With enough control rods withdrawn from a freshly-fueled reactor core, the chain reaction will grow to an intensity strong enough to damage the reactor. Control rod position thus constitutes the primary method of power control for a fission reactor, and also the first[38] means of emergency shutdown.

Due to the intense radiation flux near an operating power reactor, these control rods must be manipulated remotely rather than by direct human actuation. Nuclear reactor control rod actuators are typically special electric motors developed for this critical application.

[38] Other means of reactor shutdown exist, such as the purposeful injection of "neutron poisons" into the coolant system which act as neutron-absorbing control rods on a molecular level. The insertion of "scram" rods into the reactor, though, is by far the *fastest* method for quenching the chain-reaction.

A photograph[39] showing the control rod array at the top of the ill-fated reactor at Three-Mile Island nuclear power plant appears here, with a mass of control cables connecting the rod actuators to the reactor control system:



Rapid insertion of control rods into a reactor core for emergency shutdown purposes is called a *scram*. Accounts vary as to the origin of this term, whether it has meaning as a technical acronym or as a colloquial expression to evacuate an area. Regardless of its etymology, a "scram" is an event to be avoided if possible. Like all industrial processes, a nuclear reactor fulfills its intended purpose only when operating. Shutdowns represent not only loss of revenue for the operating company, but also loss of power to local utilities and possible disruption of critical public services (heating, cooling, water pumping, fire protection, traffic control, etc.). An emergency shutdown system at a nuclear power plant must fulfill the opposing roles of safety and availability, with an extremely high degree of instrument reliability.

The electric motor actuators intended for normal operation of control rods are generally too slow to use for scram purposes. Hydraulic actuators capable of overriding the electric motor actuation may be used for scram insertion. Some early pressurized-water reactor scram system designs used a simple mechanical latch, disengaging the control rods from their motor actuators and letting gravity draw the rods fully into the reactor core.

[39] This appears courtesy of the Nuclear Regulatory Commission's special inquiry group report following the accident at Three Mile Island, on page 159.

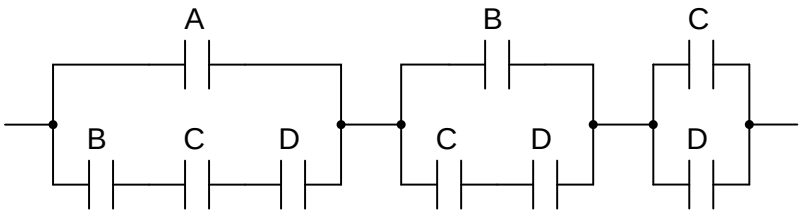
A partial list of criteria sufficient to initiate a scram is shown here:

- Detected earthquake
- Reactor pressure high
- Reactor pressure low
- Reactor water level low (BWR only)
- Reactor differential temperature high
- Main steam isolation valve shut
- Detected high radioactivity in coolant loop
- Detected high radioactivity in containment building
- Manual shutdown switch(es)
- Control system power loss
- Core neutron flux high
- Core neutron flux rate-of-change (period) high

The last two criteria bear further explanation. Since each fission event (the "splitting" of one fuel atom's nucleus by an absorbed neutron) results in a well-defined range of thermal energy release and also a well-defined range of additional neutrons released, the number of neutrons detected in the reactor core at any given moment is an approximate indication of the core's thermal power. Neutron radiation flux measurement is therefore a fundamental process variable for fission reactor control, and also for safety shutdown. If sensors detect an excessive neutron flux, the reactor should be "scrammed" to avoid damage due to overheating. Likewise, if sensors detect a neutron flux level that is *rising* at an excessive *rate*, it indicates the possibility of a runaway chain-reaction which should also initiate a reactor "scram."

In keeping with the high level of reliability and emphasis on safety for nuclear reactor shutdown controls, a common redundant strategy for sensors and logic is *two-out-of-four*, or *2oo4*. A contact logic diagram showing a 2oo4 configuration appears here:

2oo4 redundant logic for reactor scram systems



Any two contacts (A, B, C, or D) opening will interrupt power flow and "scram" the reactor

References

Adamski, Robert S., *Design Critical Control or Emergency Shut Down Systems for Safety AND Reliability*, Revision 2, Premier Consulting Services, Irvine, CA.

ANSI/ISA-84.00.01-2004 Part 1 (IEC 61151-1 Mod), "Functional Safety: Safety Instrumented Systems for the Process Industry Sector – Part 1: Framework, Definitions, System, Hardware and Software Requirements", ISA, Research Triangle Park, NC, 2004.

ANSI/ISA-84.00.01-2004 Part 2 (IEC 61151-2 Mod), "Functional Safety: Safety Instrumented Systems for the Process Industry Sector – Part 2: Guidelines for the Application of ANSI/ISA-84.00.01-2004 Part 1 (IEC 61151-1 Mod)", ISA, Research Triangle Park, NC, 2004.

Bazovsky, Igor, *Reliability Theory and Practice*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1961.

"Engineer's Guide", Pepperl+Fuchs.

"Failure Mode / Mechanism Distributions" (FMD-97), Reliability Analysis Center, Rome, NY, 1997.

Grebe, John and Goble, William, *Failure Modes, Effects and Diagnostic Analysis; Project: 3051C Pressure Transmitter*, Report number Ros 03/10-11 R100, exida.com L.L.C., 2003.

Hattwig, Martin, and Steen, Henrikus, *Handbook of Explosion Prevention and Protection*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2004.

Hicks, Tyler G., *Standard Handbook of Engineering Calculations*, McGraw-Hill Book Company, New York, NY, 1972.

"IEC 61508 Frequently Asked Questions", Rosemount website `http://mw4rosemount.usinternet.com/solution/faq61508.html`, updated December 1, 2003.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Measurement and Analysis Volume I*, Fourth Edition, CRC Press, New York, NY, 2003.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Control Volume II*, Third Edition, CRC Press, Boca Raton, FL, 1999.

Lipták, Béla G., *Instrument Engineers' Handbook – Process Software and Digital Networks*, Third Edition, CRC Press, New York, NY, 2002.

"Modern Instrumentation and Control for Nuclear Power Plants: A Guidebook", Technical Reports Series No. 387, International Atomic Energy Agency (IAEA), Vienna, 2009.

Newnham, Roger and Chau, Paul, "Safety Controls and Burner Management Systems (BMS) on Direct-Fired Multiple Burner Heaters", Born Heaters Canada Ltd.

"NFPA 70", National Electrical Code, 2008 Edition, National Fire Protection Association.

"NIOSH Pocket Guide to Chemical Hazards", DHHS (NIOSH) publication # 2005-149, Department of Health and Human Services (DHHS), Centers for Disease Control and Prevention (CDC), National Institute for Occupational Safety and Health (NIOSH), Cincinnati, OH, September 2005.

Perrow, Charles, *Normal Accidents: living with high-risk technologies*, Princeton University Press, Princeton, NJ, 1999.

Rogovin, Mitchell and Frampton, George T. Jr., *Three Mile Island Volume I, A Report to the Commissioners and to the Public*, Nuclear Regulatory Commission Special Inquiry Group, Washington DC, 1980.

Schultz, M. A., *Control of Nuclear Reactors and Power Plants*, McGraw-Hill Book Company, New York, NY, 1955.

Showers, Glenn M., "Preventive Maintenance for Burner-Management Systems", HPAC – Heating/Piping/Air Conditioning Engineering, February 2000.

Svacina, Bob, and Larson, Brad, *Understanding Hazardous Area Sensing*, TURCK, Inc., Minneapolis, MN, 2001.

"The SPEC 200 Concept", Technical Information document TI 200-100, The Foxboro Company, Foxboro, MA, 1972.

Wehrs, Dave, "Detection of Plugged Impulse Lines Using Statistical Process Monitoring Technology", Emerson Process Management, Rosemount Inc., Chanhassen, MN, December 2006.

Chapter 30

Instrument system problem-solving

The ability to solve complex problems is the most valuable technical skill an instrumentation professional can cultivate. A great many tasks associated with instrumentation work may be broken down into simple step-by-step instructions that any marginally qualified person may perform, but effective problem-solving is different. Problem-solving requires creativity, attention to detail, and the ability to approach a problem from multiple mental perspectives.

"Problem-solving" often refers to the solution of abstract problems, such as "word" problems in a mathematics class. However, in the field of industrial instrumentation it most often finds application in the form of "troubleshooting:" the diagnosis and correction of problems in instrumented systems. Troubleshooting is really just a form of problem-solving, applied to real physical systems rather than abstract scenarios. As such, many of the techniques developed to solve abstract problems work well in diagnosing real system problems. As we will see in this chapter, problem-solving in general and troubleshooting in particular are closely related to *scientific method*, where hypotheses are proposed, tested, and modified in the quest to discern cause and effect.

Like all skills, problem-solving may be improved with practice and persistence. The goal of this chapter is to outline several problem-solving tools and techniques.

30.1 Classic mistakes to avoid

Perhaps the most common mistake made by technicians attempting to diagnose a system problem is failing to gather data (i.e. taking measurements and performing simple system tests) during the troubleshooting process. Even a small amount of data gathered from a system may profoundly accelerate the process of diagnosis.

A colleague of mine has a very descriptive term for this poor habit: *Easter-egging*. The idea is that a technician goes about finding the problem in a system the same way they might go about searching for eggs hidden on Easter morning: randomly. With Easter egg hunting, the eggs could literally be hidden *anywhere*, and so there is no rational way to proceed on a search. In like manner, a technician who lacks information about the nature or source of a system problem is likely to hunt in random fashion for its source. Not only will this likely require significant time and effort, but it may very well fail entirely.

A much more efficient way to proceed is to gather new data with each and every step in the troubleshooting process. By "gathering data," I mean the following:

- Taking measurements with test equipment (multimeter, pressure gauges, etc.)
- Observing equipment indicator lights
- Stimulating the system and observing its response(s)
- Using your other senses (smell, hearing, touch) to gather clues
- Documenting new data in a notepad to help track and analyze the results of your measurements and tests

30.2 Helpful "tricks" using a digital multimeter (DMM)

The digital multimeter (DMM) is quite possibly the most useful tool in the instrument technician's collection[1]. This one piece of test equipment, properly handled, yields valuable insight into the status and operation of many electrical and electronic systems. Not only is a good-quality multimeter capable of precisely indicating electrical voltage, current, and resistance, but it is also useful for more advanced tests. The subject of this section is how to use a digital multimeter for some of these advanced tests[2].

For all these tests, I suggest the use of a top-quality field multimeter. I am personally a great fan of *Fluke* brand meters, having used this particular brand for nearly my whole professional career. The ability of these multimeters to accurately measure true RMS amplitude, discriminate between AC and DC signals, measure AC signals over a wide frequency range, and survive abuse both mechanical and electrical, is outstanding.

[1] As a child, I often watched episodes of the American science-fiction television show *Star Trek*, in which the characters made frequent use of a diagnostic tool called a *tricorder*. Week after week the protagonists of this show would avoid trouble and solve problems using this nifty device. The *sonic screwdriver* was a similar tool in the British science-fiction television show *Doctor Who*. Little did I realize while growing up that my career would make just as frequent use of another diagnostic tool: the electrical multimeter.

[2] I honestly considered naming this section "Stupid Multimeter Tricks," but changed my mind when I realized how confusing this could be for some of my readers not familiar with colloquial American English.

30.2.1 Recording unattended measurements

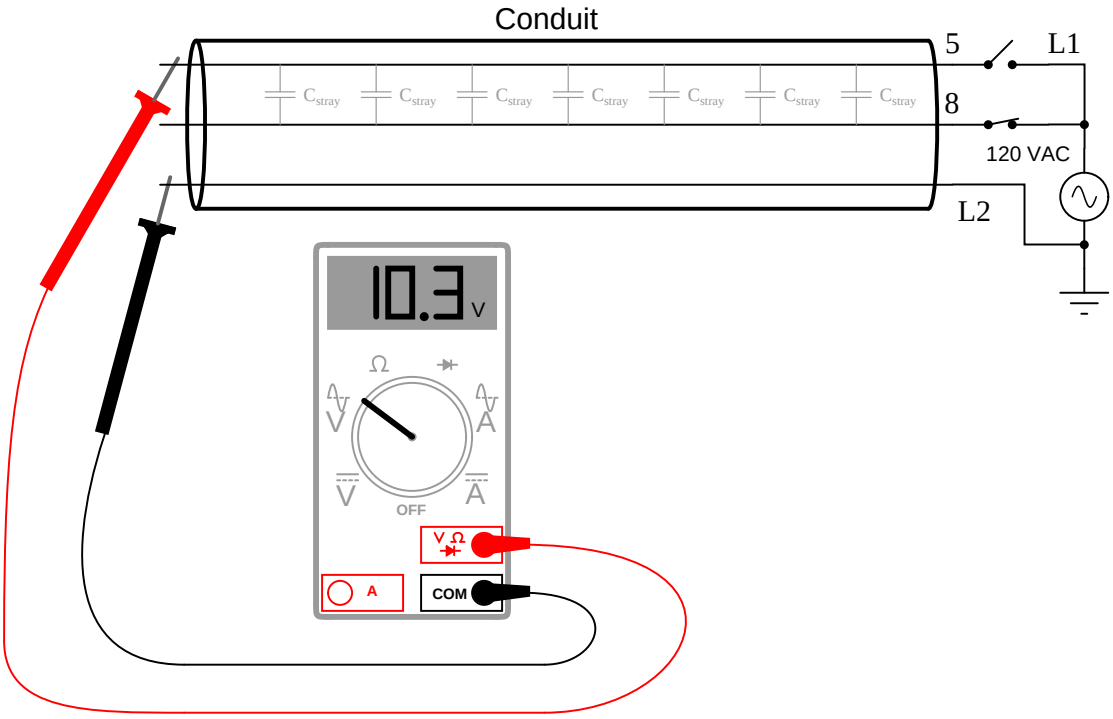
Many modern multimeters have a feature that records the highest and lowest measurements sensed during the duration of a test. On Fluke brand multimeters, this is called the *Min/Max* function. This feature is extremely useful when diagnosing intermittent problems, where the relevant voltages or currents indicating or causing the problem are not persistent, but rather come and go. Many times I have used this feature to monitor a signal with an intermittent "glitch," while I attended to other tasks.

