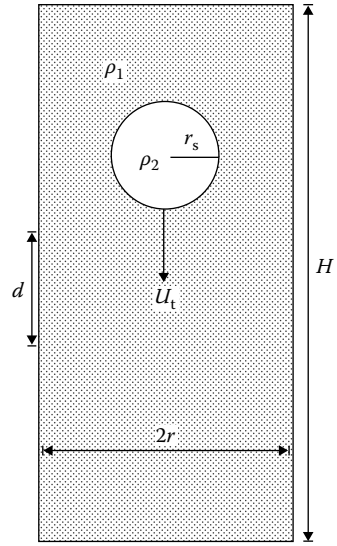


FIGURE 46.7 Schematic diagram of the falling sphere viscometer. Visual observations of the time taken for the sphere to traverse the distance d are used to determine a velocity of the sphere. The calculated velocity is then used in Equation 46.24 to determine a shear viscosity.

length of the cylinder. As the sphere passes through the marked region of length d at its terminal velocity, a measure of the time taken to traverse this distance allows the velocity of the sphere to be calculated. Using Equation 46.26, the shear viscosity of the liquid can be determined.

This method is useful for liquids with viscosities between 10^{-3} and 10^{5} Pa·s.

46.5.2 Falling Cylinder

The *falling cylinder method* is similar in concept to the falling sphere method except that a flat-ended, solid circular cylinder freely falls vertically in the direction of its longitudinal axis through a liquid sample within a cylindrical container. A schematic diagram of the configuration is shown in Figure 46.8. Taking an infinitely long cylinder of density ρ_2 and radius r_c falling through a Newtonian fluid of density ρ_1 with infinite extent, the resulting shear viscosity of the fluid is given as

$$\eta = g r_c^2 \frac{(\rho_2 - \rho_1)}{2U_t} \tag{46.29}$$

Just as with the falling sphere, a finite container volume necessitates modifying Equation 46.29 to account for the effects of container walls and ends. A correction for container wall effects can be analytically deduced by balancing the buoyancy and gravitational forces on the cylinder, of length L, with the shear force on the sides and the compressional force on the cylinder's leading end and the tensile force on the cylinder's trailing end. The resulting correction term, or geometrical factor, $G(k)$ (where $k = r_c/r$), depends on the cylinder radius and the container radius, r, and is given by

$$G(k) = \frac{[k^2(1 - \ln k) - (1 + \ln k)]}{(1 + k^2)} \tag{46.30}$$

Unlike the fluid flow around a falling sphere, the fluid motion around a falling flat-ended cylinder is very complex. The effects of container ends are minimized by creating a small gap between the

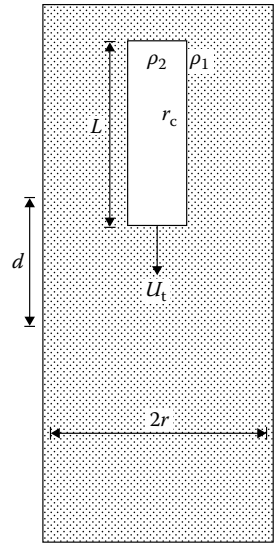


FIGURE 46.8 Schematic diagram of the falling cylinder viscometer. Using the same principle as the falling sphere, the velocity of the cylinder is obtained, which is needed to determine the shear viscosity of the fluid.

cylinder and the container wall. If a long cylinder (here, a cylinder is considered long if $\psi \geq 10$, where $\psi = L/r$) with a radius nearly as large as the radius of the container is used, then the effects of the walls would dominate, thereby reducing the end effects to a second-order effect. A major drawback with this approach is, however, if the cylinder and container are not concentric, the resulting inhomogeneous wall shear force would cause the downward motion of the cylinder to become eccentric. The potential for misalignment motivated the recently obtained analytical solution to the fluid flow about the cylinder ends [16]. An analytical expression for the end correction factor (ECF) was then deduced [17] and is given as

$$\frac{1}{\text{ECF}} = 1 + \left(\frac{8k}{\pi C_w}\right)\left(\frac{\text{G}(k)}{\psi}\right) \tag{46.31}$$

where $C_w = 1.003852 - 1.961019k + 0.9570952k^2$. C_w was derived semi-empirically [17] as a disk wall correction factor. This is based on the idea that the drag force on the ends of the cylinder can be described by the drag force on a disk. Equation 46.31 is valid for $\psi \leq 30$ and agrees with the empirically derived correction [16] to within 0.6%.

With wall and end effects taken into consideration, the working formula to determine the shear viscosity of a Newtonian fluid from a falling cylinder viscometer is

$$\eta = \frac{[gr_c^2(\rho_2 - \rho_1)\,\text{G}(k)]}{(2U_t/\text{ECF})} \tag{46.32}$$

In the past, this method was primarily used as a method to determine relative viscosities between transparent fluids. It has only been since the introduction of the ECF [16,17] that this method could be rigorously used as an absolute viscosity method. With a properly designed container and cylinder, this method is now able to provide accurate absolute viscosities from 10^{-3} to 10^{7} Pa·s.

46.5.3 Falling Methods in Opaque Liquids

The falling body methods described earlier have been extensively applied to transparent liquids where optical (often visual) observation of the falling body is possible. For *opaque liquids*, however, falling body methods require the use of some sensing technique to determine, often in situ, the position of the falling body with respect to time. Techniques have varied, but they all have in common the ability to detect the body as it moves past the sensor. A recent study at high pressure [18] demonstrated that the contrast in electric conductivity between a sphere and opaque liquid could be exploited to dynamically sense the moving sphere if suitably placed electrodes penetrated the container walls as shown in the schematic diagram in Figure 46.9. References to other similar in situ techniques are given in [18].

46.5.4 Rising Bubble/Droplet

For many industrial processes, the rising bubble viscometer has been used as a method of comparing the relative viscosities of transparent liquids (such as varnish, lacquer, and beer) for decades. Although its use was widespread, the actual behavior of the bubble in a viscous liquid was not well understood until long after the method was introduced [19]. The rising bubble method has been thought of as a derivative of the falling sphere method; however, there are fundamental physical differences between the two. The major physical differences are as follows: (1) The density of the bubble is less that of the surrounding liquid, the compressibility of the bubble leads to a change in bubble volume depending on its rise position in the fluid column, and (2) the bubble itself has some unique viscosity. Each of these differences

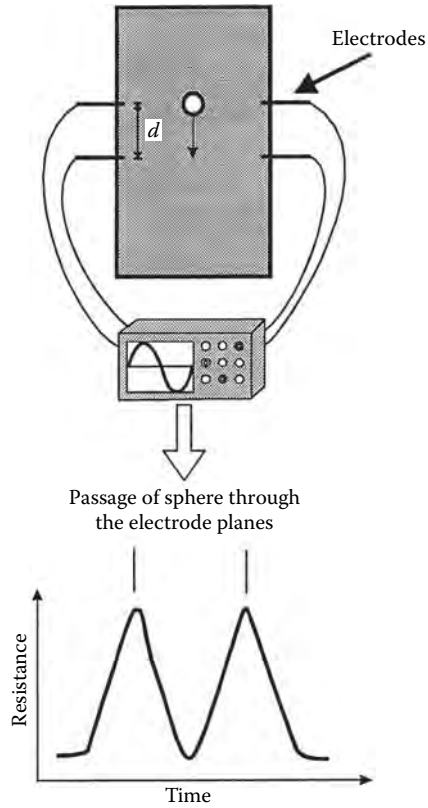


FIGURE 46.9 Diagram of one type of apparatus used to determine the viscosity of opaque liquids in situ. The electrical signal from the passage of the falling sphere indicates the time to traverse a known distance (d) between the two sensors.

can, and do, lead to significant and extremely complex rheological problems that have yet to be fully explained. If a bubble of gas or droplet of liquid with a radius, r_b, and density, ρ', is freely rising in some enclosing viscous liquid of density ρ, then the shear viscosity of the enclosing liquid is determined by

$$\eta = \left(\frac{1}{\varepsilon}\right)\frac{[2gr_b^2(\rho - \rho')]}{9U_t} \tag{46.33}$$

where

$$\varepsilon = \frac{(2\eta + 3\eta')}{3(\eta + \eta')} \tag{46.34}$$

where η' is the viscosity of the bubble. It must be noted that when $\eta' \gg \eta$, ε (as for solid spheres in a fluid), $\varepsilon = 1$, which reduces Equation 46.33 to 46.25. For $\eta' \ll \eta$ (as for gas bubbles in a fluid), ε becomes 2/3, and the viscosity calculated by Equation 46.33 is 1.5 times greater than the viscosity calculated by Equation 46.25.

During the rise, great care must be taken to avoid contamination of the bubble and its surface with impurities in the surrounding liquid. Impurities can diffuse through the surface of the bubble and combine with the fluid inside. Because the bubble has a low viscosity, the upward motion in a viscous medium induces a drag on the bubble that is responsible for generating a circulatory motion within it. This motion can efficiently distribute impurities throughout the whole of the bubble, thereby changing its viscosity and density. Impurities left on the surface of the bubble can form a "skin" that can significantly affect the rise of the bubble, as the skin layer has its own density and viscosity that are not included in Equation 46.33. These surface impurities also make a significant contribution to the inhomogeneous distribution of interfacial tension forces. A balance of these forces is crucial for the formation of a spherical bubble. The simplest method to minimize the earlier effects is to employ minute bubbles by introducing a specific volume of fluid (gas or liquid), with a syringe or other similar device, at the lower end of the cylindrical container. Very small bubbles behave like solid spheres, which make interfacial tension forces and internal fluid motion negligible.

In all rising bubble viscometers, the bubble is assumed to be spherical. Experimental studies of the shapes of freely rising gas bubbles in a container of finite extent [20] have shown that (to 1% accuracy) a bubble will form and retain a spherical shape if the ratio of the radius of the bubble to the radius of the confining cylindrical container is less than 0.2. These studies have also demonstrated that the effect of the wall on the terminal velocity of a rising spherical bubble is to cause a large decrease (up to 39%) in the observed velocity compared to the velocity measured within an unbounded medium. This implies that the walls of the container influence the velocity of the rising bubble more significantly than its geometry. In this method, end effects are known to be large. However, a rigorous, analytically or empirically derived correction factor has not yet appeared. To circumvent this, the ratio of container length to sphere diameter must be in the range of 10–100. As in other Stokesian methods, this allows the bubble's velocity to be measured at locations that experience negligible end effects.

Considering all of the earlier complications, the use of minute bubbles is the best approach to ensure a viscosity measurement that is least affected by both the liquid to be investigated and the container geometry.

46.6 Oscillation Methods

Oscillation methods are based on the measurement of viscous damping of oscillation within a fluid sample. Either the sample itself can be the wave carrying medium (acoustic method) or a separate oscillator can be used in forced oscillation or free decay mode (vibrating wire, transducer, or tuning fork methods). If a liquid is contained within a vessel suspended by some torsional system that is set

in oscillation about its vertical axis, then the motion of the vessel will experience a gradual damping. In an ideal situation, the damping of the motion of the vessel arises purely as a result of the *viscous coupling* of the liquid to the vessel and the viscous coupling between annular layers in the liquid. In any practical situation, there are also frictional losses within the mechanical components of the suspension system that contribute to damping and must be accounted for in the analysis of the measurements. From observations of the amplitudes and time periods of the resulting oscillations, a viscosity of the liquid can be calculated. A schematic diagram of the basic setup of the method is shown in Figure 46.10. Following initial oscillatory excitation, a light source (such as a low-intensity laser) can be used to measure the amplitudes and periods of the resulting oscillations by reflection of the mirror attached to the suspension rod to give an accurate measure of the logarithmic decrement of the oscillations (δ) and the periods (T).

Various working formulae have been derived that associate the oscillatory motion of a vessel of radius r to the absolute viscosity of the liquid. The most reliable formula is the following equation for a cylindrical vessel [21]:

$$\eta = \left[\frac{I\delta}{(\pi r^3 HZ)}\right]^2 \left[\frac{1}{\pi\rho T}\right] \tag{46.35}$$

where

$$Z = \left(\frac{1+r}{4H}\right)a_0 - \frac{[(3/2)+(4r/\pi\mathrm{H})]1}{p} + \frac{[(3/8)+(9r/4\mathrm{H})]a_2}{2p^2} \tag{46.36}$$

$$p = \left(\frac{\pi\rho}{\eta T}\right)^{1/2} r \tag{46.37}$$

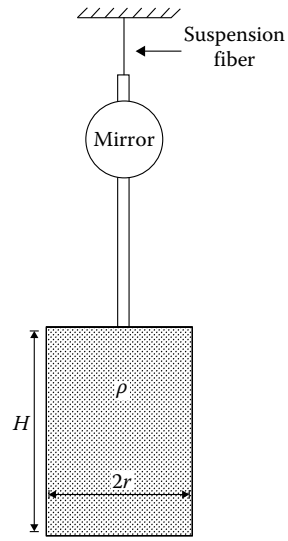


FIGURE 46.10 Schematic diagram of the oscillating cup viscometer. Measurement of the logarithmic damping of the amplitude and period of vessel oscillation is used to determine the absolute shear viscosity of the liquid.

$$a_0 = 1 - \left(\frac{\delta}{4\pi}\right) - \left(\frac{3\delta^2}{32\pi^2}\right) \tag{46.38}$$

$$a_2 = 1 + \left(\frac{\delta}{4\pi}\right) + \left(\frac{\delta^2}{32\pi^2}\right) \tag{46.39}$$

where

I is the mass moment of inertia of the suspended system

ρ is the density of the liquid

A more practical expression of Equation 46.35 is obtained by introducing a number of simplifications. First, it is a reasonable assumption to consider δ to be small (in the order of 10^{-2}–10^{-3}). This reduces a_0 and a_2 to values of 1 and –1, respectively. Second, the effects of friction from the suspension system and the surrounding atmosphere can be experimentally determined and contained within a single parameter or instrument constant, δ_0. This must then be subtracted from the measured δ. A common method of obtaining δ_0 is to observe the logarithmic decrement of the system with an empty sample vessel and subtract that value from the measured value of δ. With these modifications, Equation 46.35 becomes

$$\frac{(\delta - \delta_0)}{\rho} = \left[A\left(\frac{\eta}{\rho}\right)^{1/2} - B\left(\frac{\eta}{\rho}\right) + C\left(\frac{\eta}{\rho}\right)^{3/2} \right] \tag{46.40}$$

where

$$A = \left(\frac{\pi^{3/2}}{I}\right)\left[1 + \left(\frac{r}{4H}\right) Hr^3 T^{1/2}\right] \tag{46.41}$$

$$B = \left(\frac{\pi}{I}\right)\left[\left(\frac{3}{2}\right) + \frac{4r}{\pi H}\right] Hr^2 T \tag{46.42}$$

$$C = \left(\frac{\pi^{1/2}}{2I}\right)\left[\left(\frac{3}{8}\right) + \frac{9r}{4H}\right] HrT^{3/2} \tag{46.43}$$

It has been noted [22] that the analytical form of Equation 46.40 needs an empirically derived, instrument-constant correction factor (ζ) in order to agree with experimentally measured values of η. The discrepancy between the analytical form and the measured value arises as a result of the earlier assumptions. However, these assumptions are required as solutions of the differential equations of motion of this system are nontrivial. The correction factor is dependent on the materials, dimensions, and densities of each individual system but generally lies between the values of 1.0 and 1.08. The correction factor is obtained by comparing viscosity values of calibration materials determined by an individual system (with Equation 46.35) and viscosity values obtained by another reliable method such as the capillary method.

With the earlier considerations taken into account, the final working Roscoe's formula for the absolute shear viscosity is

$$\frac{(\delta - \delta_0)}{\rho} = \zeta\left[A\left(\frac{\eta}{\rho}\right)^{1/2} - B\left(\frac{\eta}{\rho}\right) + C\left(\frac{\eta}{\rho}\right)^{3/2} \right] \tag{46.44}$$

The *oscillating cup method* has been used and is best suited for use with low values of viscosity within the range of 10^{-5}–10^{-2} Pa·s. Its simple closed design and use at high temperatures have made this method very popular for viscosity measurement of liquid metals.

Perhaps the simplest version of the oscillation type of instrument is the vibrating-wire viscometer. Vibrating-wire viscometers are typically quite small and so require small quantities of fluid. They provide accurate measurements but may require calibration using a known fluid. They can be used in quite extreme conditions: the vibrating-wire viscometer was first used to measure the viscosity of liquid helium at temperatures around 2 K [23,24]. Related instruments have been used at pressures in the GPa range [25]. A viscometer based on a vibrating tuning fork, which is capable of accurate measurements over a wide viscosity range (A & D Engineering), is available commercially.

The theory of operation of the vibrating-wire viscometer has been described in detail in the literature [26]. It is based on the phenomenon of resonance. A wire (or more generally a thin beam) fixed at both ends will have a natural vibration frequency ω_0 determined by its length, physical properties, and tension. The vibration amplitude of the wire will be strongly peaked at that resonance frequency, and the height and width of the resonant peak depending on the magnitude of the forces that damp the wire's motion. For a vibrating wire immersed in a fluid, the width of the peak $\Delta\omega$ or equivalently the quality factor $Q = \omega_0/\Delta\omega$ thus depends on viscosity, with higher viscosity leading to a smaller, broader peak and a lower Q. In a typical instrument, the vibrating wire sits in a uniform magnetic field produced by permanent magnets. A small alternating current of angular frequency ω is passed through the wire, and the resulting Lorentz forces cause the wire to vibrate at the same frequency. The voltage across the wire will consist of a constant contribution due to the electrical impedance of the system plus an induced emf that is proportional to the amplitude of the vibration. Vibrating-wire viscometers can operate in a forced mode, in which the wire is driven over a range of frequencies and the resonant peak measured directly, or in a transient mode, in which Q is determined from measurements of the decay rate of the oscillations following a short pulse of the driving current. In some configurations, the tension on the vibrating wire is applied by a mass suspended in the fluid. In this case, the instrument can be used to simultaneously measure fluid density, since the tension and so the resonant frequency will depend on the buoyant force acting on the mass.

Several recent publications describe miniaturized versions of vibrating-object viscometers. Such instruments are of interest in applications ranging from the need for simple bedside measurements of the viscosity of blood, in which the quantity of fluid used must be minimized, to in situ measurement of viscosity in oil field reservoirs, in which the viscometer must withstand high temperatures and pressures. Very small vibrating beam and vibrating plate viscometers have been fabricated using microelectromechanical system (MEMS) technology [27,28]. Smith et al. [29] describe a MEMS oscillating plate viscometer for human blood measurements. Dehestru et al. [30] discuss a vibrating-wire viscometer designed for oil field use. It has an internal volume of a few μL and is accurate to ±10% at pressures up to 24 MPa and temperatures up to 175 °C.