The most basic high-low capture function on a multimeter only tells you what the highest and lowest measured readings were during the test interval (and that only within the meter's scan time – it is possible for a very brief transient signal to go undetected by the meter if its duration is less than the meter's scan time). More advanced multimeters actually log the *time* when an event occurs, which is obviously a more useful feature. If your tool budget can support a digital multimeter with "logging" capability, spend the extra money and take the time to learn how this feature works!

30.2.2 Avoiding "phantom" voltage readings

My first "trick" is not a feature of a high-quality DMM so much as it is a solution to a common problem *caused* by the use of a high-quality DMM. Most digital multimeters exhibit very high input impedance in their voltage-measuring modes. This is commendable, as an ideal voltmeter should have infinite input impedance (so as to not "load" the voltage signal it measures). However, in industrial applications, this high input impedance may cause the meter to register the presence of voltage where none should rightfully appear.

Consider the case of testing for the absence of AC voltage on an isolated power conductor that happens to lie near other (energized) AC power conductors within a long run of conduit:

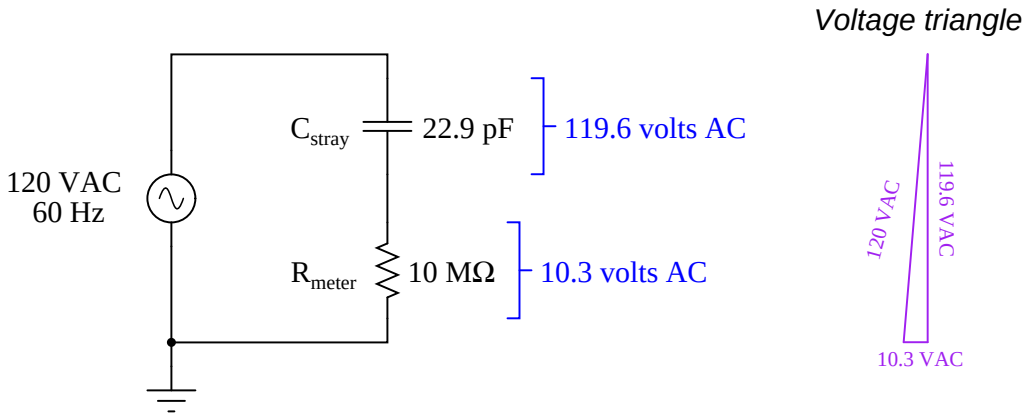


With the power switch feeding wire 5 in the open state, there should be no AC voltage measured between wire 5 and neutral (L2), yet the voltmeter registers slightly over 10 volts AC. This "phantom voltage" is due to capacitive coupling between wire 5 and wire 8 (still energized) throughout the length of their mutual paths within the conduit.

Such phantom voltages may be very misleading if the technician encounters them while troubleshooting a faulty electrical system. Phantom voltages give the impression of connection (or at least high-resistance connection) where no continuity actually exists. The example shown, where the phantom voltage is 10.3 volts compared to the source voltage value of 120 volts, is actually quite modest. With increased stray capacitance between the conductors (longer wire runs in close proximity, and/or more than one energized "neighboring" wire), phantom voltage magnitude begins

to approach that of the source voltage[3].

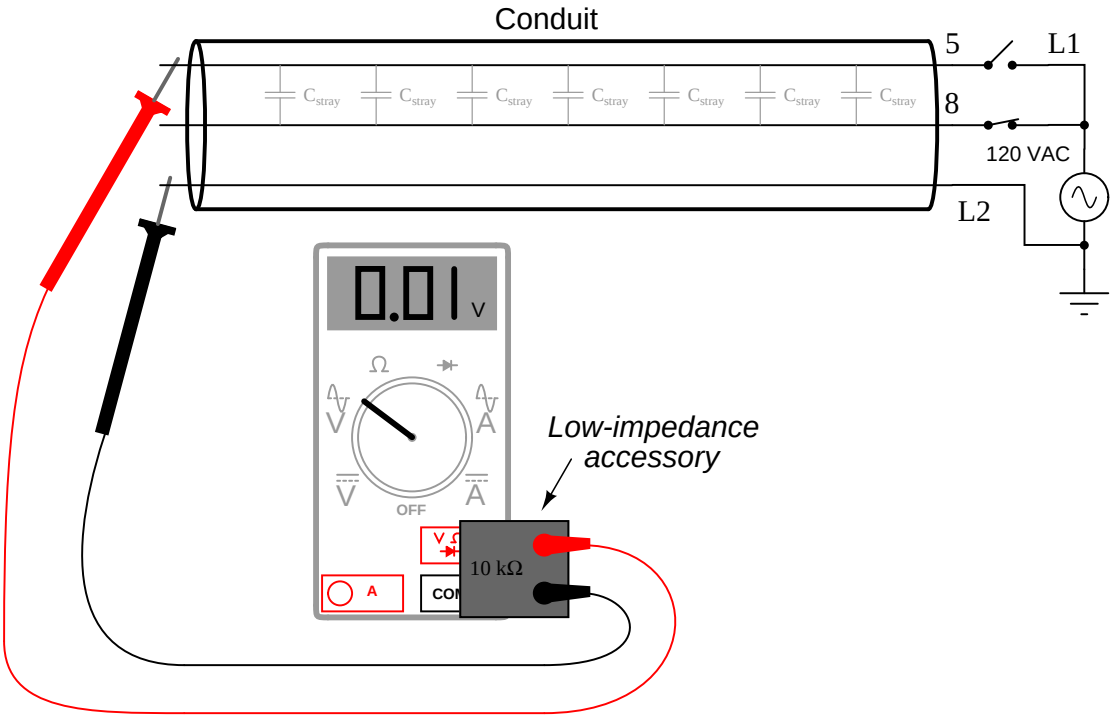
The equivalent circuit is shown here, with the DMM modeled as a 10 MΩ resistance:



An analog voltmeter would never have registered 10.3 volts under the same conditions, due to its substantially lower input impedance. Thus, "phantom voltage" readings are a product of modern test equipment more than anything else.

[3]I have personally measured "phantom" voltages in excess of 100 volts AC, in systems where the source voltage was 120 volts AC.

The obvious solution to this problem is to use a different voltmeter – one with a much lesser input impedance. But what is a technician to do if their only voltmeter is a high-impedance DMM? Connect a modest resistance in parallel with the meter input terminals, of course! Fluke happens to market just this type of accessory[4], the SV225 "Stray Voltage Adapter" for the purpose of eliminating stray voltage readings on a high-impedance DMM:



With the voltmeter's input impedance artificially decreased by the application of this accessory, the capacitive coupling is insufficient to produce any substantial voltage dropped across the voltmeter's input terminals, thus eliminating. The technician may now proceed to test for the presence of AC control signal (or power) voltages with confidence.

[4]Before there was such an accessory available, I used a 20 kΩ high-power resistor network connected in parallel with my DMM's input terminals, which I fabricated myself. It was ugly and cumbersome, but it worked well. When I made this, I took great care in selecting resistors with power ratings high enough that accidental contact with a truly "live" AC power source (up to 600 volts) would not cause damage to them. A pre-manufactured device such as the Fluke SV225, however, is a much better option.

30.2.3 Non-contact AC voltage detection

While the last multimeter "trick" was the elimination of a parasitic effect, this trick is the exploitation of that same effect: "phantom voltage" readings obtained through capacitive coupling of a high-impedance voltmeter to a conductor energized with AC voltage (with respect to ground). You may use a high-impedance AC voltmeter to perform qualitative measurements of ground-referenced AC power voltage by setting the meter to the most sensitive AC range possible, grounding one test lead, and simply touching the other test lead to the insulation of the conductor under test. The presence of voltage (usually in the range of millivolts AC) upon close proximity to the energized conductor will indicate the energization of that conductor.

This trick is useful for determining whether or not particular AC power or control wires are energized in a location where the only access you have to those wires is their insulating sheaths. An example of where you might encounter this situation is where you have removed the cover from a conduit elbow or other fitting to gain access to a wire bundle, and you find those wires labeled for easy identification, but the wires do not terminate to any exposed metal terminals for you to contact with your multimeter's probe tips. In this case, you may firmly connect one probe to the metal conduit fitting body, while individually touching the other probe tip to the desired conductors (one at a time), watching the meter's indication in AC millivolts.

Several significant caveats limit the utility of this "trick:"

- The impossibility of quantitative measurement
- The potential for "false negative" readings (failure to detect a voltage that is present)
- The potential for "false positive" readings (detection of a "phantom voltage" from an adjacent conductor)
- The exclusive applicability to AC voltages of significant magnitude ($\geq$ 100 VAC)

Being a qualitative test only, the millivoltage indication displayed by the high-impedance voltmeter tells you nothing about the actual magnitude of AC voltage between the conductor and ground. Although the meter's input impedance is quite constant, the parasitic capacitance formed by the surface area of the test probe tip and the thickness (and dielectric constant) of the conductor insulation is quite variable. However, in conditions where the validity of the measurement may be established (e.g. cases where you can touch the probe tip to a conductor known to be energized in order to establish a "baseline" millivoltage signal), the technique is useful for quickly checking the energization status of conductors where ohmic (metal-to-metal) contact is impossible.

For the same reason of wildly variable parasitic capacitance, this technique should *never* be used to establish the de-energization of a conductor for safety purposes. The only time you should trust a voltmeter's non-indication of line voltage is when that same meter is validated against a known source of similar voltage in close proximity, and when the test is performed with direct metal-to-metal (probe tip to wire) contact. A non-indicating voltmeter *may* indicate the absence of dangerous voltage, or it may indicate an insensitive meter.

30.2.4 Detecting AC power harmonics

The presence of *harmonic* voltages[5] in an AC power system may cause many elusive problems. Power-quality instruments exist for the purpose of measuring harmonic content in a power system, but a surprisingly good qualitative check for harmonics may be performed using a multimeter with a frequency-measuring function.

Setting a multimeter to read AC voltage (or AC current, if that is the quantity of interest) and then activating the "frequency" measurement function should produce a measurement of exactly 60.0 Hz in a properly functioning power system (50.0 Hz in Europe and some other parts of the world). The only way the meter should ever read anything significantly different from the base frequency is if there is significant harmonic content in the circuit. For example, if you set your multimeter to read frequency of AC voltage, then obtained a measurement of 60 Hz that intermittently jumped up to some higher value (say 78 Hz) and then back down to 60 Hz, it would suggest your meter was detecting harmonic voltages of sufficient amplitude to make it difficult for your meter to "lock on" to the fundamental frequency.

It it very important to note that this is a crude test of power system harmonics, and that measurements of "solid" base frequency do not guarantee the absence of harmonics. Certainly, if your multimeter produces unstable readings when set to measure frequency, it suggests the presence of strong harmonics in the circuit. However, the absence of such instability does not necessarily mean the circuit is free of harmonics. In other words, a stable reading for frequency is *inconclusive*: the circuit might be harmonic-free, or the harmonics may be weak enough that your multimeter ignores them and only displays the fundamental circuit frequency.

[5] These are AC voltages having frequencies that are integer-multiples of the fundamental powerline frequency. In the United States, where 60 Hz is standard, harmonic frequencies would be whole-number multiples of 60: 120 Hz, 180 Hz, 240 Hz, 300 Hz, etc.

30.2.5 Identifying noise in DC signal paths

An aggravating source of trouble in analog electronic circuits is the presence of AC "noise" voltage superimposed on DC signals. Such "noise" is immediately evident when the signal is displayed on an oscilloscope screen, but how many technicians carry a portable oscilloscope with them for troubleshooting?

A high-quality multimeter exhibiting good discrimination between AC and DC voltage measurement is very useful as a qualitative noise-detection instrument. Setting the multimeter to read AC voltage, and connecting it to an signal source where pure (unchanging) DC voltage is expected, should yield a reading of nearly zero millivolts. If noise is superimposed on this DC signal, it will reveal itself as an AC voltage, which your meter will display.

Not only is the AC voltage capability of a high-quality (discriminating) multimeter useful in detecting the presence of "noise" voltage superimposed on analog DC signals, it may also give clues as to the source of the noise. By activating the frequency-measuring function of the multimeter while measuring AC voltage (or AC millivoltage), you will be able to track the frequency of the noise to see its value and stability.

Once on a job I was diagnosing a problem in an analog power control system, where the control device was acting strangely. Suspecting that noise on the measurement signal line might be causing the problem, I set my Fluke multimeter to measure AC volts, and read a noise voltage of several tenths of a volt (superimposed on a DC signal a few volts in magnitude). This told me the noise *was* indeed a significant problem. Pressing the "Hz" button on my multimeter, I measured a noise frequency of 360 Hz, which happens to be the "ripple" frequency of a six-pulse (three-phase) AC-to-DC rectifier operating on a base frequency of 60 Hz. This told me where the likely source of the noise was, which led me to the physical location of the problem (a bad shield on a cable run near the rectified power output wiring).

30.2.6 Generating test voltages

Modern digital multimeters are fantastically capable measurement tools, but did you know they are also capable of *generating* simple test signals? Although this is not the design purpose of the resistance and diode-check functions of a multimeter, the meter does output a low DC voltage in each of these settings.

This is useful when qualitatively testing certain instruments such as electronic indicators, recorders, controllers, data acquisition modules, and alarm relays, all designed to input a DC voltage signal from a 250 ohm resistor conducting the 4-20 mA electronic transmitter signal. By setting a multimeter to either the resistance (Ω) or diode check function and then connecting the test leads to the input terminals of the instrument, the instrument's response may be noted.

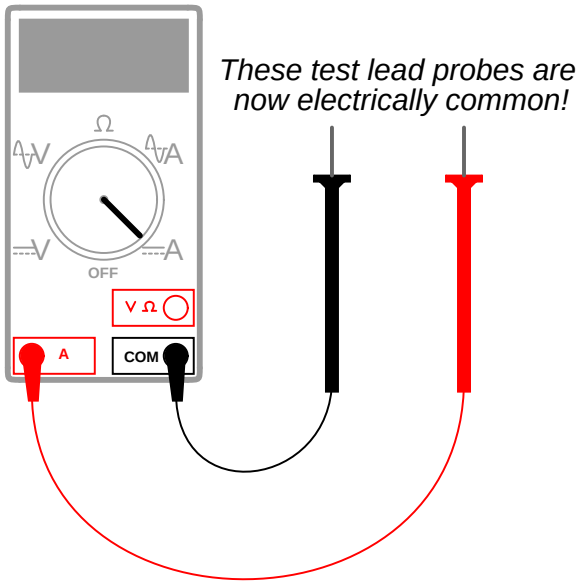
Of course, this is a *qualitative* test only, since multimeters are not designed to output any precise amount of voltage in either the resistance or diode-check modes. However, for testing the basic response of a process indicator, recorder, controller, data acquisition channel, DCS input, or any other DC-signal-receiving devices, it is convenient and useful. In every multimeter I have ever tried this with, the diode-check function outputs *more* voltage than the resistance measurement function[6]. This gives you two levels of "test signal" generation: a low level (resistance) and a high level (diode check). If you are interested in using your multimeter to generate test voltages, I recommend you take the time to connect your multimeter to a high-impedance voltmeter (such as another digital multimeter set to measure DC volts) and note just how much voltage your meter outputs in each mode. Knowing this will allow you to perform tests that are more quantitative than qualitative.

[6] There is a design reason for this. Most digital multimeters are designed to be used on semiconductor circuits, where the minimum "turn-on" voltage of a silicon PN junction is approximately 500 to 700 millivolts. The diode-check function must output more than that, in order to force a PN junction into forward conduction. However, it is useful to be able to check ohmic resistance in a circuit *without* activating any PN junctions, and so the resistance measurement function typically uses test voltages *less than 500 millivolts*.

30.2.7 Using the meter as a temporary jumper

Often in the course of diagnosing problems in electrical and electronic systems, there is a need to temporarily connect two or more points in a circuit together to force a response. This is called "jumpering," and the wires used to make these temporary connections are called *jumper wires*.

More than once I have found myself in a position where I needed to make a temporary "jumper" connection between two points in a circuit, but I did not have any wires with me to make that connection. In such cases, I learned that I could use my multimeter test leads while plugged into the *current-sensing* jacks of the meter. Most digital multimeters have a separate jack for the red test lead, internally connected to a low-resistance *shunt* leading to the common (black) test lead jack. With the red test lead plugged into this jack, the two test leads are effectively common to one another, and act as a single length of wire.



Touching the meter's test leads to two points in a circuit will now "jumper" those two points together, any current flowing through the shunt resistance of the multimeter. If desired, the meter may be turned on to monitor how much current goes through the "jumper" if this is diagnostically relevant.

An additional benefit to using a multimeter in the current-measuring mode as a test jumper is that this setting is usually current-protected by a fuse inside the meter. Applying jumper wires to a live circuit may harbor some danger if significant potential and current-sourcing capability exist between those two points: the moment a jumper wire bridges those points, a dangerous current may develop within the wire. Using the multimeter in this manner gives you a *fused* jumper wire: an added degree of safety in your diagnostic procedure.

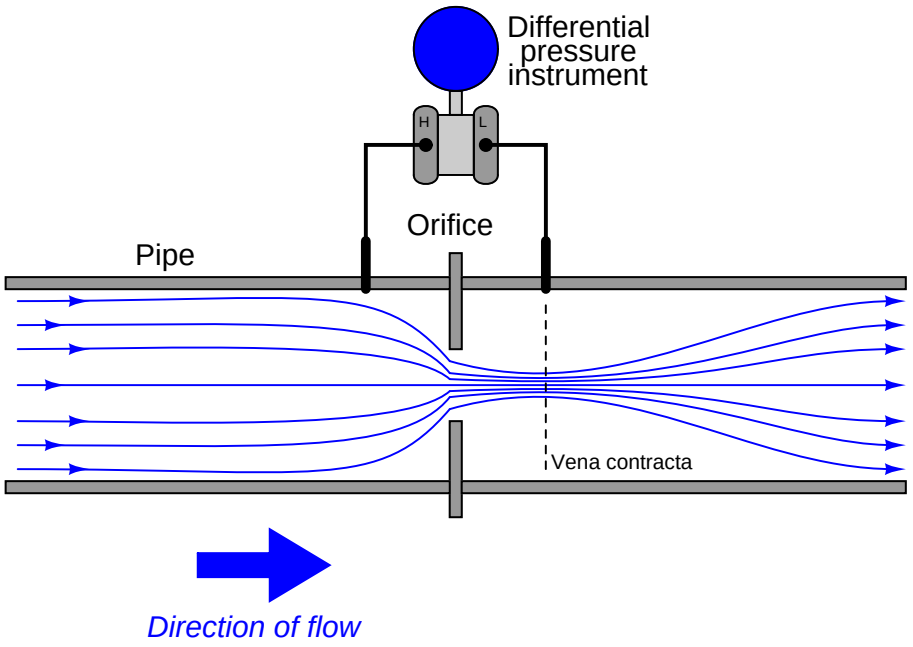
Appendix A

Doctor Strangeflow, or how I learned to relax and love Reynolds numbers

Of all the non-analytical (non-chemistry) process measurements students encounter in their Instrumentation training, flow measurement is one of the most mysterious. Where else would we have to *take the square root* of a transmitter signal just to measure a process variable in the simplest case? Since flow measurement is so vital to many industries, it cannot go untouched in an Instrumentation curriculum. Students must learn how to measure flow, and how to do it accurately. The fact that it is a fundamentally complex thing, however, often leads to oversimplification in the classroom. Such was definitely the case in my own education, and it lead to a number of misunderstandings that were corrected after a lapse of 15 years, in a sudden "Aha!" moment that I now wish to share with you.

The orifice plate is to flow measurement what a thermocouple is to temperature measurement: an inexpensive yet effective primary sensing element. The concept is disarmingly simple. Place a restriction in a pipe, then measure the resulting pressure drop (ΔP) across that restriction to infer flow rate.

You may have already seen a diagram such as the following, illustrating how an orifice plate works:



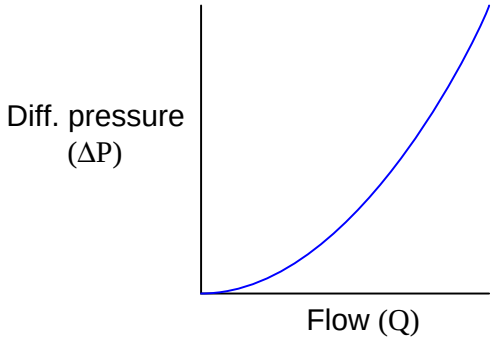
Now, the really weird thing about measuring flow this way is that the resulting ΔP signal does not linearly correspond to flow rate. Double the flow rate, and the ΔP quadruples. Triple the flow rate and the ΔP increases by a factor of nine. To express this relationship mathematically:

$$Q^2 \propto \Delta P$$

In other words, differential pressure across an orifice plate (ΔP) is proportional to the *square* of the flow rate (Q^2). To be more precise, we may include a coefficient (k) with a precise value that turns the proportionality into an equality:

$$Q^2 = k(\Delta P)$$

Expressed in graphical form, the function looks like one-half of a parabola:



To obtain a linear flow measurement signal from the differential pressure instrument's output signal, we must "square root" that signal, either with a computer inside the transmitter, with a computer inside the receiving instrument, or a separate computing instrument (a "square root extractor"). We may see mathematically how this yields a value for flow rate (Q), following from our original equation:

$$Q^2 = k(\Delta P)$$

$$\sqrt{Q^2} = \sqrt{k(\Delta P)}$$

$$Q = \sqrt{k(\Delta P)}$$

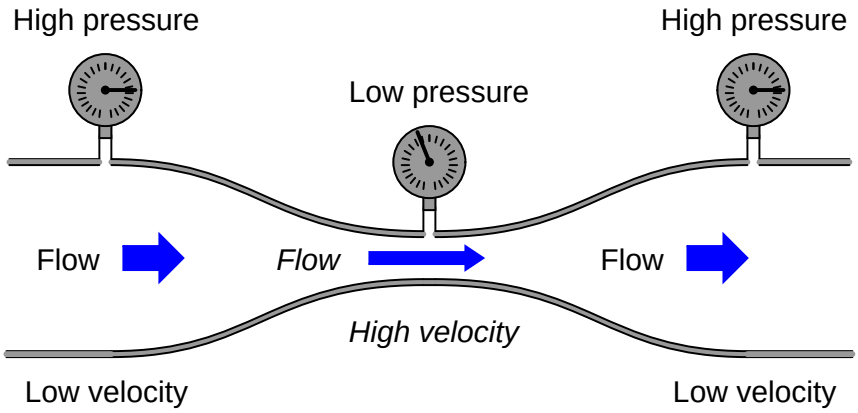
. . . substituting a new coefficient value k[1] . . .

$$Q = k\sqrt{\Delta P}$$

Students are taught that the differential pressure develops as a consequence of energy conservation in the flowing liquid stream. As the liquid enters a constriction, its velocity must increase to account for the same volumetric rate through a reduced area. This results in kinetic energy increasing, which must be accompanied by a corresponding decrease in potential energy (i.e. pressure) to conserve total fluid energy.

[1] Since we get to choose whatever k value we need to make this an equality, we don't have to keep k inside the radicand, and so you will usually see the equation written as it is shown in the last step with k outside the radicand.

Pressure measurements taken in a venturi pipe confirm this:



In all honesty, this did not make sense to me when I heard this. My "common sense" told me the fluid pressure would *increase* as it became crammed into the constriction, not decrease. Even more, "common sense" told me that whatever pressure was lost through the constriction would never be regained, contrary to the pressure indication of the gauge furthest downstream. Accepting this principle was an act of faith on my part, putting preconceived notions aside for something new. A leap of faith, however, is not the same as a leap in understanding. I believed what I was told, but I really didn't understand *why* it was true.

The problem intensified when my teacher showed a more detailed flow equation. This new equation contained a term for fluid density (ρ):

$$Q = k\sqrt{\frac{\Delta P}{\rho}}$$

What this equation showed us is that orifice plate flow measurement depended on density. If the fluid density changed, our instrument calibration would have to change in order to maintain good accuracy of measurement. Something disturbed me about this equation, though, so I raised my hand. The subsequent exchange between me and my teacher went something like this:

Me: What about viscosity?

Teacher: What?

Me: Doesn't fluid viscosity have an effect on flow measurement, just like density?

Teacher: You don't see a variable for viscosity in the equation, do you?

Me: Well, no, but it's *got* to have some effect on flow measurement!

Teacher: How come?

Me: Imagine clean water flowing through a venturi, or through the hole of an orifice plate. At a certain flow rate, a certain amount of ΔP will develop across the orifice. Now imagine

that same orifice flowing an equal rate of liquid honey: approximately the same density as water, but much thicker. Wouldn't the increased "thickness," or viscosity, of the honey result in more friction through the orifice, and thus more of a pressure drop than what the water would create?

Teacher: I'm sure viscosity has some effect, but it must be minimal since it isn't in the equation.

Me: Then why is honey so hard to suck through a straw?

Teacher: Come again?

Me: A straw is a narrow pipe, similar to the throat of a venturi or the hole of an orifice, right? The difference in pressure between the suction in my mouth and the atmosphere is the ΔP across that orifice. The result is flow through the straw. If viscosity is of such little effect, then why is liquid honey so much harder to suck through a straw than water? The pressure is the same, the density is about the same, then why isn't the flow rate the same according to the equation you just gave us?

Teacher: In industry, we usually don't measure fluids as thick as honey, and so it's safe to ignore viscosity in the flow equation . . .

My teacher's smokescreen – that thick fluid flow streams were rare in industry – did nothing to alleviate my confusion. Despite my ignorance of the industrial world, I could very easily imagine liquids that were more viscous than water, honey or no honey. Somewhere, somehow, someone had to be measuring the flow rate of such liquids, and there the effects of viscosity on orifice ΔP must be apparent. Surely my teacher knew this. But then why did the flow equation not have a variable for viscosity in it? How could this parameter be unimportant? Like most students, though, I could see that arguing would get me nowhere and it was better for my grade to just go along with what the teacher said than to press for answers he couldn't give. In other words, I swept my doubts under the carpet of "learning" and made a leap of faith.

After that, we studied different types of orifice plates, different types of pressure tap locations, and other inferential primary sensing elements (annubars, target meters, pipe elbows, etc.). They all worked on Bernoulli's principle of decreased pressure through a restriction, and they all required square root extraction of the pressure signal to obtain a linearized flow measurement. In fact, this became the sole criterion for determining whether or not we needed square root extraction on the signal: did the flow measurement originate from a differential pressure instrument? If so, then we needed to "square root" the signal. If not, we didn't. A neat and clean distinction, separating ΔP-based flow measurements from all the others (magnetic, vortex shedding, Coriolis effect, thermal, etc.). Nice, clean, simple, neat, and only 95% correct, as I was to discover later.

Fast-forward fifteen years. I was now a teacher in a technical college, teaching Instrumentation to students just like myself a decade and a half ago. It was my first time preparing to teach flow measurement, and so I brushed up on my knowledge by consulting one of the best technical references I could get my hands on: Béla Lipták's *Process Measurement and Analysis*, third edition. Part of the *Instrument Engineers' Handbook* series, this wonderful work was to be our primary text as we

explored the world of process measurement during the 2002-2003 academic year.

It was in reading this book that I had an epiphany. Section 2.8 of the text discussed a type of flowmeter I had never seen or heard of before: the *laminar* flowmeter. As I read this section of the book, my jaw hit the floor. Here was a differential-pressure-based flowmeter that was linear! That is, there was no square root extraction required at all to convert the ΔP measurement into a flow measurement. Furthermore, its operation was based on some weird equation called the *Hagen-Poiseuille* Law rather than Bernoulli's Law.

Early in the section's discussion of this flowmeter, a couple of paragraphs explained the meaning of something called *Reynolds number* of a flow stream, and how this was critically important to laminar flowmeters. Now, I had heard of Reynolds number before when I worked in industry, but I never knew what it meant. All I knew is that it had something to do with the selection of flowmeter types: one must know the Reynolds number of a fluid before one could properly select which type of flow-measuring instrument to use in a particular application. Since this determination typically fell within the domain of instrument engineers and not instrument technicians (as I was), I gave myself permission to remain ignorant about it and blissfully went on my way. Little did I know that Reynolds number held the key to understanding my "honey-through-a-straw" question of years ago, as well as comprehending (not just believing) how orifice plates actually worked.