46.7 Acoustic Methods

Viscosity plays an important role in the absorption of energy of an *acoustic wave* traveling through a liquid. By using ultrasonic waves (10^4 Hz < f < 10^8 Hz), the elastic, viscoelastic, and viscous response of a liquid can be measured down to times as short as 10 ns. When the viscosity of the fluid is low, the resulting time scale for structural relaxation is shorter than the ultrasonic wave period, and the fluid is probed in the relaxed state. High-viscosity fluids subjected to ultrasonic wave trains respond as a stiff fluid because structural equilibration due to the acoustic perturbation does not go to completion before the next wave cycle. Consequently, the fluid is said to be in an unrelaxed state that is characterized by dispersion (frequency-dependent wave velocity) and elastic moduli that reflect a much stiffer liquid. The frequency dependence of the viscosity relative to some reference viscosity (η_0) at low frequency, η/η_0, and of the absorption per wavelength, $\alpha\lambda$, where α is the absorption coefficient of the liquid and λ is the wavelength of the compressional

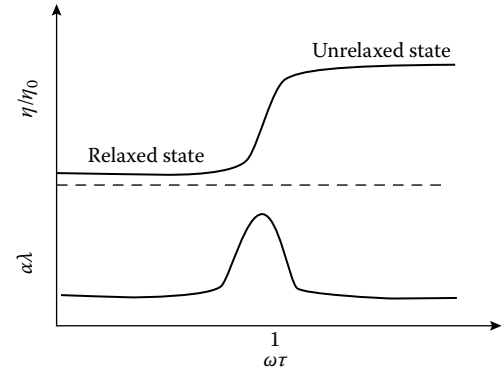


FIGURE 46.11 Effects of liquid relaxation (relaxation frequency corresponds to $\omega\tau = 1$ where $\omega = 2\pi f$) on relative viscosity (upper) and absorption per wavelength (lower) in the relaxed elastic ($\omega\tau < 1$) and unrelaxed viscoelastic ($\omega\tau > 1$) regimes.

wave, for a liquid with a single relaxation time, t, is shown in Figure 46.11. The maximum absorption per wavelength occurs at the *relaxation frequency* when $\omega\tau = 1$ and is accompanied by a step in η/η_0, as well as in other properties such as velocity and compressibility. Depending on the application of the measured properties, it is important to determine if the liquid is in a relaxed or unrelaxed state.

A schematic diagram of a typical apparatus for measuring viscosity by the ultrasonic method is shown in Figure 46.12. Mechanical vibrations in a *piezoelectric* transducer travel down one of the buffer rods (BR-1 in Figure 46.12) and into the liquid sample and are received by a similar transducer mounted on the other buffer rod, BR-2. In the fixed buffer rod configuration, once steady-state conditions have been reached, the applied signal is turned off quickly. The decay rate of the received and amplified signal,

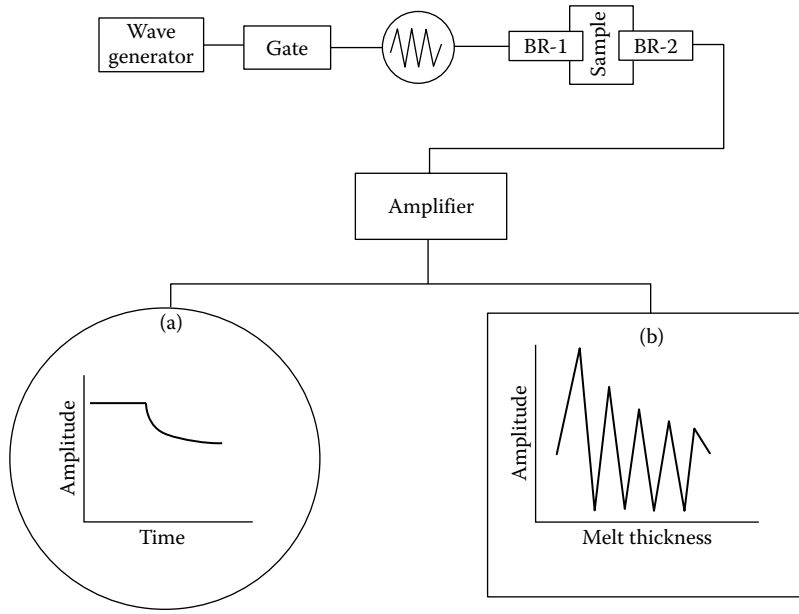


FIGURE 46.12 Schematic diagram of apparatus for liquid shear and volume viscosity determination by ultrasonic wave attenuation measurement showing the received signal amplitude through the exit buffer rod (BR-2) using (a) a fixed buffer rod (BR-2) configuration and (b) an interferometric technique with moveable buffer rod.

displayed on an oscilloscope on an amplitude versus time plot as shown in Figure 46.12a, gives a measure of α. The received amplitude decays as

$$A = A_0 e^{-(b+\alpha c)t'} \tag{46.45}$$

where
A is the received decaying amplitude
A_0 is the input amplitude
b is an apparatus constant that depends on other losses in the system such as due to the transducer and container that can be evaluated by measuring the attenuation in a standard liquid
c is the compressional wave velocity of the liquid
t' is the time

At low frequencies, the absorption coefficient is expressed in terms of volume and shear viscosity:

$$\left(\eta_v + \frac{4\eta}{3}\right) = \frac{\alpha\rho c^3}{2\pi^2 f^2} \tag{46.46}$$

One of the earliest ultrasonic methods of measuring attenuation in liquids is based on *acoustic interferometry* [31]. Apart from the instrumentation needed to move and determine the position of one of the buffer rods accurately, the experimental apparatus is essentially the same as for the fixed buffer rod configuration [32]. The measurement, however, depends on the continuous acoustic wave interference of transmitted and reflected waves within the sample melt as one of the buffer rods is moved away from the other rod. The attenuation is characterized by the decay of the maxima amplitude as a function of melt thickness as shown on the interferogram in Figure 46.12b. Determining α from the observed amplitude decrement involves numerical solution to a system of equations characterizing complex wave propagation [33]. The ideal conditions represented in the theory do not account for conditions such as wave front curvature, buffer rod end nonparallelism, surface roughness, and misalignment. These problems can be addressed in the amplitude fitting stage, but they can be difficult to overcome. The interested reader is referred to [33] for further details.

Ultrasonic methods have not been and are not likely to become the mainstay of fluid viscosity determination simply because they are more technically complicated than conventional viscometry techniques. And although ultrasonic viscometry supplies additional related elastic property data, its niche in viscometry is its capability of providing volume viscosity data. Since there is no other viscometer to measure η_v, ultrasonic absorption measurements play a unique role in the study of volume viscosity.

46.8 Microrheology

On the microscopic scale, viscosity is related to the phenomenon of diffusion. A small (i.e., micron sized) particle suspended in a viscous fluid such as water undergoes random motion as a result of thermal fluctuations in the forces acting on it. This is the well-known Brownian motion. Einstein showed in a famous 1905 [34] paper that the mean squared displacement $\langle r^2 \rangle$ of a particle undergoing Brownian motion is given by

$$\left\langle r^2(\tau) \right\rangle = 2dD\tau, \tag{46.47}$$

where
τ is the time
d is the dimensionality of the motion
The angle brackets indicate an average over an ensemble of particles

D is the diffusion constant, which is related to the viscosity η of the fluid by the Stokes–Einstein equation:

$$D = \frac{k_B T}{6\pi\eta a} \tag{46.48}$$

Here
k_B is the Boltzmann constant
T is the absolute temperature

If the radius a of the diffusing particles is known, these equations can be used to determine the viscosity from measurements of their mean squared displacement. In a viscoelastic fluid, the effects of elasticity make the mean squared displacement grow more slowly than linearly. In this case, Equations 46.47 and 46.48 can be generalized to relate the frequency-dependent viscous and elastic moduli to $\langle r^2 \rangle$. This is discussed in detail in Refs. [35–39].

Several techniques collectively referred to as microrheology have been developed to make use of this relationship to measure viscosity or, more generally, the viscoelastic properties of fluids. The most commonly used of these will be discussed briefly in the following. The development and application of microrheological techniques have been discussed in a number of review papers [35,40]. Microrheology can be used to measure the properties of very small quantities of fluid and so is useful for characterizing expensive or scarce fluids. On the other hand, while at least one commercial instrument is available, microrheology is still largely a research laboratory technique. For quantitative measurements of the bulk viscous and elastic moduli, the fluids being tested must be homogeneous on the scale of the particle motion, since, if the fluid has small-scale structure, the local viscoelastic properties felt by the tracer particles can differ dramatically from the bulk properties. This is not an issue for most Newtonian fluids, but it can be for many complex fluids such as suspensions and gels.

46.8.1 Particle-Tracking Microrheology

Particle-tracking microrheology is based on direct imaging of small tracer particles undergoing Brownian motion in a fluid. This limits its application to fluids that are transparent or at least sufficiently so that the positions of the tracer particles can be detected. The maximum viscosity that can be measured with particle tracking is limited by the requirement that the motion of the tracers must be large enough to be detectable. This limit depends on the optical resolution of the microscope and image analysis procedure, as well as on the size of the tracers. For a typical experimental system, this limit is on the order of tens of Pa · s.

Particle-tracking microrheology is done using a high-quality microscope fitted with a video camera interfaced to a computer. Tracer particles are suspended in the fluid of interest, which is placed in a sample holder and mounted on the stage of the microscope. Sample holders can be fabricated to contain any volume of fluid, but 10–100 μL is typical. The microscope must be focused on the center of the holder to avoid any influence of the walls on the motion of the tracer particles. The video camera records images of the tracer particles at a set frame rate, which are stored on the computer, and image analysis software is later used to reconstruct the particles' trajectories and to calculate the mean squared displacement. To obtain good statistics in the data, the trajectories of many particles are tracked simultaneously, and the concentration of tracers in the fluid is typically chosen to give about 50 particles in the microscope's field of view.

The implementation details are discussed more fully elsewhere [35,40], but several points are worth mentioning. The tracer particles should be close in density to the fluid of interest so that they do not settle over the course of the measurement. Polystyrene latex spheres with a density of 1.05 g cm^{-3} can be obtained in a range of sizes and with a variety of chemical coatings and are typically used for aqueous solutions. It is important to ensure that the tracer particles do not interact chemically or

electrostatically with the fluid being tested. Using bright-field microscopy, particles down to 0.5 μm in diameter can be tracked, while fluorescence microscopy can be used to track fluorescently dyed particles as small as 0.1 μm in size. The shortest time over which the mean squared displacement can be measured is limited by the video frame rate and the longest time by computer memory or the stability of the fluid.