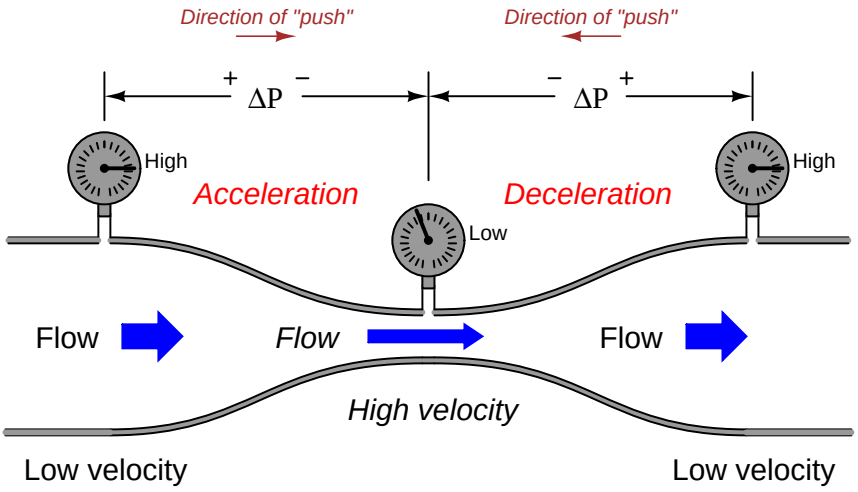
According to Lipták, laminar flowmeters were effective only for low Reynolds numbers, typically below 1200. Cross-referencing the orifice plate section of the same book told me that Reynolds numbers for typical orifice-plate flow streams were much greater (10,000 or higher). Furthermore, the orifice plate section contained an insightful passage on page 152 which I will now quote here. Italicized words indicate my own emphasis, locating the exact points of my "Aha!" moments:

> The basic equations of flow assume that the velocity of flow is uniform across a given cross-section. In practice, flow velocity at any cross section approaches zero in the boundary layer adjacent to the pipe wall, and varies across the diameter. *This flow velocity profile has a significant effect on the relationship between flow velocity and pressure difference developed in a head meter.* In 1883, Sir Osborne Reynolds, an English scientist, presented a paper before the Royal Society, proposing a single, dimensionless ratio now known as Reynolds number, as a criterion to describe this phenomenon. This number, Re, is expressed as
>
> $$Re = \frac{VD\rho}{\mu}$$
>
> where V is velocity, D is diameter, ρ is density, and μ is absolute viscosity. Reynolds number expresses the ratio of inertial forces to viscous forces. At a very low Reynolds number, viscous forces predominate, and the inertial forces have little effect. *Pressure difference approaches direct proportionality to average flow velocity and to viscosity.* At high Reynolds numbers, inertial forces predominate and viscous drag effects become negligible.

What the second paragraph is saying is that for slow-moving, viscous fluids (such as honey in a straw), the forces of friction (fluid "dragging" against the pipe walls) are far greater than the forces of inertia (fluid momentum). This means that the pressure difference required to move such a fluid through a pipe primarily works to overcome the friction of that fluid against the walls of the pipe. For most industrial flows, where the flow velocities are fast and the fluids have little viscosity (like clean water), flow through an orifice plate is assumed to be frictionless. Thus, the pressure dropped across a constriction is *not* the result of friction between the fluid and the pipe, but rather it is a consequence of having to *accelerate* the fluid from a low velocity to a high velocity through the narrow orifice.

My mistake, years ago, was in assuming that water flowing through an orifice generated substantial friction, and that this is what created the ΔP across an orifice plate. This is what my "common sense" told me. In my mind, I imagined the water having to rub past the walls of the pipe, past the face of the orifice plate, and through the constriction of the orifice at a very high speed, in order to make it through to the other side. I memorized what my teacher told us about energy exchange and how pressure had to drop as velocity increased, but I never really internalized it because I still held to my faulty assumption of friction being the dominant mechanism of pressure drop in an orifice plate. In other words, while I could parrot the doctrine of kinetic and potential energy exchange, I was still *thinking* in terms of friction, which is a totally different phenomenon. The difference between these two phenomena is the difference between energy *exchanged* and energy *dissipated*. To use an electrical analogy, it is the difference between *reactance* (X) and *resistance* (R). Incidentally, many electronics students experience the same confusion when they study reactance, mistakenly thinking it is the same thing as resistance where in reality it is quite different in terms of energy, but that is a subject for another essay!

In a frictionless flow stream, fluid pressure decreases as fluid velocity increases in order to conserve energy. Another way to think of this is that a pressure differential must develop in order to provide the "push" needed to *accelerate* the fluid from a low speed to a high speed. Conversely, as the fluid slows back down after having passed through the constriction, a reverse pressure differential must develop in order to provide the "push" needed for that *deceleration*:



A moving mass does not simply slow down on its own! There must be some opposing force to decelerate a mass from a high speed to a low speed. This is where the pressure recovery downstream of the orifice plate comes from. If the pressure differential across an orifice plate originated primarily from friction, as I mistakenly assumed when I first learned about orifice plates, then there would be no reason for the pressure to *ever* recover downstream of the constriction. The presence of friction means energy *lost*, not energy *exchanged*. Although both inertia and friction are capable of creating pressure drops, the lasting effects of these two different phenomena are definitely not the same.

There is a quadratic ("square") relationship between velocity and differential pressure precisely because there is a quadratic relationship between velocity and kinetic energy as all first-quarter physics students learn ($E_k = \frac{1}{2}mv^2$). This is why ΔP increases with the square of flow rate (Q^2)

and why we must "square-root" the ΔP signal to obtain a flow measurement. This is also why fluid density is so important in the orifice-plate flow equation. The denser a fluid is, the more work will be required to accelerate it through a constriction, resulting in greater ΔP, all other conditions being equal:

$$Q = k\sqrt{\frac{\Delta P}{\rho}} \qquad \text{(Our old friend, the "orifice plate" equation)}$$

This equation is only accurate, however, when fluid friction is negligible: when the viscosity of the fluid is so low and/or its speed is so high that the effects of potential and kinetic energy exchange completely overshadow[2] the effects of friction against the pipe walls and against the orifice plate. This is indeed the case for most industrial flow applications, and so this is what students first study as they learn how flow is measured. Unfortunately, this is often the *only* equation two-year Instrumentation students study with regard to flow measurement.

In situations where Reynolds number is low, fluid friction becomes the dominant factor and the standard "orifice plate" equation no longer applies. Here, the ΔP generated by a viscous fluid moving through a pipe really does depend primarily on how "thick" the fluid is. And, just like electrons moving through a resistor in an electric circuit, the pressure drop across the area of friction is directly proportional to the rate of flow ($\Delta P \propto Q$ for fluids, $V \propto I$ for electrons). This is why laminar flowmeters – which work only when Reynolds number is low – yield a nice *linear* relationship between ΔP and flow rate and therefore do not require square root extraction of the ΔP signal. These flowmeters do, however, require temperature compensation (and even temperature *control* in some cases) because flow measurement accuracy depends on fluid viscosity, and fluid viscosity varies according to temperature. The Hagen-Poiseuille equation describing flow rate and differential pressure for laminar flow (low Re) is shown here for comparison:

$$Q = k\left(\frac{\Delta P D^4}{\mu L}\right)$$

Where,

Q = Flow rate (gallons per minute)
k = Unit conversion factor = 7.86 $\times 10^5$
ΔP = Pressure drop (inches of water column)
D = Pipe diameter (inches)
μ = Liquid viscosity (centipoise) – this is a temperature-dependent variable!
L = Length of pipe section (inches)

Note that if the pipe dimensions and fluid viscosity are held constant, the relationship between flow and differential pressure is a direct proportion:

$$Q \propto \Delta P$$

[2] In engineering, this goes by the romantic name of *swamping*. We say that the overshadowing effect "swamps" out all others because of its vastly superior magnitude, and so it is safe (not to mention simpler!) to ignore the smaller effect(s). The most elegant cases of "swamping" are when an engineer intentionally designs a system so the desired effect is many times greater than the undesired effect(s), thereby forcing the system to behave more like the ideal. This application of swamping is prevalent in electrical engineering, where resistors are often added to circuits for the purpose of overshadowing the effects of stray (undesirable) resistance in wiring and components.

In reality, there is no such thing as a frictionless flow (excepting superfluidic cases such as Helium II which are well outside the bounds of normal experience), just as there is no such thing as a massless flow (no inertia). In normal applications there will always be both effects at work. By not considering fluid friction for high Reynolds numbers and not considering fluid density for low Reynolds numbers, engineers draw simplified models of reality which allow us to more easily measure fluid flow. As in so many other areas of study, we exchange accuracy for simplicity, precision for convenience. Problems arise when we forget that we've made this Faustian exchange and wander into areas where our simplistic models are no longer accurate.

Perhaps the most practical upshot of all this for students of Instrumentation is to realize exactly why and how orifice plates work. Bernoulli's equation does *not* include any considerations of friction. To the contrary, we must assume the fluid to be completely frictionless in order for the concept to make sense. This explains several things:

- There is little permanent pressure drop across an orifice: most of the pressure lost at the vena contracta is regained further on downstream as the fluid returns to its original (slow) speed. Permanent pressure drop will occur only where there is energy *lost* through the constriction, such as in cases where fluid friction is substantial. Where the fluid is frictionless there is no mechanism in an orifice to dissipate energy, and so with no energy lost there must be full pressure recovery as the fluid returns to its original speed.

- Pressure tap location makes a difference: to ensure that the downstream tap is actually sensing the pressure at a point where the fluid is moving significantly faster than upstream (the "vena contracta"), and not just anywhere downstream of the orifice. If the pressure drop were due to friction alone, it would be permanent and the downstream tap location would not be as critical.

- Standard orifice plates have knife-edges on their upstream sides: to minimize contact area (friction points) with the high-speed flow.

- Care must be taken to ensure Reynolds number is high enough to permit the use of an orifice plate: if not, the linear $Q/\Delta P$ relationship for viscous flow will assert itself along with the quadratic potential/kinetic energy relationship, causing the overall $Q/\Delta P$ relationship to be polynomial rather than purely quadratic, and thereby corrupting the measurement accuracy.

- Sufficient upstream pipe length is needed to condition flow for orifice plate measurement, not to make it "laminar" as is popularly (and wrongly) believed, but to allow natural turbulence to "flatten" the flow profile for uniform velocity. *Laminar flow* is something that only happens when viscous forces overshadow inertial forces (e.g. flow at low Reynolds numbers), and is totally different from the *fully developed turbulent flow* that orifice plates need for accurate measurement.

In a more general sense, the lesson we should learn here is that blind faith is no substitute for understanding, and that a sense of confusion or disagreement during the learning process is a sign of one or more misconceptions in need of correction. If you find yourself disagreeing with what you are being taught, either you are making a mistake and/or your teacher is. Pursuing your questions to their logical end is the key to discovery, while making a leap of faith (simply believing what you are told) is an act of avoidance: escaping the discomfort of confusion and uncertainty at the expense of a deeper learning experience. This is an exchange no student should feel they have to make.

References

Lipták, Béla G., *Instrument Engineers' Handbook – Process Measurement and Analysis Volume I*, Third Edition, CRC Press, New York, NY.

Appendix B

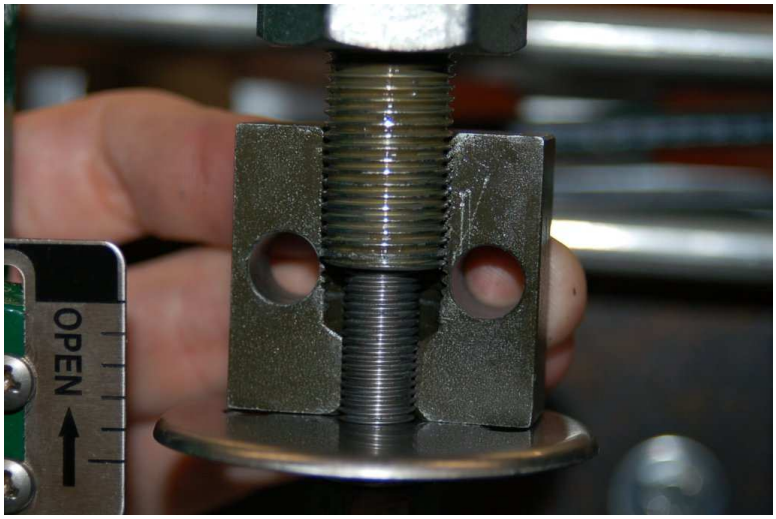
Disassembly of a sliding-stem control valve

The following collection of photographs chronicles the complete disassembly of a Fisher E-body globe valve with pneumatic diaphragm actuator. This control valve design is quite mature, but nevertheless enjoys wide application in modern industrial settings.

This is the complete control valve, without a positioner attached. What you see here is the actuator (painted green) and the valve body (painted grey), mounted on a steel plate for student learning in a laboratory setting. The left-hand photograph shows the complete control valve assembly, while the right-hand photograph shows a student loosening the spanner nut holding the valve actuator yoke to the valve body:

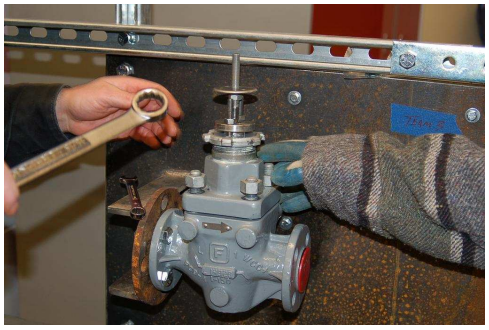
The next step is to un-couple the actuator stem from the valve stem. On Fisher sliding-stem valves, this connection is made by a split block with threads matching those on each stem. Removing two bolts from the block allows it to be taken apart (left-hand photograph). Nuts threaded on to the valve stem, jammed up against the coupling block, must also be loosened before the stems may be uncoupled (right-hand photograph):

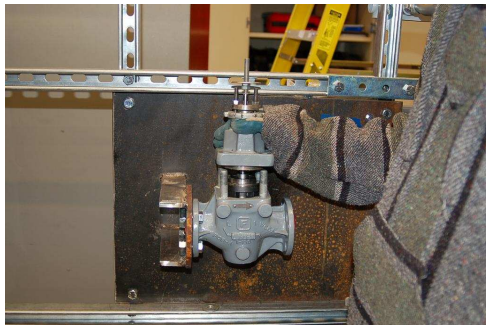
A close-up photograph of this stem connector block, with the front half removed for inspection, shows how it engages both threaded stems (valve and actuator) in a single nut-like assemblage. The solid valve stem (below) slides into the hollow actuator stem (above), while the split connector "nut" engages the threads of both, holding the two stems together so they move up and down as one piece:



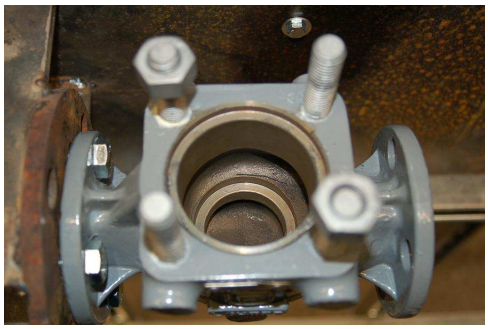
Once the actuator and valve body stems have been uncoupled, the actuator may be removed from the valve body entirely:

The bonnet is held to the rest of the valve body (in this case) by four large studs. Removing the nuts on these studs allows the bonnet to be lifted off the body, exposing the valve trim for view:



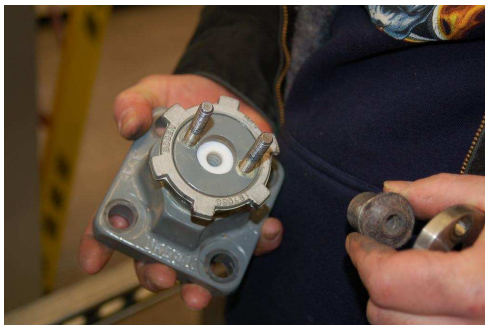


Seats in Fisher E-body globe valves rest in the bottom of the body, held in place by the cage surrounding the valve plug. Once the bonnet is removed from the body, the seat may be removed without need of any specialized tools (left-hand photograph). A view inside the body shows the place where the seat normally rests (right-hand photograph):



With the bonnet removed, the plug and cage may be easily removed for inspection:

The packing follower (between the student's fingers) has been removed from the valve bonnet, and you can see the upper Teflon packing rings within the bonnet. The student is also holding the packing flange in the same hand as the packing follower (left-hand photograph). In the right-hand photograph, we see the student using a screwdriver to gently push the Teflon packing rings out of the bonnet, from the bottom side. Care should be taken not to gouge or otherwise damage these rings during removal:



The left-hand photograph shows all the packing components stacked on top of each other on the concrete floor, next to the bonnet. From top to bottom you see the following components: a felt wiper, the packing follower, five (5) Teflon packing rings, a coil spring, and the packing box ring. The right-hand photograph shows the same packing components stacked on the valve stem:

Turning to the actuator, we begin disassembly by loosening the diaphragm hold-down bolts (left-hand photograph) and removing the upper half of the diaphragm casing (right-hand photograph). A single bolt secures the upper diaphragm plate to the top of the actuator stem:

In the left-hand photograph you see the student removing the spring seat, having previously loosened the spring adjustor nut. With the spring seat removed, the spring may be removed from the actuator assembly. In the right-hand photograph the spring adjustor and spring seat have been removed from the actuator stem. The student is now pointing at the valve spring, partially removed:

Sliding the actuator diaphragm, plate, and stem out of the actuator assembly from the top of the actuator makes it easy to remove the large actuator spring (left-hand photograph). The right-hand photograph shows all the moving actuator components re-assembled in their proper order outside of the yoke:

The left-hand photograph shows the lower half of the actuator casing, with the student removing six (6) hold-down bolts joining this casing half to the actuator yoke. The right-hand photograph shows the actuator casing half completely removed from the yoke, revealing a gasket and the bronze stem bushing (which serves to both guide the actuator stem and seal air pressure, since this is a reverse-acting actuator):

A circular spring clip holds the stem bushing in the yoke casting. The left-hand photograph shows the student using pliers to squeeze this spring clip and remove it from its groove cut into the cast iron of the yoke. In the right-hand photograph, we see the student using the wooden handle of a hammer to gently tap the bushing out of the yoke. The bushing has rubber O-ring seals between it and the yoke casting, so a small amount of force will be necessary to dislodge it. Using the hammer's wooden handle to drive the bushing instead of a metal tool protects the relatively soft bronze bushing from impact damage. Note how the student's right hand is waiting to catch the bronze bushing when it emerges from the hole, to protect it from falling against the hard concrete floor:

The final photograph shows the bushing removed from its hole:

Appendix C

How to use this book – some advice for teachers

If you would like to maximize your students' learning in a field of study that emphasizes critical thinking as much as Instrumentation, I have one simple piece of advice: *engage your students, don't just present information to them.* Do not make the mistake so many teachers do, of thinking it is their role to provide information in pre-digested form to their students, and that it is each student's responsibility to passively absorb this information.

High achievement happens only in an atmosphere of high expectations. If you design coursework allowing students to expend minimal effort, your students will achieve minimal learning. Alternatively, if you require students to think deeply about their subject of study, challenge them with interesting and relevant assignments, and hold them accountable to rigorous standards of demonstrated competence, your students can and will move mountains.

In this appendix I present to you some concepts and models for achieving high standards of learning in the field of Instrumentation. The ideas documented here have all been proven to work in my own instruction, and I continue to use them on a daily basis. However, this is not a rigid blueprint for success – I invite and encourage others to experiment with variations on the same themes. More than anything else, I hope to encourage educators with examples of unconventional thinking and unconventional curricula, to show what may be accomplished if you allow yourself to be creative and objective-driven in your instructional design.

C.1 Teaching technical theory

Learning is not merely a process of information transfer. It is first and foremost a transformation of one's thinking. When we learn something substantial, it alters the way we perceive and interact with the world around us. Learning any subject also involves a substantial accumulation of facts in one's memory, but memorization alone is not really learning (at least it isn't learning at the college level). Transmitting facts into a student's memory is easy – so easy, in fact, that I believe it is a waste of a teacher's time to overly participate in the process. A well-written book does a far better job of conveying facts and concepts than an instructor's live presentation[1]. The instructor's focus during class time should be the development of higher-order thinking skills. This includes (but is not limited to) problem-solving, logical reasoning, diagnostic techniques, and metacognition (critiquing one's own thinking).

Rather than devote most of your classroom time to lecture-style presentations – where the flow of information goes primarily from you to your students – place the responsibility for fact-gathering on your students. Have them read books such as this[2] and arrive at the classroom *prepared* to discuss what they have already studied.

When students are with you in the classroom and in the lab, probe their understanding with questions – lots of questions. Give them realistic problems to solve. Challenge them with projects requiring creative thought. Get your students to reveal how they think, both to you and to their peers. This will transform your classroom atmosphere from a monologue into a dialogue, where you engage with the minds of your students as partners in the learning process instead of lecturing to them as subordinates.

A format I have used with great success is to assign homework to students covering the next topic, so they research that topic in advance of our coverage of it in class. This homework comes in the form reading assignments and in the form of question sets. Some of these questions point students directly to specific texts to read, while others allow students to choose their own research material. When my students arrive for class, I quiz them on some of the basic points of their research (this ensures my students will actually do the research). After the quiz, students choose which sections of the reading assignment or which of the homework questions they would like to present on in class that day, and then they spend a short time working in teams to prepare their presentations. When this small-group time has finished, I call upon those student teams to present what they have found. During this presentation period, my role is to probe students' knowledge with further questions (Socratic dialogue) and "fill in" any important points they may have missed. If at any point in time during the Socratic discussion my students become "stuck" on a difficult concept, I have them break into small groups to discuss and resolve the difficulty. This almost never fails to bear fruitful ideas and re-start the dialogue. The class period concludes with another quiz covering one or more points from the day's discussion.

[1] To be sure, there are some gifted lecturers in the world. However, rather than rely on a human being's live performance, it is better to capture the brilliance of an excellent presentation in static form where it may be peer-reviewed and edited to perfection, then placed into the hands of an unlimited number of students in perpetuity. In other words, if you think you're a great presenter, do us all a favor and translate that brilliance into a format that will reach more people!

[2] It would be arrogant of me to suggest my book is the best source of information for your students. Have them research information on instrumentation from other textbooks, from manufacturers' literature, from whitepapers, from reference manuals, from encyclopedia sets, or whatever source(s) you deem most appropriate. If you possess knowledge that your students need to know that isn't readily found in any book, *publish it for everyone's benefit!*

This teaching method shifts the burden of transmitting facts and concepts from myself to static sources such as textbooks[3]. This shift in responsibility frees valuable class time for more important tasks, namely the refinement of higher-order thinking skills. It is an utter waste of an instructor's talent and time to exhaust a class period transmitting facts to students, when that same talented instructor can use the time to engage students in problem-solving processes, critical thinking, and other cognitive activities of greater challenge and greater importance.

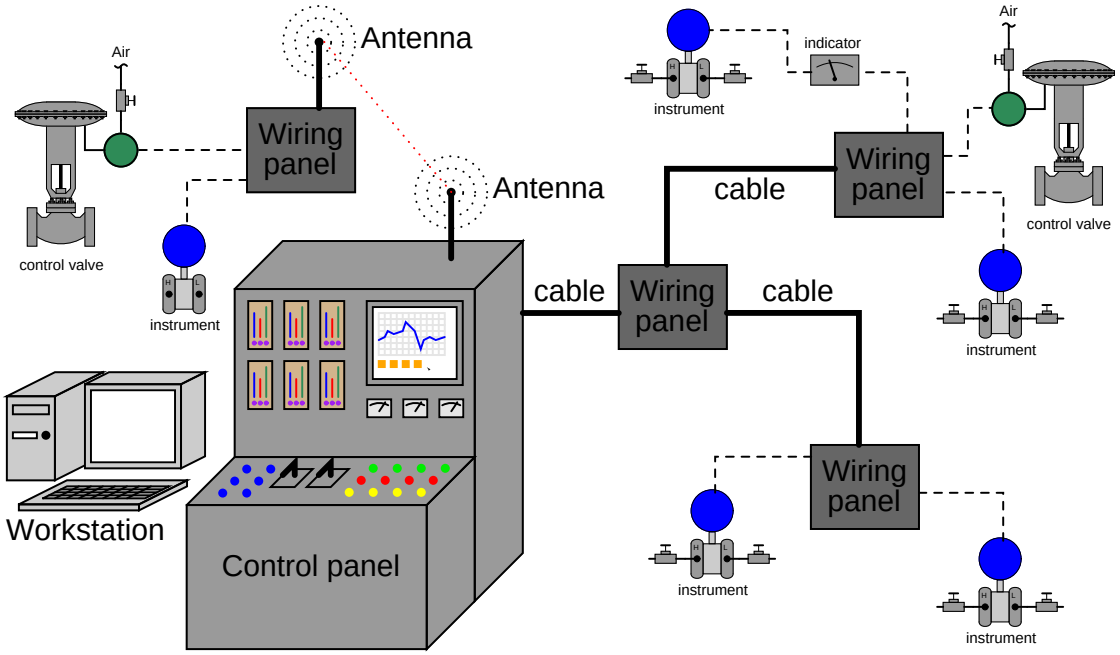
A further benefit of having students perform daily research in preparation for class is students learn how to research, which is no small accomplishment. In a complex field where technology advances on a daily basis, your students will need to be able to learn new facts on their own (without your assistance!) after they graduate. In fact, I would argue this is the single most important skill any person can learn in school: how to efficiently acquire new knowledge on their own. Such a skill not only prepares them for excellence in their chosen career, but it also brings great benefit to every other area of life where the acquisition of new information is essential to decision-making (e.g. participation in the democratic process, legal proceedings, medical decision-making, investing, parenting, etc.).

[3] And multimedia resources, too! With all the advances in multimedia presentations, there is no reason why an instructor cannot build a library of videos, computer simulations, and other engaging resources to present facts and concepts to students outside of class time.

C.2 Teaching technical practices (labwork)

Labwork is an essential part of any science-based curriculum. Here, much improvement may be made over the "standard" educational model to improve student learning. In my students' Instrumentation courses, I forbid the use of pre-built "trainer" systems and lab exercises characterized by step-by-step instructions. Instead, I have my students construct real working instrumentation systems. The heart of this approach is a "multiple-loop" system spanning as large a geographic area as practically possible, with instruments of all kinds connecting to a centralized control room area. None of the instruments need perform any practical purpose, since the goal of the multiple-loop system is for students to learn about the instruments themselves.

A model for a multiple-loop system might look something like this:



Instruments may or may not be grouped together to form complete control systems, since process control is not necessarily the purpose of this system. The primary purpose of a multiple-loop instrument system is to provide an infrastructure for students to investigate instrumentation apart from the dynamics of a functioning process. The separation of controls from process may seem counter-productive at first, but it actually provides a rich and flexible learning experience. Students are able to measure instrument signals and correlate them to actual physical measurements, take instruments in and out of service, check instrument calibration, see the effects of calibration on measurement accuracy and resolution, practice lock-out and tag-out procedures, diagnose instrument problems introduced by the instructor, practice installing and removing instruments, remove old wire and pull new wire into place, practice sketching and editing loop diagrams, and many other practical tasks without having to balance the needs of a working process. The system may be altered at any time as needed, since there are no process operating constraints to restrict maintenance operations.

The fundamental advantage of a process-less instrument system is there are no process limitations restricting educational objectives. In this sense it is as flexible as a computer simulation, but with the advantage of using real-world components.

The first academic year I attempted to build such a system with my students was 2002-2003. Our system cost almost nothing[4], with a control panel fabricated from a discarded fiberglass electrical enclosure and 4-20 mA loop wiring salvaged from discarded spools of category-5 data communications cable (four twisted pairs per cable). We stapled the cable runs to the lab room wall, and used cheap terminal block assemblies to provide connection points between the cat-5 trunk cables and individual instrument cables. Our first loops built with this system were as follows:

- Air compressor receiver tank pressure measurement – *measurement only*
- Air compressor temperature measurement – *measurement only*
- Regulated (service) air pressure measurement – *measurement only*
- Wash basin water level measurement – *measurement only*
- Water column level and temperature control – *measurement and control*
- Air reservoir pressure control – *measurement and control*

The first four of these instrument loops were "permanent" in that they were never disconnected once installed. The water level and temperature control system was a later addition made toward the end of the academic year. It began as a pneumatic system, then was upgraded to electronic (single-loop digital controller), then as a PLC-controlled process, then finally as a DCS-controlled process. The air pressure control system was much the same. All the time we left the process vessels and field instruments in place, used the same signal tubing and wiring, but merely changed the control instruments at the other end of that tubing and wiring.