The stored video images of the diffusing particles must be analyzed to extract the trajectory of each particle and ultimately the mean squared displacement of the ensemble of particles. Several software packages can be used to do this analysis. Particle-tracking software by Crocker and Grier [41] that uses the interactive data language (IDL)® programming language (Excelis Inc.) is freely available [42] and has been used extensively. This package has also been translated into MATLAB®, and other image analysis software such as ImageJ could also be used. The image analysis involves filtering the images to remove background intensity variations, identifying candidate particles by thresholding the images, determining their positions to subpixel accuracy from the centroid of their brightness distribution, and finally reconstructing particle trajectories by matching particles from one image to the next. Once the trajectories of all recorded particles have been determined, their squared displacements are calculated and averaged. The viscosity, or the viscous and elastic moduli in the case of a non-Newtonian fluid, can then be calculated. Typically measurement uncertainties limit the observable mean squared displacement to about 10^{-4} μm^2.

An individual tracer particle moving in a viscoelastic fluid interacts hydrodynamically with other tracer particles. Two-particle microrheology exploits this fact by using correlations between the motions of pairs of particles to obtain viscoelastic properties of the fluid on the scale of the separation between the particles [43]. Two-particle microrheology gives moduli that are independent of the interactions between the tracer particles and the material under study [43] and so gives the bulk properties even in microscopically inhomogeneous fluids. Two-particle microrheology requires the same equipment as described earlier. The difference comes in the data analysis, which, since it involves analyzing the motion of pairs of particles, requires much more data than in single-particle measurements. This technique has been used to study length-scale-dependent rheology in inhomogeneous materials such as actin networks [43], but for now it remains a research tool rather than a practical technique for routine measurements of material properties.

46.8.2 Dynamic Light Scattering

A second microrheological technique uses dynamic light scattering to measure the mean squared displacement of tracer particles in the fluid. In this technique, incident laser light is scattered by particles suspended in the fluid of interest. A photomultiplier mounted at an angle θ with respect to the incident beam measures the scattered light intensity, which fluctuates in time due to the random motion of the scattering particles. A digital correlator calculates the autocorrelation function of the scattered intensity, $g_2(\tau)$, where τ is the time and g_2 is related to g_1, the autocorrelation function of the scattered electric field, which decays in time as

$$g_1(\tau) = e^{-(q^2/6)\langle r^2(\tau)\rangle}, \tag{46.49}$$

where $q = (4\pi n/\lambda)\sin(\theta/2)$, n is the refractive index of the fluid, and λ is the wavelength of the laser light. The decay rate of $g_1(\tau)$ thus gives the mean squared displacement of the tracer particles, from which the viscosity or viscous and elastic moduli can be calculated using Equations 46.47 and 46.48, as shown earlier. An example of data obtained using this technique is shown in Figure 46.13, where the mean squared displacement is plotted as a function of time for 110 nm polystyrene spheres in water.

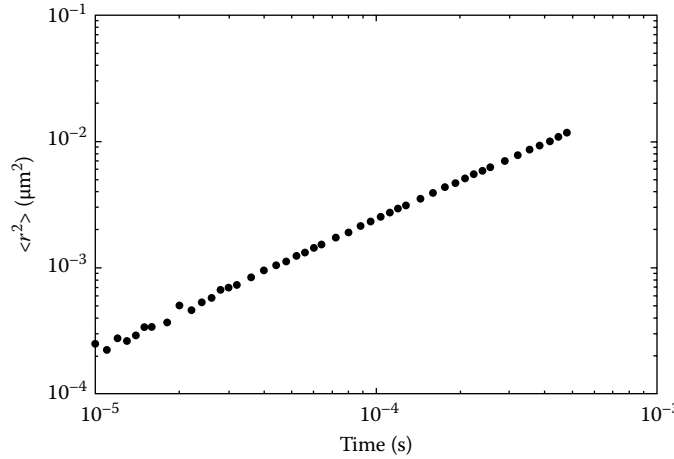


FIGURE 46.13 Mean squared displacement of 100 nm polystyrene sphere diffusion in water, measured using dynamic light scattering. These data can be used to obtain the diffusion constant, from which the viscosity of the fluid can be determined. (From Yang, N. and deBruyn, J.R., unpublished.)

A range of light scattering instruments, including laser, photodetector, and correlator, are available commercially, although many "black-box" scattering systems are intended for particle size determination assuming a known viscosity. The tracer particles must again be neutrally buoyant in the fluid under study, and their concentration must be low enough that the laser light is scattered at most once as it passes through the sample holder. Typically a few milliliters of fluid is required, and the sample must again be transparent. Light scattering microrheology measures the viscoelastic properties of fluids at much higher frequencies than particle tracking, which is useful for characterizing complex fluids such as polymer solutions.

A related microrheological technique uses diffusing-wave spectroscopy. This is also a light scattering technique but requires the light to be scattered many times in the sample, so that the light effectively diffuses through it. This requires a higher concentration of tracer particles. As shown earlier, the decay rate of the electric field autocorrelation function is related to the mean squared displacement of the tracer particles, although the equations involved are slightly more complicated. Details are given in a number of references. Diffusing-wave spectroscopy allows measurements at even higher frequencies than the single-scattering technique. An instrument that does microrheological measurements based on this technique is available commercially.

46.9 High-Pressure Rheology

Measurement of viscosity of fluids at high pressures began with Bridgman [45]. His primary interest was to investigate liquids that would remain fluid at high pressures so they could be used as hydrostatic pressure-transmitting media in his experiments. Since that time, continued development of high-pressure viscometry methods remained in the scientific research realm until recently. Over the past decade, there has been increasing commercial and industrial interest in the application of high-pressure technology in the processing of fluids. A wide range of foods, for example, is now being processed using high-pressure technology principally as a nonthermal alternative for extending shelf life. The technology also provides the food industry with a variety of new product development opportunities able to exploit the functional properties of ingredients such as hydrocolloids and proteins. High-pressure technology is proving increasingly attractive and beneficial in the destruction of microorganisms; activation and deactivation of enzymes; inactivation kinetics of both vegetative

and pathogenic microorganisms; change of functional properties of biopolymers such as proteins and polysaccharides used in foams, gels, and emulsions; and the control of phase change such as fat solidification and ice melting point. The rheology of these fluids and soft solids is important to control process function and speed. In other applications such as in the automotive industry, understanding the high-pressure rheology of petroleum-based fuels is critical for the safe and efficient design and performance of engines. High-pressure rheology has also been recognized as vital information for studies of planetary interiors, where fluids of mainly silicate or metallic composition play significant roles in heat and mass flow and therefore fluid rheology is a pivotal parameter in controlling planetary evolution.

46.9.1 High-Pressure Rheometry

Rheometry methods employed at high pressure are mainly based on traditional methods and can be grouped into four categories: concentric cylinder, capillary, oscillatory, and falling/rolling body. The theoretical treatment given earlier for each of these methods applies, but additional corrections may be required for the container geometry effects due to the limited sample volumes imposed by the high-pressure environment. The approaches to measure motion or stress in situ are often applied outside the high-pressure environment because of the high loads required and can involve detailed experimental setups. Also, depending on the pressure range, the equipment required to generate high pressure can be much more elaborate than the equipment required to measure viscosity. Although there are several high-pressure viscometers commercially available for use below a pressure of 1 GPa and examples are listed in Table 46.5, measurement of viscosity at pressures above 1 GPa is carried out mainly in the research laboratory. There is no implied endorsement of any manufacturer listed in Table 46.5.

Recent reviews of high-pressure viscosity measurement have appeared [46,47] and describe in detail the pressure devices and methods used up to ~1 GPa. This appears to be the pressure limit of the concentric cylinder method because of the engineering challenges presented by the friction effects in rotating high-pressure seals and the difficulties in precise measurement of torque when parts are subjected to high mechanical loads. Most high-pressure rotating systems use the Searle principle with a rotating inner cylinder and stationary outer cylinder. Methods of measuring the resulting shear stress include elastic torsion tubes, magnetic or inductive coupling, strain gauge load cell, or variable torque induction motor. The capillary method is based on pressure-driven flow. The pressure drop along the capillary is normally measured with pressure transducers, which must detect small pressure differences along the capillary on top of a large pressure signal arising from the static high-pressure environment. High-pressure oscillation methods are based on the measurement of viscous damping of an oscillator located within a pressurized fluid sample. Piezoelectric quartz crystals, which can operate at pressures exceeding 1 GPa, show a shift in the

TABLE 46.5 High-Pressure Viscometer/Rheometer Manufacturers

Manufacturers	Maximum Pressure (Bars)	Method
PSL Systemtechnik	1200	Rotational
Cambridge Viscosity, Inc.	1379	Oscillating piston/EM
PAC	2040	Oscillating piston/EM
Chandler Engineering	2721	Rotational
Stony Brook Scientific, Inc.	4082	Falling needle/magnetic
F5 Technologie	5000	Torsional oscillation

Note: All the above-mentioned manufacturers/distributors can be found on the Internet, along with their most recent contact information, product description, and, in some cases, pricing. There is no implied endorsement of any manufacturer listed in this table.

crystal resonant frequency reflecting the mechanical shear impedance of the sample. A second way to measure fluid viscosity is by observation of the free decay of a disk or wire-shaped oscillator. A third way to measure viscosity using the oscillation method at high pressure is by passage of transducer-generated and transducer-detected ultrasonic waves through the pressurized fluid. The viscosity is calculated via the wave absorption coefficient of the sample, which is measured by signal attenuation at the receiving transducer.

At pressures above 1 GPa, the falling/rolling body method is the preferred method although at least one study of a capillary method employed to 6 GPa has been reported [48]. The simplicity and reliability of construction of components for the Stokesian method make it capable of providing viscosity measurements at the highest pressures, where the sample fluid volume ranges from $\sim 10^{-2}$ mm^3 to a few mm^3. In principle, viscosity can be measured with this method at pressures up to several 10's of GPa using the diamond anvil cell high-pressure device. At such high pressures, high temperature is also required in order to melt the pressure-frozen liquid or because this range of pressures is mainly of interest to the Earth and planetary sciences and the relevant materials to be studied have melting temperatures in excess of 1500 °C. Such high temperatures require judicious selection of the material for the falling/rolling body to ensure solid-phase stability of the body at the test conditions, chemical inertness of the body in the fluid, appropriate density contrast of the body and fluid, and ease of fabrication of the desired geometry. The velocimetry method used depends on the body and fluid and on the observational access to the pressurized sample region. Since the diamond anvil cell has the great advantage of optical access through the diamond anvils, this is the method of choice for velocimetry of the moving body between two fiducial points. In large-volume presses, which contain optically opaque components (usually ceramics and metals) in the pressure cell, there are two main approaches to measuring the speed of the moving body. The first involves allowing the body to move for a measured time in the pressurized fluid and then quenching the temperature to freeze the fluid, thereby capturing the body in position. Follow-up measurement of the body position is made by sectioning the recovered sample and measuring the distance the body has moved. This is an effective method for measuring viscosities in the range of approximately 10^{-1}–10^5 Pa s if the density contrast of the body and fluid is chosen appropriately. A technique for tailoring the density of a probe sphere, by cladding a metallic core sphere with a refractory mantle to make a composite sphere [49], allows for adjustment of the fall time of the body to within acceptable ranges considering the fluid density and its expected viscosity. However, for measuring much lower viscosities, the control on quenching the temperature and accurate determination of the body start/stop position are inadequate, and therefore, in situ velocimetry methods are required. Many in situ velocimetry methods have been devised based on the following with either in situ or ex situ components for detection: electromagnetic induction using a two-coil external arrangement and an electrically conducting body in a nonconducting fluid, detecting passage of an insulating body moving through two electrodes placed in a metallic fluid by measuring electrical resistance of the fluid, detecting passage of a weakly radioactive body such as a ^{60}Co sphere sensed by an external scintillation counter, and x-ray radiographic imaging of a body with sufficient absorption x-ray contrast with the fluid. Over the last decade, the last method has become the method of choice for accurate high-pressure viscometry at pressures above 1 GPa [50] on low-viscosity fluids and is described in further detail.

Experiments to measure the viscosity of a fluid at high temperatures and at pressures above 1 GPa are frequently carried out using the x-ray radiography imaging technique in real time at a synchrotron facility. The high photon flux at a synchrotron is needed in order to produce images of sufficient quality and at high enough frequency (or ideally a movie) at various positions during travel of the body in order to derive a velocity. The x-ray radiographic technique images the pressurized region according to the relative x-ray attenuation coefficients of the experimental materials. A diagram of the experimental imaging setup for a high-pressure Stokesian viscometry measurement using synchrotron x-ray radiography, the resultant sequence of time lapse images, and the sphere displacement and velocity data for a falling composite sphere is shown in Figure 46.14.

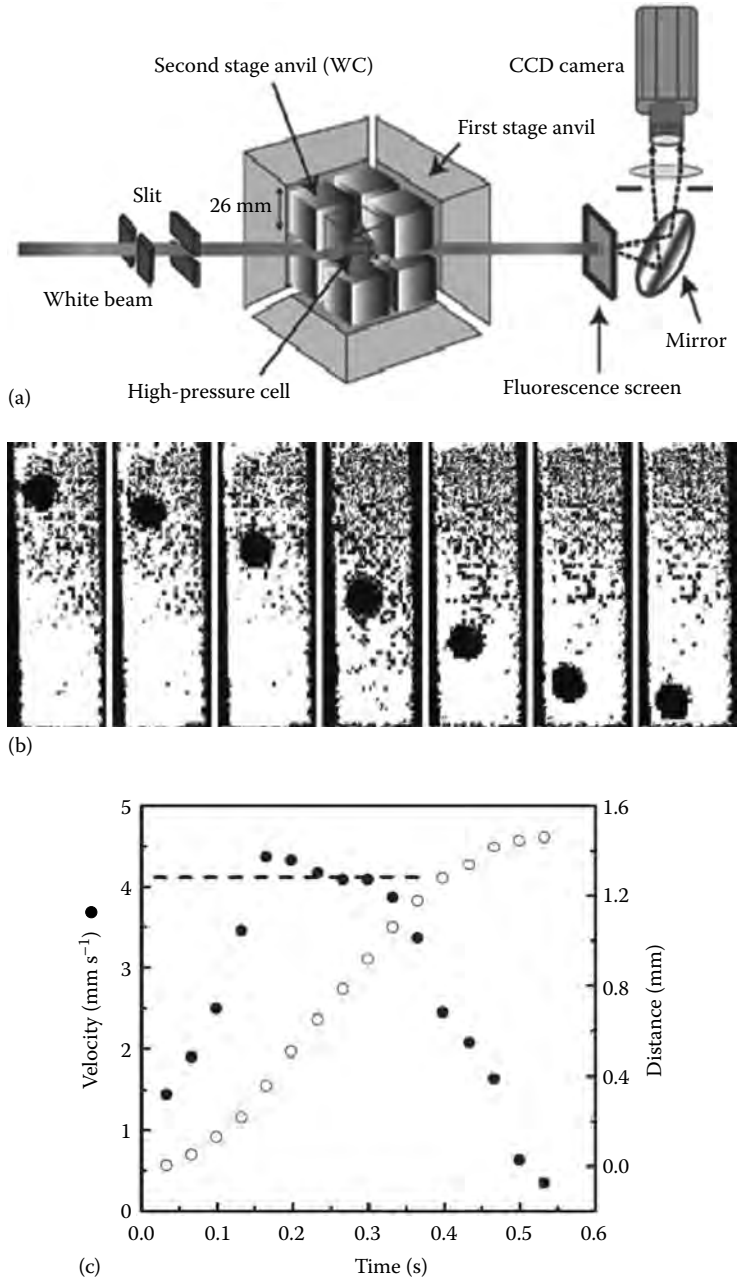


FIGURE 46.14 (a) Synchrotron x-ray radiography setup for a high-pressure Stokesian viscometry measurement [51]. (b) Synchrotron radiographic images of a Pt core—ruby mantle composite sphere falling in liquid Fe–29 wt% S at 2.6 GPa and 1373 K. Images are separated in time by 62 ms [52]. (c) Distance and velocity of falling composite sphere versus time in Fe–29 wt% S liquid, at 2.6 GPa and 1373 K. Dashed line is terminal velocity [52].

References

1. T. A. Litovitz and C. M. Davis, Structural and shear relaxation in liquids: In: W. P. Mason (Ed.), *Physical Acoustics: Principles and Methods*, Vol. II, Part A, *Properties of Gases, Liquids and Solutions*, New York: Academic Press, 1965, pp. 281–349.
2. Y. Bottinga and P. Richet, Silicate melts: The "anomalous" pressure dependence of the viscosity, *Geochim. Cosmochim. Acta*, 59, 2725–2731, 1995.
3. J. Ferguson and Z. Kemblowski, *Applied Fluid Rheology*, New York: Elsevier, 1991.
4. R. W. Whorlow, *Rheological Techniques*, 2nd edn., New York: Ellis Horwood, 1992.
5. K. Walters, *Rheometry*, London: Chapman and Hall, 1975.
6. J. M. Dealy, *Rheometers for Molten Plastics*, New York: Van Nostrand Reinhold, 1982.
7. J. R. Van Wazer, J. W. Lyons, K. Y. Kim, and R. E. Colwell, *Viscosity and Flow Measurement*, New York: Interscience, 1963.
8. C. W. Macosko, *Rheology: Principles, Measurements, and Applications*, New York: VCH, 1994.
9. W. A. Wakeham, A. Nagashima, and J. V. Sengers (Eds.), *Measurement of the Transport Properties of Fluids*, Oxford, U.K.: Blackwell Scientific, 1991.
10. J. A. Himenez and M. Kostic, A novel computerized viscometer/rheometer, *Rev. Sci. Instrum*., 65(1), 229–241, 1994.
11. DIN 53018 (Parts 1 and 2), 53019, German National Standards.
12. Marks' *Standard Handbooks for Mechanical Engineers*, New York: McGraw-Hill, 1978.
13. ASTM D445-10, Standard Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (and Calculation of Dynamic Viscosity), available at http://enterprise.astm.org/filtrexx40.cgi?HISTORICAL/D445-10.htm (accessed on March 15, 2013).
14. W. D. Kingery, *Viscosity in Property Measurements at High Temperatures*, New York: John Wiley & Sons, 1959.
15. H. Faxen, Die Bewegung einer Starren Kugel Langs der Achsee eines mit Zaher Flussigkeit Gefullten Rohres: Arkiv for Matematik, *Astronomi och Fysik*, 27(17), 1–28, 1923.
16. F. Gui and T. F. Irvine, Jr., Theoretical and experimental study of the falling cylinder viscometer, *Int. J. Heat Mass Transfer*, 37(1), 41–50, 1994.
17. N. A. Park and T. F. Irvine, Jr., Falling cylinder viscometer end correction factor, *Rev. Sci. Instrum*., 66(7), 3982–3984, 1995.
18. G. E. LeBlanc and R. A. Secco, High pressure stokes' viscometry: A new *in-situ* technique for sphere velocity determination, *Rev. Sci. Instrum*., 66(10), 5015–5018, 1995.
19. R. Clift, J. R. Grace, and M. E. Weber, *Bubbles, Drops, and Particles*, San Diego: Academic Press, 1978.
20. M. Coutanceau and P. Thizon, Wall effect on the bubble behavior in highly viscous liquids, *J. Fluid Mech*., 107, 339–373, 1981.
21. R. Roscoe, Viscosity determination by the oscillating vessel method I: Theoretical considerations, *Proc. Phys. Soc*., 72, 576–584, 1958.
22. T. Iida and R. I. L. Guthrie, *The Physical Properties of Liquid Metals*, Oxford, UK: Clarendon Press, 1988.
23. J. T. Tough, W. D. McCormick, and J. G. Dash, Viscosity of liquid He II, *Phys. Rev*., 132, 2373–2378, 1963.
24. J. T. Tough, W. D. McCormick, and J. G. Dash, Vibrating wire viscometer, *Rev. Sci. Instrum*., 35, 1345–1348, 1964.
25. P. S. vander Gulik, R. Mostert, and H. R. vanden Berg, The viscosity of methane at 273 K up to 1 GPa, *Fluid Phase Equil*., 79, 301–311, 1992.
26. T. Retsina, S. M. Richardson, and W. A. Wakeham, The theory of a vibrating-rod viscometer *Appl. Sci. Res*., 43, 325–346, 1987.
27. C. Riesch, E. K. Reichel, A. Jachimowicz, J. Schalko, P. Hudek, B. Jakoby, and F. Keplinger, A suspended plate viscosity sensor featuring in-plane vibration and piezoresistive readout, *J. Micromech. Microeng*., 19, 075010 1–10, 2009.