In addition to these six permanent and semi-permanent loops, students used the system throughout the year to connect individual instruments for loop calibration. Usually there was no control involved, as they were simply studying individual instruments and were not ready for a complete control system yet. Every time they had a transducer to calibrate, a control valve to test, or a transmitter to configure, I required them to tie it into the loop system and document the loop using ISA standard loop diagrams. Then, I would fault their loops (usually electrically by creating opens or shorts in signal wiring, or pneumatically by plugging tubes with foam earplugs) and have them troubleshoot the loops using real test equipment, documenting their diagnostic steps for grading purposes. After successful commissioning, calibration, and troubleshooting, students disassembled the loop so the instruments could be used again in a different loop.

Our multiple-loop instrument system – despite its crude appearance and low cost – was extremely successful as an educational tool. My students gained a tremendous amount of practical knowledge and skill in addition to the basic theory. Abstract principles of measurement and instrument application "came alive" for them as they saw the pieces fit together to make a working system.

[4]Of course, we had to have plenty of instruments to install in this loop system, and industrial instruments are not cheap. My point is that the *infrastructure* of control panel, trunk cabling, field wiring, terminal blocks, etc. was very low-cost. If an Instrumentation program already has an array of field instruments for students to work with in a lab setting, it will not cost much at all to integrate these instruments into a realistic multi-loop system as opposed to having students work with individual instruments on the benchtop or installed in dedicated "trainer" modules.

The intentionally distributed nature of the system – with the control panel located in one far corner of the room and field instruments scattered around the rest of the room – forced students to think and work in a manner much more similar to the real work environment. There were days they were so excited about working on this system that I had to coax them out of class when the school day was over!

In the summer of 2006 I upgraded the loop system to include a 12 foot by 8 foot metal control room panel (donated by a local paper mill), a set of computer workstations for DCS and SCADA system consoles, industry-standard terminal block assemblies located in electrical enclosures, with plenty of electrical conduit runs between different locations in the lab facility to allow pulling of new wires and cables. Students still must connect each instrument they learn about into the system, configuring either a panel-mounted or computer-based display to register the measured variable in proper units (or to receive a control signal if the instrument in question is a final control element). Construction of working control systems (transmitter, controller, valve or motor) is quite easy with this infrastructure in place. The geographically distributed nature of the system lends itself well to realistic troubleshooting, with students working in teams (communicating via hand-held radios) to diagnose problems I intentionally place into the system.

The following photographs show the appearance of the new (2006) multiple loop system, beginning with the control panel and computer workstation cluster. These two elements comprise the "control room area" of the lab:

In another area of the lab room is a pneumatic control panel and a cabinet housing the distributed control system (DCS) I/O rack:

The rest of the lab room is dedicated as a "field area" where field instruments are mounted and wires (or tubes) run to connect those instruments to remote indication and/or control devices:

Note the use of metal strut hardware to form a frame which instruments may be mounted to, and the use of flexible liquid-tight conduit to connect field instruments to rigid conduit pieces so loop wiring is never exposed.

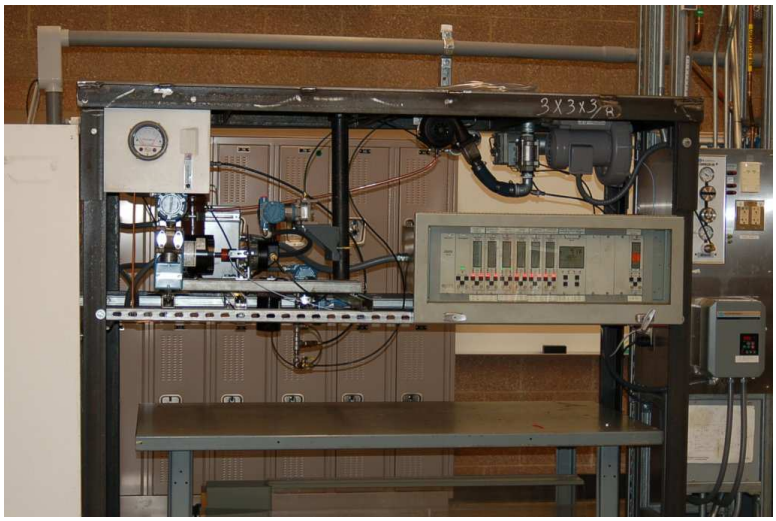
A less expensive alternative[5] to metal strut is standard *industrial pallet racking*, examples shown here with 2 inch pipe attached for instrument mounting, and enclosures attached for instrument cable routing and termination:

The multiple-loop system is designed to be assembled, disassembled, and reassembled repeatedly as each student team works on a new instrument. As such, it is in a constant state of flux. It is not really a *system* so much as it is an *infrastructure* for students to build working loops and control systems within.

[5] When I built my first fully-fledged educational loop system in 2006 at Bellingham Technical College (I built a crude prototype in 2003), I opted for Cooper B-Line metal strut because it seemed the natural choice for the application. It wasn't until 2009 when I needed to expand and upgrade the loop system to accommodate more students that I happened to come up with the idea of using pallet racking as the framework material. Used pallet racking is plentiful, and very inexpensive compared to building a comparable structure out of metal strut. As these photographs show, I still used Cooper B-Line strut for some portions, but the bulk of the framework is simply pallet racking adapted for this unconventional application.

In addition to the multiple-loop system, my students' lab contains working processes (also student-built!) which we improve upon every year. One such process is a water flow/level/temperature control system, shown here:

Another is a turbocompressor system, built around a diesel engine turbocharger (propelled by the discharge of a 2 horsepower air blower) and equipped with a pressurized oil lubrication system and temperature/vibration monitor:



The process piping and equipment is altered only when necessary, but the control systems on these processes undergo major revisions each year when a new group of students takes the coursework relevant to those systems. Having a set of functioning process systems present in the lab at all times also gives students examples of working instrument systems to study as they plan construction of their temporary loops in the multiple-loop system.

C.3 Teaching diagnostic principles and practices

Diagnostic ability is arguably the most difficult skill to develop within a student, and also the most valuable skill a working technician can possess[6]. In this section I will outline several principles and practices teachers may implement in their curricula to teach the science and art of troubleshooting to their students.

First, we need to define what "troubleshooting" is and what it is not. It is *not* the ability to follow printed troubleshooting instructions found in equipment user's manuals[7]. It is *not* the ability to follow one rigid sequence of steps ostensibly applicable to any equipment or system problem[8]. Troubleshooting is first and foremost the practical application of *scientific thinking* to repair of malfunctioning systems. The principles of hypothesis formation, experimental testing, data collection, and re-formulation of hypotheses is the foundation of any detailed cause-and-effect analysis, whether it be applied by scientists performing primary research, by doctors diagnosing their patients' illnesses, or by technicians isolating problems in complex electro-mechanical-chemical system. In order for anyone to attain mastery in troubleshooting skill, they need to possess the following traits:

- A rock-solid understanding of relevant, fundamental principles (e.g. how electric circuits work, how feedback control loops work)
- Close attention to detail
- An open mind, willing to pursue actions led by data and not by preconceived notions

The first of these points is addressed by any suitably rigorous curriculum. The other points are habits of thought, best honed by months of practice. Developing diagnostic skill takes a lot of time, and so the educator must plan for this in curriculum design. It is not enough to sprinkle a few troubleshooting activities throughout a curriculum, or to devote a single class to the topic. Troubleshooting should be a topic tested on every exam, present in every lab activity, and (ideally) touched upon in every day of the student's technical education.

Scientific, diagnostic thinking is characterized by a repeating cycle of *inductive* and *deductive* reasoning. Inductive reasoning is the ability to reach a general conclusion by observing specific details. Deductive reasoning is the ability to predict details from general principles. For example, a student engages in deductive reasoning when they conclude an "open" fault in a series DC circuit will cause current in that circuit to stop. That same student would be thinking inductively if they measured zero current in a DC series circuit and thus concluded there was an "open" fault somewhere in it. Of these two cognitive modes, inductive is by far the more difficult because multiple solutions exist for any one set of data. In our zero-current series circuit example, inductive reasoning might lead the troubleshooter to conclude an open fault existed in the circuit. However, an unpowered source could also be at fault, or for that matter a malfunctioning ammeter falsely registering zero

[6] One of the reasons diagnostic skill is so highly prized in industry is because so few people are actually good at it. This is a classic case of supply and demand establishing the value of a commodity. Demand for technicians who know how to troubleshoot will always be high, because technology will always break. Supply, however, is short because the skill is difficult to teach. This combination elevates the value of diagnostic skill to a very high level.

[7] Yes, I have actually heard people make this claim!

[8] The infamous "divide and conquer" strategy of troubleshooting where the technician works to divide the system into halves, isolating which half the problem is in, is but *one particular procedure: merely one tool in the diagnostician's toolbox*, and does not constitute the whole of diagnostic ability.

current when in fact there is current. Inductive conclusions are *risky* because the leap from specific details to general conclusions always harbor the potential for error. Deductive conclusions are *safe* because they are as secure as the general principles they are built on (e.g. *if* an "open" exists in a series DC circuit, there will be *no* current in the circuit, guaranteed). This is why inductive conclusions are always validated by further deductive tests, not visa-versa. For example, if the student induced that an unpowered voltage source might cause the DC series circuit to exhibit zero current, they might elect to test that hypothesis by measuring voltage directly across the power supply terminals. If voltage is present, then the hypothesis of a dead power source is incorrect. If no voltage is present, the hypothesis is provisionally true[9].

Scientific method is a cyclical application of inductive and deductive reasoning. First, an hypothesis is made from an observation of data (inductive). Next, this hypothesis is checked for validity – an experimental test to see whether or not a prediction founded on that hypothesis is correct (deductive). If the data gathered from the experimental test disproves the hypothesis, the scientist revises the hypothesis to fit the new data (inductive) and the cycle repeats.

Since diagnostic thinking requires both deductive and inductive reasoning, and deductive is the easier of the two modes to engage in, it makes sense for teachers to focus on building deductive skill first. This is relatively easy to do, simply by adding on to the theory and practical exercises students already engage in during their studies.

Both deductive and inductive diagnostic exercises lend themselves very well to Socratic discussions in the classroom, where the instructor poses questions to the students and the students in turn suggest answers to those questions. The next two subsections demonstrate specific examples showing how deductive and inductive reasoning may be exercised and assessed, both in a classroom environment and in a laboratory environment.

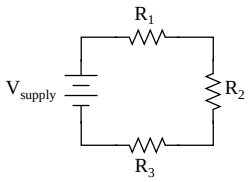
[9] Other things could be at fault. An "open" test lead on the multimeter for example could account for both the zero-current measurement and the zero-voltage measurement. This scientific concept eludes many people: it is far easier to *disprove* an hypothesis than it is to *prove* one. To quote Albert Einstein, "No amount of experimentation can ever prove me right; a single experiment can prove me wrong."

C.3.1 Deductive diagnostic exercises

For example, consider a simple three-resistor series DC circuit, the kind of lab exercise one would naturally expect to see within the first month of education in an Instrumentation program. A typical lab exercise would call for students to construct a three-resistor series DC circuit on a solderless breadboard, predict voltage and current values in the circuit, and validate those predictions using a multimeter. A sample exercise is shown here:

Competency: **Series DC resistor circuit** Version:

Schematic



Given conditions

V_{supply} = R_1 = R_2 = R_3 =

Parameters

	Predicted	Measured		Predicted	Measured
I_{supply}			I_{R1}		
V_{R1}			I_{R2}		
V_{R2}			I_{R3}		
V_{R3}					

Analysis

Relationship between resistor voltage drops and total voltage:

Fault analysis

Suppose component [ ] fails ☐ open ☐ shorted ☐ other _____

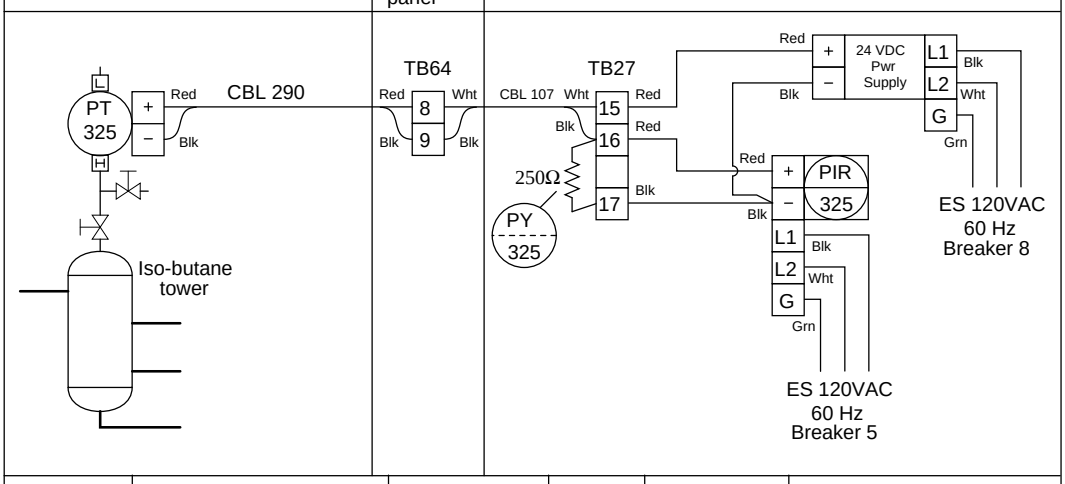
What will happen in the circuit?

Note the **Fault Analysis** section at the end of this page. Here, after the instructor has verified the correctness of the student's mathematical predictions and multimeter measurements, he or she would then challenge the student to predict the effects of a random component fault (either quantitatively or qualitatively), perhaps one of the resistors failing open or shorted. The student makes their predictions, then the instructor simulates that fault in the circuit (either by pulling the resistor out of the solderless breadboard to simulate an "open" or placing a jumper wire in parallel with the resistor to simulate a "short"). The student then uses his or her multimeter to verify the predictions. If the predicted results do not agree with the real measurements, the instructor

works with the student to identify why their prediction(s) were faulty and hopefully correct any misconceptions leading to the incorrect result(s). Finally, a different component fault is chosen by the instructor, predictions made by the student, and verification made using a multimeter. The actual amount of time added to the instructor's validation of student lab completion is relatively minor, but the benefits of exercising deductive diagnostic processes are great.