28. Y. Zhao, S. Li, A. Davidson, B. Yang, Q. Wang, and Q. Lin, A MEMS viscometric sensor for continuous glucose monitoring, *J. Micromech. Microeng*., 17, 2528–2537, 2007.
29. P. D. Smith, R. C. D. Young, and C. R. Chatwin, A MEMS viscometer for unadulterated human blood, *Measurement*, 43, 144–151, 2010.
30. G. Dehestru, M. Leman, J. Jundt, P. Dryden, M. Sullivan, and C. Harrison, A microfluidic vibrating wire viscometer for operation at high pressure and high temperature, *Rev. Sci. Instrum*., 82, 035113 1–5, 2011.
31. H. J. McSkimin, Ultrasonic methods for measuring the mechanical properties of liquids and solids, In: W. P. Mason (Ed.), *Physical Acoustics: Principles and Methods*, Vol. I, Part A, *Properties of Gases, Liquids and Solutions*, New York: Academic Press, 1964, pp. 271–334.
32. P. Nasch, M. H. Manghnani, and R. A. Secco, A modified ultrasonic interferometer for sound velocity measurements in molten metals and alloys, *Rev. Sci. Instrum*., 65, 682–688, 1994.
33. K. W. Katahara, C. S. Rai, M. H. Manghnani, and J. Balogh, An interferometric technique for measuring velocity and attenuation in molten rocks, *J. Geophys. Res*., 86, 11779–11786, 1981.
34. A. Einstein, Ueberdievondermolekularkinetis chentheoriederwaermegefordertebewegung voninruhendenfluessigkeitensuspendiertenteilchen, *Ann. Phys*., 17, 549–560, 1905.
35. M. L. Gardel, M. T. Valentine, and D. A. Weitz, Microrheology, In: K. Breuer (Ed.), *Microscale Diagnostic Techniques*, New York: Springer, 2005.
36. T. G. Mason and D. A. Weitz, Optical measurements of frequency-dependent linear viscoelastic moduli of complex fluids, *Phys. Rev. Lett*., 74, 1250–1253, 1995.
37. T. G. Mason, Estimating the viscoelastic moduli of complex fluids using the generalized Stokes–Einstein equation, *Rheol. Acta*., 39, 371–378, 2000.
38. A. J. Levine and T. C. Lubensky, One- and two-particle microrheology, *Phys. Rev. Lett*., 85, 1774–1777, 2000.
39. A. J. Levine and T. C. Lubensky, Response function of a sphere in a viscoelastic two-fluid medium, *Phys. Rev. E*, 63, 041510 1–12, 2001.
40. T. A. Waigh, Microrheology of complex fluids, *Rep. Prog. Phys*., 68, 685–742, 2005.
41. J. Crocker and D. G. Grier, Methods of digital video microscopy for colloidal studies, *J. Colloid Interface Sci*., 179, 298–310, 1996.
42. J. Crocker and E. Weeks, Particle tracking using IDL, 1996, http://www.physics.emory.edu/~weeks/idl/
43. J. C. Crocker, M. T. Valentine, E. R. Weeks, T. Gisler, P. D. Kaplan, A. G. Yodh, and D. A. Weitz, Two-point microrheology of inhomogeneous soft materials, *Phys. Rev. Lett*., 85, 888–891, 2000.
44. N. Yang and J. R. deBruyn, unpublished.
45. P. W. Bridgman, The viscosity of liquids under pressure, *Proc. Natl. Acad. Sci*., 11, 603–606, 1925.
46. L. Kulisiewicz and A. Delgado, High pressure rheological measurement methods: A review, *Appl. Rheol*., 20, 13018, 2010.
47. C. J. Schaschke, High pressure viscosity measurement with falling body type viscometers, *Int. Rev. Chem. Eng*., 2, 564–576, 2010.
48. J. D. Barnett and C. D. Bosco, Viscosity measurements on liquids to pressures of 60 kbar, *J. Appl. Phys*., 40, 3144–3150, 1969.
49. R. A. Secco, R. F. Tucker, S. P. Balog, and M. D. Rutter, Tailoring sphere density for high pressure physical property measurements on liquids, *Rev. Sci. Instrum*., 71, 2114–2116, 2001.
50. T. Uchida, Y. Wang, M. L. Rivers, S. R. Sutton, D. J. Weidner, M. T. Vaughan, J. Chen, B. Li, R. A. Secco, M. D. Rutter, and H. Liu, A large-volume press facility at the Advanced Photon Source: Diffraction and imaging studies on materials relevant to the cores of planetary bodies, *J. Phys. Condens. Matter*., 14, 11517–11523, 2002.
51. K. Funakoshi, A. Suzuki, and H. Terasaki, In situ viscosity measurements of albite melt under high pressure, *J. Phys. Condens. Matter*., 14, 11343–11347, 2002.
52. R. A. Secco, M. D. Rutter, S. P. Balog, H. Liu, D. C. Rubie, T. Uchida, D. Frost, Y. Wang, M. Rivers, and S. R. Sutton, Viscosity and density of Fe–S liquids at high pressures, *J. Phys. Condens. Matter*., 14, 11325–11330, 2002.

Further Information

Brooks, R. F., Dinsdale, A. T., and Quested, P. N., The measurement of viscosity of alloys—A review of methods, data and models, *Meas. Sci. Technol.*, 16, 354–362, 2005, a review of methods used to measure viscosity of liquid metals including methods adaptable to "in plant" measurement of fluid flow in metallurgical industrial processes.

Hou, Y. Y. and Kassim, H. O., Instrument techniques for rheometry, *Rev. Sci. Instrum.*, 76. 101101, 2005, a review of some of the latest advances in rheology measuring techniques including new approaches to conventional velocimetry (e.g., nuclear magnetic resonance, ultrasound pulse Doppler mapping) as well as other techniques that are important for industrial processes (e.g., extensional rheometry and interfacial rheometry).

O'Neill, M. E. and Chorlton, F., *Viscous and Compressible Fluid Dynamics*, Chichester, U.K.: Ellis Horwood, 1989, mathematical methods and techniques and theoretical description of flows of Newtonian incompressible and ideal compressible fluids.

Ryan, M. P. and Blevins, J. Y. K., The viscosity of synthetic and natural silicate melts and glasses at high temperatures and 1 bar (10^5 Pascals) pressure and at higher pressures, *U.S. Geological Survey Bulletin* 1764, Denver, CO, 1987, p. 563, an extensive compilation of viscosity data in tabular and graphic format and the main techniques used to measure shear viscosity.

Van Wazer, J. R., Lyons, J. W., Kim, K. Y., and Colwell, R. E., *Viscosity and Flow Measurement: A Laboratory Handbook of Rheology*, New York: Interscience Publishers, Division of John Wiley & Sons, 1963, A comprehensive overview of viscometer types and simple laboratory measurements of viscosity for liquids.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

47

Surface Tension Measurement

47.1 Mechanics of Fluid Surfaces **47**-2
47.2 Standard Methods and Instrumentation **47**-5
Capillary Rise Method • Wilhelmy Plate and du Noüy Ring Methods • Maximum Bubble Pressure Method • Pendant Drop and Sessile Drop Methods • Drop Weight or Volume Method • Spinning Drop Method
47.3 Specialized Methods **47**-11
Dynamic Surface Tension • Surface Viscoelasticity • Measurements at Extremes of Temperature and Pressure • Interfacial Tension
Defining Terms **47**-12
Acknowledgments **47**-12
References **47**-13
Further Readings **47**-14

David B. Thiessen
California Institute of Technology

Kin F. Man
California Institute of Technology

The effect of surface tension is observed in many everyday situations. For example, a slowly leaking faucet drips because the force of surface tension allows the water to cling to it until a sufficient mass of water is accumulated to break free. Surface tension can cause a steel needle to "float" on the surface of water although its density is much higher than that of water. The surface of a liquid can be thought of as having a skin that is under tension. A liquid droplet is somewhat analogous to a balloon filled with air. The elastic skin of the balloon contains the air inside at a slightly higher pressure than the surrounding air. The surface of a liquid droplet likewise contains the liquid in the droplet at a pressure that is slightly higher than ambient. A clean liquid surface, however, is not elastic like a rubber skin. The tension in a piece of rubber increases as it is stretched and will eventually rupture. A clean liquid surface can be expanded indefinitely without changing the surface tension.

The mechanical model of the liquid surface is that of a skin under tension. Any given patch of the surface thus experiences an outward force tangential to the surface everywhere on the perimeter. The force per unit length of the perimeter acting perpendicular to the perimeter is defined as the *surface tension*, γ. Molecules in the interfacial region have a higher potential energy than molecules in the bulk phases because of an imbalance of intermolecular attractive forces. This will result in an excess free energy per unit area associated with the surface that is numerically equivalent to the surface tension, as shown later. Consider a flat rectangular patch of fluid interface of width W and length L. In order to expand the length to $L + \Delta L$, an amount of work $\gamma W \Delta L$ must be done at the boundary. The product $W\Delta L$ is just the change in area ΔA of the surface. The work done to increase the area is thus $\Delta A\gamma$, which corresponds to the increase in surface free energy. Thus, the surface tension γ is seen to be equivalent to the surface free energy per unit area. Room-temperature organic liquids typically have surface tensions in the range of 20–40 mN m^{-1},

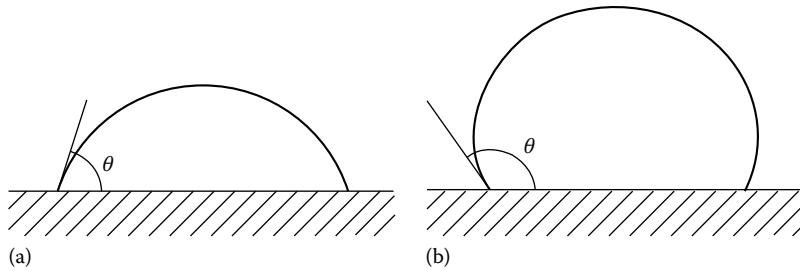


FIGURE 47.1 Illustration of contact angles and wetting. The liquid in (a) wets the solid better than that in (b).

while pure water has a value of 72 mN m^{-1} at 25 °C. The interface between two immiscible liquids, such as oil and water, also has a tension associated with it, which is generally referred to as the *interfacial tension*.

The surface energy concept is useful for understanding the shapes adopted by liquid surfaces. An isolated system in equilibrium is in a state of minimum free energy. Because the surface free energy contributes to the total free energy of a multiphase system, the surface free energy is minimized at equilibrium subject to certain constraints. Also, because the surface free energy is directly proportional to the surface area, surface area is also minimized. In the absence of gravity, a free-floating liquid droplet assumes a spherical shape because, for a given volume of liquid, the sphere has the least surface area. However, a droplet suspended from a needle tip on Earth does not form a perfect sphere because the minimum free energy configuration involves a trade-off between a reduction of the surface energy and a reduction of the gravitational potential. The droplet elongates to reduce its gravitational potential energy.

The surface energy concept is also useful for understanding the behavior of so-called surface active agents or *surfactants*. A two-component liquid mixture in thermodynamic equilibrium exhibits preferential adsorption of one component at the surface if the adsorption causes a decrease in the surface energy. The term surfactant is reserved for molecular species that strongly adsorb at the surface even when their concentration in the bulk liquid is very low. Surfactants are common in natural waters and are very important in many biological and industrial processes.

The interface between a solid and a fluid also has a surface free energy associated with it. Figure 47.1a shows a liquid droplet at rest on a solid surface surrounded by air. This system contains three different types of interfaces: solid–gas, solid–liquid, and liquid–gas, each with a characteristic surface free energy per unit area. The state of minimum free energy for the system then involves trade-offs in the surface area for the various interfaces. The region of contact between the gas, liquid, and solid is termed the *contact line*. The liquid–gas surface meets the solid surface with an angle θ measured through the liquid, which is known as the *contact angle*. The contact angle attains a value that minimizes the free energy of the system and is thus a characteristic of a particular solid–liquid–gas system. The system shown in Figure 47.1a has a smaller contact angle than that shown in Figure 47.1b. The smaller the contact angle, the better the liquid is said to wet the solid surface. For $\theta = 0$, the liquid is said to be perfectly wetting.

Measurement of surface tension is important in many fields of science and engineering, as well as in medicine. A number of standard methods exist for its measurement. In many systems of interest, the surface tension changes with time, perhaps, for example, because of adsorption of surfactants. Several standard methods can be used to measure dynamic surface tension if it changes slowly with time. Special techniques have been developed to measure dynamic surface tensions for systems that evolve very rapidly.

47.1 Mechanics of Fluid Surfaces

Some methods of measuring surface tension depend on the mechanics at the line of contact between a solid, liquid, and gas. When the system is in static mechanical equilibrium, the contact line is motionless, meaning that the net force on the line is zero. Forces acting on the contact line

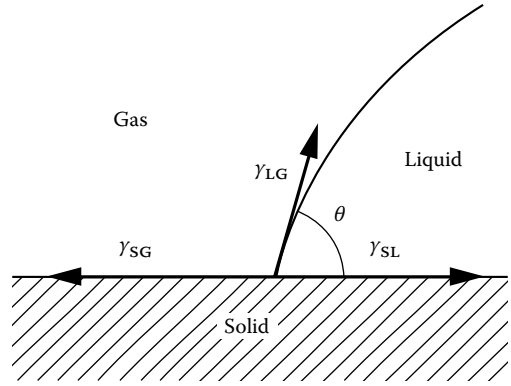


FIGURE 47.2 Surface tension forces acting on the contact line.

arise from the surface tensions of the converging solid–gas, solid–liquid, and liquid–gas interfaces, denoted by γ_{SG}, γ_{SL}, and γ_{LG}, respectively (Figure 47.2). The condition of zero net force along the direction tangent to the solid surface gives the following relationship between the surface tensions and contact angle θ:

$$\gamma_{SG} = \gamma_{SL} + \gamma_{LG}\cos\theta \tag{47.1}$$

This is known as *Young's equation*. The contact angle is thus seen to be dependent on the surface tensions between the various phases present in the system and is therefore an intrinsic property of the system.

As discussed in the introduction, the surface tension of a droplet causes an increase in pressure in the droplet. This can be understood by considering the forces acting on a curved section of surface as illustrated in Figure 47.3a. Because of the curvature, the surface tension forces pull the surface toward the concave side of the surface. For mechanical equilibrium, the pressure must then be greater on the concave side of the surface. Figure 47.3b shows a saddle-shaped section of surface in which surface tension forces oppose each other, thus reducing or eliminating the required pressure difference across the surface. The mean curvature of a 2D surface is specified in terms of the two principal radii of curvature, R_1 and R_2, which are measured in perpendicular directions. A detailed mechanical analysis of curved

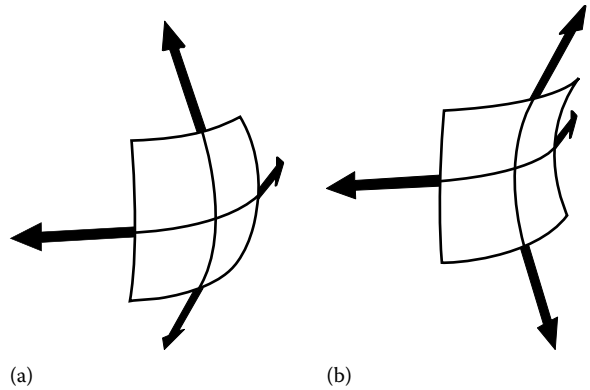


FIGURE 47.3 Mechanics of curved surfaces that have principal radii of curvature of (a) the same sign and (b) the opposite sign.

tensile surfaces shows that the pressure change across the surface is directly proportional to the surface tension and to the mean curvature of the surface:

$$P_A - P_B = \gamma\left(\frac{1}{R_1} + \frac{1}{R_2}\right) \tag{47.2}$$

where γ is the surface tension and the quantity in brackets is twice the mean curvature. The sign of the radius of curvature is positive if its center of curvature lies in phase A and negative if it lies in phase B. Equation 47.2 is known as the *Young–Laplace equation*, and the pressure change across the interface is termed the *Laplace pressure*. Measurement of the Laplace pressure for a surface of known curvature then allows a determination of the surface tension.

Several methods of surface tension measurement are based on the measurement of the static shape of an axisymmetric drop or bubble or on the point of mechanical instability of such drops or bubbles. In a gravitational field, a drop or bubble that is attached to a solid support assumes a nonspherical shape. Figure 47.4a shows the shape of a hanging droplet, also known as a pendant drop, and Figure 47.4b shows a so-called sessile drop. Axisymmetric air bubbles in water attain the same shapes as water drops in air, except that they are inverted. A bubble supported from below is thus called a hanging or pendant bubble, and a bubble supported from above is called a captive or sessile bubble. The reason for the deviation of the shape from that of a sphere can be understood from Equation 47.2. The hydrostatic pressure changes with depth more rapidly in a liquid than in a gas. The pressure difference across the surface of a pendant drop in air therefore increases from top to bottom, requiring an increase in the mean curvature of the surface according to Equation 47.2. The drop in Figure 47.4a has a neck at the top, which means that the two principal radii of curvature have opposite signs and cancel to some extent. At the bottom of the drop, the radii of curvature have the same sign, thus making the mean curvature larger. The Young–Laplace equation can be

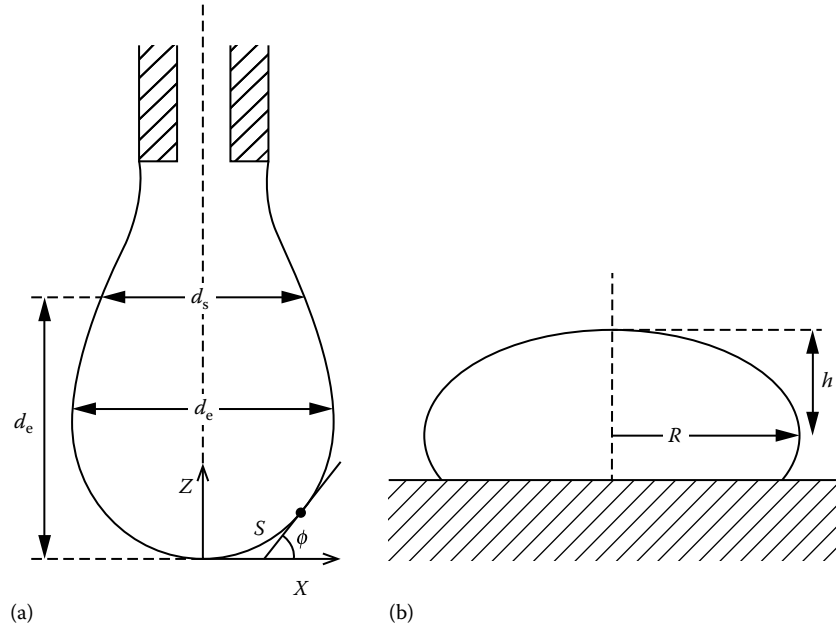


FIGURE 47.4 (a) A pendant drop showing the characteristic dimensions, d_e and d_s, and the coordinates used in the Young–Laplace equation. (b) A sessile drop showing the characteristic dimensions R and h.