An example of a more advanced deductive diagnostic exercise appropriate to later phases of a student's Instrumentation education appears here. A loop diagram shows a pressure recording system for an iso-butane distillation column:

Loop Diagram: Iso-butane tower pressure | **Revised by:** D.P. Cell | **Date:** April 1, 2003



Tag number	Description	Manufacturer	Model	Calibration	Notes
PT-325	Gauge pressure transmitter	Rosemount	1151GP	0-100 PSIG 4-20 mA	Check calibration monthly
PY-325	250 Ω resistor	n/a	n/a	+/- 0.1 %	
PIR-325	Pressure indicating recorder	Leeds & Northrup	165	1-5 VDC	Shared by TIR 244 and AIR 300

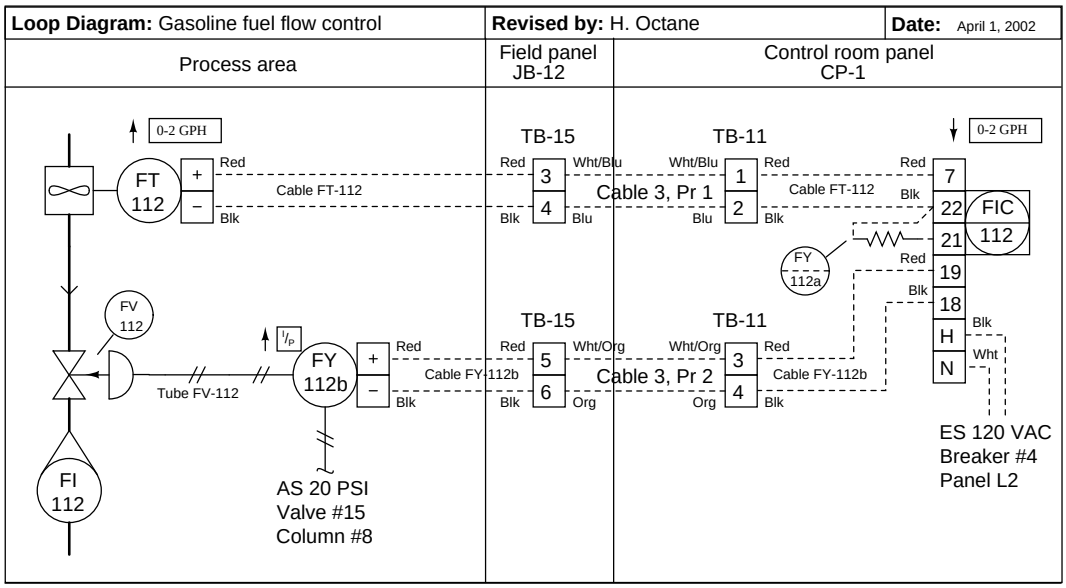
A set of questions accompanying this diagram challenge each student to predict effects in the instrument system resulting from known faults, such as:

- PT-325 block valve left shut and bleed valve left open (*predict voltage between TB27-16 and TB27-17*)
- Loose wire connection at TB64-9 (*predict pressure indication at PIR-325*)
- Circuit breaker #5 shut off (*predict loop current at applied pressure of 50 PSI*)

Given each hypothetical fault, there is only one correct conclusion for any given question. This makes deductive exercises unambiguous to assess.

C.3.2 Inductive diagnostic exercises

Building inductive diagnostic reasoning skill in students requires a lot of practice troubleshooting faulted systems. This is best done through hands-on troubleshooting exercises (e.g. students work to locate and identify faults placed by the instructor in a working instrument loop, built using a loop system for increased realism), although it is not impossible to implement on a written exam. Take for instance this exam question:



This system used to work just fine, but now it has a problem: the controller registers zero flow, and its output signal (to the valve) is saturated at 100% (wide open) as though it were trying to "ask" the valve for more flow. Your first diagnostic step is to check to see if there actually is gasoline flow through the flowmeter and valve by looking at the rotameter. The rotameter registers a flow rate in excess of 2 gallons per hour.

Identify possible faults in this system that could account for the controller's condition (no flow registered, saturated 100% output), depending on what you find when you look at the rotameter:

- Possible fault:
- Possible fault:

Here, the student must identify two probably faults to account for all exhibited symptoms. More than two different kinds of faults are possible[10], but the student need only identify two faults independently capable of causing the controller to register zero flow when it should be registering more than 2 GPH.

[10] Jammed turbine wheel in flowmeter, failed pickup coil in flowmeter, open wire in cable FT-112 or pair 1 of cable 3 (assuming the flow controller's display was not configured to register below 0% in an open-loop condition), etc.

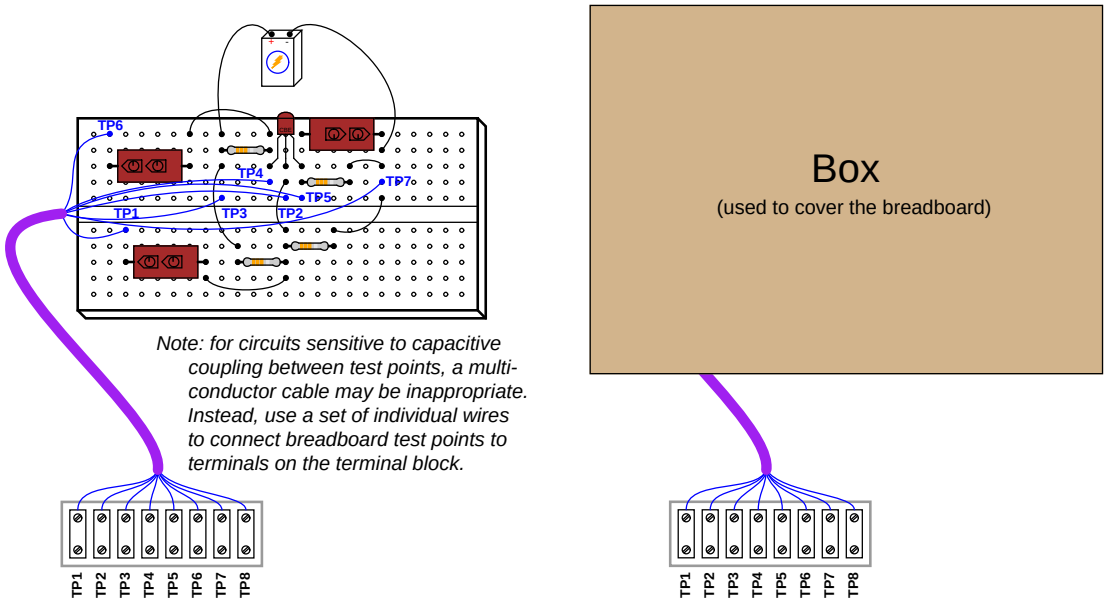
Creating realistic troubleshooting exercises for students is a matter of building (or having students build) working systems that may be faulted by the instructor without the fault being visually evident to the student. A couple of possibilities will be explored here:

- Troubleshooting "harnesses" for solderless breadboard circuits
- Multi-loop instrument system

Solderless breadboards are universally used in the teaching of basic electronics, because they allow students to quickly and efficiently build different circuits using replaceable components. As wonderful as breadboards are for fast construction of electronic circuits, however, it is virtually impossible to create a realistic component fault without the fault being evident to the student simply by visual inspection. In order for a breadboard to provide a realistic *diagnostic* scenario, you must find a way to hide the circuit while still allowing access to certain test points in the circuit.

A simple way to accomplish this is to build a "troubleshooting harness" consisting of a multi-terminal block connected to a multi-conductor cable. Students are given instructions to connect various wires of this cable to critical points in the circuit, then cover up the breadboard with a five-sided box so that the circuit can no longer be seen. Test voltages are measured between terminals on the block, not by touching test leads to component leads on the breadboard (since the breadboard is now inaccessible).

The following illustration shows what this looks like when applied to a single-transistor amplifier circuit:



If students cannot visually detect a fault, they must rely on voltage measurements taken from terminals on the block. This is quite challenging, as not even the shapes of the components may be seen with the box in place. The only guide students have for relating terminal block test points to points in the circuit is the schematic diagram, which is good practice because it forces students to interpret and follow the schematic diagram.

Of course this technique works well for electronic circuits, but what about whole instrument measurement and control systems? Here, the only realistic solution is to build (or have students build) working instrument systems for you (the instructor) to fault. The construction of a large system may be expedited by having students work in teams, but in order to ensure individual diagnostic competence, students must be tested individually.

In my lab courses, students work in teams to build functioning measurement and control loops using the infrastructure of a multiple-loop system (see Appendix section C.2 beginning on page 1756 for a detailed description). Teamwork helps expedite the task of constructing each loop, such that even an inexperienced team is able to assemble a working loop (transmitter connected to an indicator or controller, with wires pulled through conduits and neatly landed on terminal blocks) in just a few hours.

Each student creates their own loop diagram showing all instruments, wires, and connection points, following ISA standards. These loop diagrams are verified by doing a "walk-through" of the loop with all student team members present. The "walk-through" allows the instructor to inspect work quality and ensure any necessary corrections are made to the diagrams. After each team's loop has been inspected and all student loop diagrams edited, the diagrams are placed in a document folder accessible to all students in the lab area.

Once the loop is wired, calibrated, inspected, and documented, it is ready to be faulted. When a student is ready to begin their diagnostic exercise, they gather their team members and approach the instructor. The instructor selects a loop diagram from the document folder *not* drawn by that student, ideally of a loop constructed by another team. The student and teammates leave the lab room, giving the instructor time to fault the loop. Possible faults include:

- Loosen wire connections
- Short wire connections (loose strands of copper strategically placed to short adjacent terminals together)
- Cut cables in hard-to-see locations
- Connect wires to the wrong terminals
- Connect wire pairs backward
- Mis-configure instrument calibration ranges
- Insert square root extraction where it is not appropriate
- Mis-configure controller action or display
- Insert unrealistically large damping constants in either the transmitter, indicator, or final element
- Plug pneumatic signal lines with foam earplugs
- Turn off hand valves
- Trip circuit breakers

After the fault has been inserted, the instructor calls the student team back into the lab area (ideally using a hand-held radio, simulating the work environment of a large industrial facility where technicians carry two-way radios) to describe the symptoms. This part of the exercise works best when the instructor acts the part of a bewildered operator, describing what the system is not doing correctly, without giving any practical advice on the location of the problem or how to fix it[11]. An important detail for the instructor to include is the "history" of the fault: is this a new loop which has never worked correctly, or was it a working system that failed? Faults such as mis-connected wires are realistic of improper installation (new loop), while faults such as loose connections are perfectly appropriate for previously working systems. Whether the instructor freely offers this "history" or waits for the student to ask, it is important to include in the diagnostic scenario because it is an extremely useful piece of information to know while troubleshooting actual systems in industry. Virtually anything may be wrong (including multiple faults) in a brand-new installation, whereas previously working systems tend to fail in fewer ways.

After this introduction, the one student begins his or her diagnosis, with the other team members acting as scribes to document the student's steps. The diagnosing student may ask a teammate for

[11] I must confess to having a lot of fun here. Sometimes I even try to describe the problem incorrectly. For instance, if the problem is a huge damping constant, I might tell the student that the instrument simply does not respond, because that it what it looks like it you do not take the time to watch it respond *very slowly*.

manual assistance (e.g. operating a controller while the student observes a control valve's motion), but no one is allowed to help the one student diagnose the problem. After a set period of time of instructor observation, if the student has not been able to at least locate the approximate nature of the fault (e.g. "it's a problem in the transmitter"), the exercise is aborted and the instructor reviews the student's actions (as documented by the teammates) to help the student understand where they went wrong in their diagnosis. Otherwise, the student is given more time[12] to pinpoint the nature of the fault.

Depending on the sequencing of your students' coursework, some diagnostic exercises may include components unfamiliar to the student. For example, a relatively new student familiar only with the overall function of a control loop but intimately familiar with the workings of measurement devices may be asked to troubleshoot a loop where the fault is in the control valve positioner rather than in the transmitter. I still consider this to be a fair assessment of the student's diagnostic ability, so long as the expectations are commensurate with the student's knowledge. I would not expect a student to precisely locate the nature of a positioner fault if they had never studied the function or configuration of a valve positioner, but I would expect them to be able to broadly identify the location of the fault (e.g. "it's somewhere in the valve") so long as they knew how a control signal is supposed to command a control valve to move. That student should be able to determine by manually adjusting the controller output and measuring the output signal with the appropriate loop-testing tools that the valve was not responding as it should despite the controller properly performing its function. The ability to diagnose problems in instrument systems where some components of the system are mysterious "black boxes" is a very important skill, because your students *will* have to do exactly that when they step into industry and work with specific pieces of equipment they never had time to learn about in school[13].

I find it nearly impossible to fairly assign a letter or percentage grade to any particular troubleshooting effort, because no two scenarios are quite the same. Mastery assessment (either pass or fail, with multiple opportunities to re-try) seems a better fit. Mastery assessment with no-penalty retries also has the distinct advantage of directing more attention and providing more practice for weaker students: the weaker a student is in troubleshooting, the more they must exercise their skills.

Successfully passing a troubleshooting exercise requires not only that the fault be correctly identified and located in a timely manner, but that all steps leading to the diagnosis are logically justified. Random "trial and error" tests by the student will result in a failed attempt, even if the student was eventually able to locate the fault. A diagnosis with no active tests such as multimeter or test gauge measurements, or actions designed to stimulate system components, will also fail to pass. For example, a student who successfully locates a bad wiring connection by randomly tugging

[12] The instructor may opt to step away from the group at this time and allow the student to proceed unsupervised for some time before returning to observe.

[13] I distinctly remember a time during my first assignment as an industrial instrument technician that I had to troubleshoot a problem in a loop where the transmitter was an oxygen analyzer. I had no idea how this particular analyzer functioned, but I realized from the loop documentation that it measured oxygen concentration and output a signal corresponding to the percentage concentration (0 to 21 percent) of O_2. By subjecting the analyzer to known concentrations of oxygen (ambient air for 21%, inert gas for 0%) I was able to determine the analyzer was responding quite well, and that the problem lie elsewhere in the system. If the analyzer had failed my simple calibration test, I would have known there was something wrong with it, which would have led me to either get help from other technicians working at that facility or simply replace the analyzer with a new unit and try to learn about and repair the old unit in the shop. In other words, my ignorance of the transmitter's specific workings did not prevent me from diagnosing the loop in general.

at every wire connection should *not* pass the troubleshooting exercise because such actions do not demonstrate diagnostic thinking[14].

With a multiple-loop system infrastructure placed in an educational lab setting, more than one student is able to perform a diagnostic exercise at any given time. In fact, the only hard limit to "student throughput" is the number of constructed loops in the system at any given time. Time limits placed on each effort help avoid wasted time and increase the number of exercises that may be performed in any given time period. With the teammate-scribe method, each student gets assessed on individual effort yet enjoys the further educational benefit of seeing how other people solve problems. Student engagement is maximized and instructor burden minimized when multiple teams are working simultaneously to locate faults in a multiple-loop system.

To summarize key points of diagnostic exercises using a multiple-loop system:

- Students work in teams to build each loop
- Loop inspection and documentation finalized by a "walk-through" with the instructor
- Instructor placement of faults (it is important no student knows what is wrong with the loop!)
- Each student individually diagnoses a loop, with team members acting merely as scribes
- Students must use loop diagrams drawn by someone else, ideally diagnosing a loop built by a different team
- Time limit for each student to narrow the scope of the problem
- Passing a diagnostic exercise requires:

  → Accurate identification of the problem

  → Each diagnostic step logically justified by previous results

  → Tests (measurements, component response checks) performed before reaching conclusions
- Mastery (pass/fail) assessment of each attempt, with multiple opportunities for re-tries if necessary

[14] Anyone can (eventually) find a fault if they check every detail of the system. Randomly probing wire connections or aimlessly searching through a digital instrument's configuration is not troubleshooting. I have seen technicians waste incredible amounts of time on the job randomly searching for faults, when they could have proceeded much more efficiently by taking a few multimeter measurements and/or stimulating the system in ways revealing what and where the problem is. One of your tasks as a technical educator is to discourage this bad habit by refusing to tolerate non-diagnostic behavior in a troubleshooting exercise!

C.4 Assessing student learning

When the time comes to assess your students' learning, prioritize performance assessment over written or verbal response. In other words, require that your students *demonstrate* their competence rather than merely explain it. Performance assessment takes more time than written exams, but the results are well worth it. Not only will you achieve a more valid measurement of your students' learning, but they will experience greater motivation to learn because they know they must put their learning into action.

Make liberal use of *mastery* assessments in essential knowledge and skill domains, where students must repeat a demonstration of competence as many times as necessary to achieve perfect performance. Not only does this absolutely guarantee students will learn what they should, but the prospect of receiving multiple opportunities to demonstrate knowledge or skill has the beneficial effect of relieving psychological stress for the student. Mastery assessment lends itself very well to the measurement of diagnostic ability.

An idea I picked up through a discussion on an online forum with someone from England regarding engineering education is the idea of breaking exams into two parts: a *mastery* portion and a *proportional* portion. In each course section, students must pass the mastery exam with 100% accuracy before they can take the proportional exam. A limited number of opportunities are given to re-take the mastery exam, with points deducted from the proportional exam score if the mastery exam is not passed on the first try. Mastery exams cover all the basic concepts, with very straight-forward questions (no tricks or ambiguous wording). The proportional exam, by contrast, is a single-effort test filled with challenging problems requiring high-level thinking. By dividing exams into two parts, it is possible to guarantee the entire class has mastered basic concepts while challenging even the most capable students.

Another unconventional assessment strategy is to create multi-stage exams, where the grade or score received for the exam depends on the highest level passed. This is how I assess students on PLC programming: a large number of programming projects are provided as examples, each one fitting into one of four categories of increasing difficulty. The first level is the minimum required to pass the course, while the fourth level is so challenging that only a few students will be able to pass it in the time given. For each of these levels, the student is given the design parameters (e.g. "program a motor start-stop system with a timed lockout preventing a re-start until at least 15 seconds has elapsed"); a micro-PLC; a laptop computer with the PLC programming software; the necessary switches, relays, motors, and other necessary hardware; and 1 hour of time to build and program a working system. There are too many example projects provided for any student to memorize solutions to them all, especially when no notes are allowed during the assessment (only manufacturer's documentation for the PLC and other hardware). This means the student must demonstrate both mastery of the basic PLC programming and wiring elements, as well as creative design skills to arrive at their own solution to the programming problem. There is no limit to the number of attempts a student may take to pass a given level, and no penalty for failed efforts. Best of all, this assessment method demands little of the instructor, as the working project "grades" itself.

My philosophy on assessment is that good assessment is actually more important than good instruction. If the assessments are valid and rigorous, student learning (and instructor teaching!) will rise to meet the challenge. However, even the best instruction will fail to produce consistently high levels of student achievement if students know their learning will never be rigorously assessed. In a phrase, *good assessment drives the learning process.*

For those who might worry about an emphasis on assessment encouraging teachers to "teach to the test," I offer this advice: there is nothing wrong with teaching to the test so long as the test is valid! Educators usually avoid teaching to the test out of a fear students might pass the test(s) without actually learning what they are supposed to gain from taking the course. If this is even possible, it reveals a fundamental problem with the test: it does not actually measure what you want students to know. A valid test is one that cannot be "foiled" by teaching in any particular way. Valid tests challenge students to think, and cannot be passed through memorization. Valid tests avoid asking for simple responses, demanding students articulate reasoning in their answers. *Valid tests are passable only by demonstrated competence.*

Another important element of assessment is long-term review. You should design the courses in such a way that important knowledge and skill areas are assessed on an ongoing basis up through graduation. Frequent review is a key element to attaining mastery.

C.5 Summary

To summarize some of the key points and concepts for teaching:

- Do not waste class time transmitting facts to students – let the students research facts outside of class
- Use class time to develop high-level thinking skills (e.g. problem-solving, diagnostic techniques, metacognition).
- Use Socratic dialogue and small-group collaboration to get students engaged with the subject matter.
- Make labwork as realistic as possible.
- Build diagnostic skill by first exercising deductive reasoning, as a prelude to inductive reasoning.
- Incorporate frequent troubleshooting exercises in the lab, with students diagnosing realistic faults in instrument systems.
- Assess student learning validly and rigorously.
- Review important knowledge and skill areas continually until graduation. Build this review into the program courses themselves (homework, quizzes, exams) rather than relying on ad hoc review.

One final piece of advice for educators at every level: *it is better to teach a few things well than a lot of things poorly!* If external constraints force you to "cover" too much material in too little time, focus on making each learning exercise as integrative as possible, so that at least students get to experience different topics in ways that reinforce and give context to each other.

Appendix D

Contributors

This is an open-source book, which means everyone has a legal write to modify it to their liking. As the author, I freely accept input from readers that will make this book better. This appendix exists to give credit to those readers who have helped me improve the book.

D.1 Error corrections

Sangani, Champa

- Identified calculation error in milliamp-to-pH scaling problem (*Analog Electronic Instrumentation* chapter), September 2009.

Thompson, Brice

- Identified errors in high/low select and high/low limit function illustrations (*Basic Process Control Strategies* chapter), June 2009.

D.2 New content

Goertz, Kevin

- Took photographs of various flowmeters, control valves, and an insertion pH probe assembly, 2006-2007.

Appendix E

Creative Commons Attribution License

E.1 A simple explanation of your rights

This is an "open-source" textbook, meaning the digital files used to create the final form (PDF, printed paper, or other) are freely available for your perusal, reproduction, distribution, and even modification. These files reside at the following website:

`http://openbookproject.net/books/socratic/sinst/book/`

The Creative Commons Attribution license grants you (the recipient), as well as anyone who might receive my work from you, the right to freely use it. This license also grants you (and others) the right to modify my work, so long as you properly credit my original authorship. My work is copyrighted under United States law, but this license grants everyone else in the world certain freedoms not customarily available under full copyright. This means no one needs to ask my permission, or pay any royalties to me, in order to read, copy, distribute, publish, or otherwise use this book.

If you choose to modify my work, you will have created what legal professionals refer to as a *derivative work*. The Creative Commons license broadly groups derivative works under the term *adaptations*. In simple terms, the fundamental restriction placed on you when you do this is you must properly credit me for the portions of your adaptation that are my original work. Otherwise, you may treat your adaptation the same way you would treat a completely original work of your own. This means you are legally permitted to enjoy full copyright protection for your adaptation, up to and including exclusive rights of reproduction and distribution. In other words, this license does *not* bind your derivative work under the same terms and conditions I used to release my original work.

The practical upshot of this is you may modify my work and re-publish it as you would any other book, with the full legal right to demand royalties, restrict distributions, etc. This does not compromise the freedom of my original work, because that is still available to everyone under the terms and conditions of the Attribution license[1]. It does, however, protect the investment(s) you make in creating the adaptation by allowing you to release the adaptation under whatever terms you see fit (so long as those terms comply with current intellectual property laws, of course).

In summary, the following "legalese" is actually a very good thing for you, the reader of my book. It grants you permission to do so much more with this text than what you would be legally allowed to do with any other (traditionally copyrighted) book. It also opens the door to open collaborative development, so it might grow into something far better than what I alone could create.

[1] You *cannot* pass my original work to anyone else under different terms or conditions than the Attribution license. That is called *sublicensing*, and the Attribution license forbids it. In fact, any re-distribution of my original work must come with a notice to the Attribution license, so anyone receiving the book through you knows their rights.

E.2 Legal code

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE ("CCPL" OR "LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

1. "Adaptation" means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this License. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered an Adaptation for the purpose of this License.

2. "Collection" means a collection of literary or artistic works, such as encyclopedias and anthologies, or performances, phonograms or broadcasts, or other works or subject matter other than works listed in Section 1(f) below, which, by reason of the selection and arrangement of their contents, constitute intellectual creations, in which the Work is included in its entirety in unmodified form along with one or more other contributions, each constituting separate and independent works in themselves, which together are assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined above) for the purposes of this License.

3. "Distribute" means to make available to the public the original and copies of the Work or Adaptation, as appropriate, through sale or other transfer of ownership.

4. "Licensor" means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License.

5. "Original Author" means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i) in the case of a performance the actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast.

6. "Work" means the literary and/or artistic work offered under the terms of this License including without limitation any production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression including digital form, such as a book, pamphlet and other writing; a lecture, address, sermon or other work of the same nature; a dramatic or dramatico-musical work; a choreographic work or entertainment in dumb show; a musical composition with

or without words; a cinematographic work to which are assimilated works expressed by a process analogous to cinematography; a work of drawing, painting, architecture, sculpture, engraving or lithography; a photographic work to which are assimilated works expressed by a process analogous to photography; a work of applied art; an illustration, map, plan, sketch or three-dimensional work relative to geography, topography, architecture or science; a performance; a broadcast; a phonogram; a compilation of data to the extent it is protected as a copyrightable work; or a work performed by a variety or circus performer to the extent it is not otherwise considered a literary or artistic work.

7. "You" means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

8. "Publicly Perform" means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images.

9. "Reproduce" means to make copies of the Work by any means including without limitation by sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage of a protected performance or phonogram in digital form or other electronic medium.

2. Fair Dealing Rights. Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.

3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

1. to Reproduce the Work, to incorporate the Work into one or more Collections, and to Reproduce the Work as incorporated in the Collections;

2. to create and Reproduce Adaptations provided that any such Adaptation, including any translation in any medium, takes reasonable steps to clearly label, demarcate or otherwise identify that changes were made to the original Work. For example, a translation could be marked "The original work was translated from English to Spanish," or a modification could indicate "The original work has been modified.";

3. to Distribute and Publicly Perform the Work including as incorporated in Collections; and,

4. to Distribute and Publicly Perform Adaptations.

5. For the avoidance of doubt:

1. Non-waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;

2. Waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor waives the exclusive right to collect such royalties for any exercise by You of the rights granted under this License; and,

3. Voluntary License Schemes. The Licensor waives the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License.

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. Subject to Section 8(f), all rights not expressly granted by Licensor are hereby reserved.

4. Restrictions. The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

1. You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(b), as requested. If You create an Adaptation, upon notice from any Licensor You must, to the extent practicable, remove from the Adaptation any credit as required by Section 4(b), as requested.

2. If You Distribute, or Publicly Perform the Work or any Adaptations or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution ("Attribution Parties") in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if supplied; (iii) to the extent reasonably practicable, the URI, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and (iv) , consistent with Section 3(b), in the case of an Adaptation, a credit identifying the use of the Work in the Adaptation (e.g., "French translation of the Work by Original Author," or "Screenplay based on original Work by Original Author"). The credit required by this Section 4 (b) may be implemented in any reasonable manner; provided, however, that in the case of a Adaptation or Collection, at a minimum such credit will appear, if a credit for all contributing authors of the Adaptation or Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written

permission of the Original Author, Licensor and/or Attribution Parties.

3. Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Adaptations or Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author's honor or reputation. Licensor agrees that in those jurisdictions (e.g. Japan), in which any exercise of the right granted in Section 3(b) of this License (the right to make Adaptations) would be deemed to be a distortion, mutilation, modification or other derogatory action prejudicial to the Original Author's honor and reputation, the Licensor will waive or not assert, as appropriate, this Section, to the fullest extent permitted by the applicable national law, to enable You to reasonably exercise Your right under Section 3(b) of this License (right to make Adaptations) but not otherwise.

5. Representations, Warranties and Disclaimer

UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTIBILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

1. This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Adaptations or Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

2. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

1. Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.

2. Each time You Distribute or Publicly Perform an Adaptation, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.

3. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

4. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

5. This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.

6. The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced according to the corresponding provisions of the implementation of those treaty provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights not granted under this License, such additional rights are deemed to be included in the License; this License is not intended to restrict the license of any rights under applicable law.

Creative Commons Notice

Creative Commons is not a party to this License, and makes no warranty whatsoever in connection with the Work. Creative Commons will not be liable to You or any party on any legal theory for any damages whatsoever, including without limitation any general, special, incidental or consequential damages arising in connection to this license. Notwithstanding the foregoing two (2) sentences, if Creative Commons has expressly identified itself as the Licensor hereunder, it shall have all rights and obligations of Licensor.

Except for the limited purpose of indicating to the public that the Work is licensed under the CCPL, Creative Commons does not authorize the use by either party of the trademark "Creative Commons" or any related trademark or logo of Creative Commons without the prior written consent of Creative Commons. Any permitted use will be in compliance with Creative Commons' then-current trademark usage guidelines, as may be published on its website or otherwise made available upon request from time to time. For the avoidance of doubt, this trademark restriction does not form part of this License.

Creative Commons may be contacted at `http://creativecommons.org/`

Index