written as coupled first-order differential equations in terms of the coordinates of the interface for an axisymmetric surface in a gravitational field as

$$\frac{dx}{ds} = \cos\phi$$
$$\frac{dz}{ds} = \sin\phi \tag{47.3}$$
$$\frac{d\phi}{ds} = \frac{2}{b} + \left(\frac{\Delta\rho g}{\gamma}\right) z - \frac{\sin\phi}{x}$$
$$x(0) = z(0) = \phi(0) = 0$$

where

x and z are the horizontal and vertical coordinates, respectively, with the origin at the drop apex
s is the arc length along the drop surface measured from the drop apex
ϕ is the angle between the surface tangent and the horizontal (Figure 47.4a)
The parameter b is the radius of curvature at the apex of the drop or bubble
$\Delta\rho$ is the density difference between the two fluid phases
g is the acceleration of gravity

Numerical integration of Equation 47.3 allows one to compute the shape of an axisymmetric fluid interface. Comparison of computed shapes with experimentally measured shapes of drops or bubbles is a useful method of measuring surface tension. If all lengths in Equation 47.3 are made dimensionless by dividing them by b, the resulting equation contains only one parameter, $\beta = \Delta\rho g b^2/\gamma$, which is called the Bond number (or shape factor). The shape of an axisymmetric drop, bubble, or meniscus depends only on this one dimensionless parameter. The Bond number can also be written as $\beta = 2b^2/a^2$ where $a = \sqrt{2\gamma/\Delta\rho g}$ is known as the capillary constant and has units of length.

Several dynamic methods of measuring surface tension are based on capillary waves. Capillary waves result from oscillations of the liquid surface for which surface tension is the restoring force. The frequency of the surface oscillation is thus dependent on the surface tension and wavelength. Very low amplitude capillary waves with a broad range of frequencies are always present on liquid surfaces owing to thermal fluctuations. Larger-amplitude capillary waves can be excited by purposely perturbing the surface.

47.2 Standard Methods and Instrumentation

A number of commonly used methods of measuring surface tension exist. The choice of a method depends on the system to be studied, the degree of accuracy required, and possibly on the ability to automate the measurements. In the discussion that follows, these methods are grouped according to the kind of instruments used in the measurements. Because the information presented for each method is necessarily brief, readers who are interested in constructing their own apparatus should consult the more detailed treatises in [1–4]. A list of commercially available instruments is given in Table 47.1, together with manufacturer names and approximate prices. Vendors can be contacted at the addresses given in Table 47.2.

TABLE 47.1 Commercially Available Instruments

Method	Instrument Type	Manufacturer/Product Name	Approximate Price (Range)
Capillary rise	Manual	Fisher	$79
Wilhelmy plate/du Noüy ring	Manual, mechanical balance	CSC, Fisher, Kahl	$2,000–$4,000
Wilhelmy plate/du Noüy ring	Manual, electrobalance	KSV, Lauda, NIMA	$4,000–$11,000
Wilhelmy plate/du Noüy ring	Automatic, electrobalance	Cahn, Krüss, KSV, NIMA	$9,000–$24,000
Maximum bubble pressure	Automatic	Krüss, Lauda, Sensa Dyne	$5,000–$23,000
Pendant/sessile drop	Manual	Krüss, Rame-Hart	$7,000–$10,000
Pendant/sessile drop	Automatic	ADSA, AST, FTA, Krüss, Rame-Hart, Temco	$10,000–$100,000
Drop weight/volume	Automatic	Krüss, Lauda	$16,000–$21,000
Spinning drop	Manual	Krüss	$20,100

Note: Price ranges reflect differences in degree of automation, the number of accessories included, or variation in price between manufacturers.

TABLE 47.2 Manufacturers and Suppliers of Instruments for Surface Tension Measurement

AST Products, 9 Linnell Circle, Billerica, MA 01821-3902, Tel.: (508) 663-7652, http://www.astp.com/

Applied Surface Thermodynamics Research Associates (ASTRA), 15 Brendan Rd., Toronto, Ontario, Canada M4G 2W9, Tel.: (416) 978-3601

Brinkmann Instruments, Inc. (distributor for Lauda tensiometers), One Catiaque Road, P.O. Box 1019, Westbury, NY 11590-0207, Tel.: (800) 645-3050

Cahn Instruments, 5225 Verona Rd., Bldg. 1, Madison, WI 53711, Tel.: (800) 244-6305

CSC Scientific Company, Inc., 8315 Lee Highway, Fairfax, VA 22031, Tel.: (800) 458-2558, http://www.cscscientific.com/

CTC Technologies, Inc. (distributor for NIMA tensiometers), 7925-A North Oracle Road, Suite 364, Tucson, AZ 85704-6356, Tel.: (800) 282-8325, http://www.ctctechnologies.com/

First Ten Angstroms (FTA), 465 Dinwiddie Street, Portsmouth, VA 23704, Tel.: (800) 949-4110, http://firsttenangstroms.com/

Fisher Scientific, 711 Forbes Ave., Pittsburgh, PA 15219-4785, Tel.: (800) 766-7000, http://www.fishersci.com/ecomm/servlet/home?storeId=10652

Kahl Scientific Instrument Corp., P.O. Box 1166, El Cajon, CA 92022-1166, Tel.: (619) 444-2158

Krüss USA., 9305-B Monroe Road, Charlotte, NC 28270-1488, Tel.: (704) 847-8933, http://www.kruss.de/en/contact/kruess-worldwide/usa.html

KSV Instruments USA, P.O. Box 192, Monroe, CT 06468, Tel.: (800) 280-6216

Rame-Hart, Inc., 8 Morris Ave., Mountain Lakes, NJ 07046, Tel.: (201) 335-0560, http://ramehart.com/

Sensa Dyne Instrument Div., Chem-Dyne Research Corp., P.O. Box 30430, Mesa, AZ 85275-0430, Tel.: (602) 924-1744, http://www.sensadyne.com/

Temco, Inc., 4616 North Mingo, Tulsa, OK 74117-5901, Tel.: (918) 834-2337

47.2.1 Capillary Rise Method

If a glass capillary tube is brought into contact with a liquid surface, and if the liquid wets the glass with a contact angle of less than 90°, then the liquid is drawn up into the tube as shown in Figure 47.5a. The surface tension is directly proportional to the height of rise, h, of the liquid in the tube relative to the flat liquid surface in the larger container. By applying Equation 47.2 to the meniscus in the capillary tube, the following relationship is obtained:

$$\Delta\rho g h = \frac{2\gamma}{b} \tag{47.4}$$

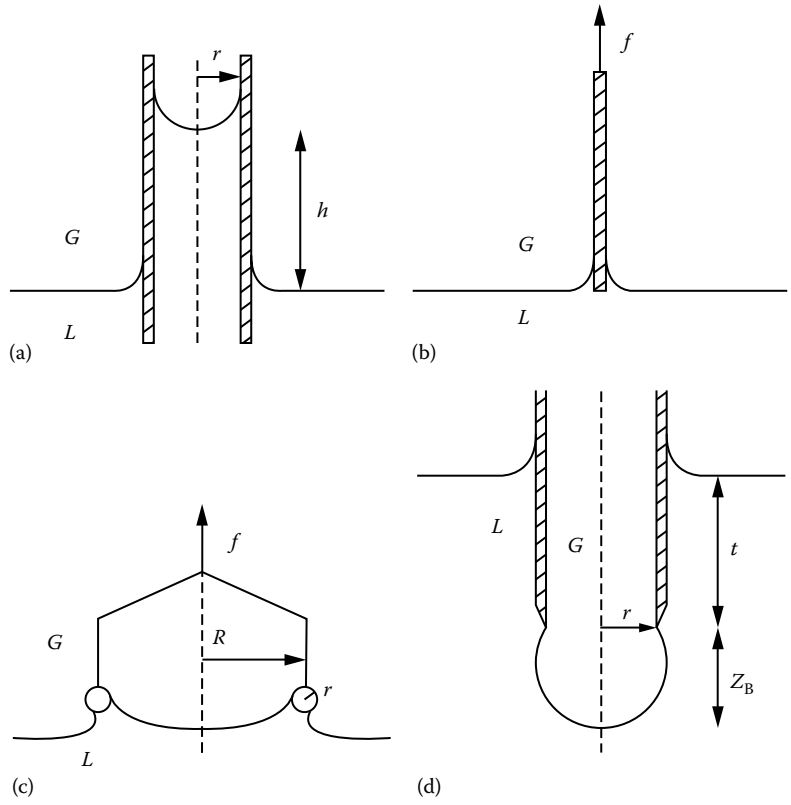


FIGURE 47.5 Geometries for (a) capillary rise method, (b) Wilhelmy plate method, (c) du Noüy ring method, and (d) maximum bubble pressure method.

where

b is the radius of curvature at the center of the meniscus

$\Delta\rho$ is the density difference between liquid and gas

For small capillary tubes, b is well approximated by the radius of the tube itself, assuming that the contact angle of the liquid on the tube is zero. For larger tubes or for increased accuracy, the value of b must be corrected for gravitational deformation of the meniscus (p. 12 of [1]). Obtaining accurate results with the capillary rise method requires using a thoroughly clean glass capillary tube with a very uniform diameter of less than 1 mm. The container for the liquid should be at least 8 cm in diameter, and the liquid must wet the capillary tube with a contact angle of zero. This method is primarily useful for pure liquids and is capable of high accuracy at relatively low cost.

47.2.2 Wilhelmy Plate and du Noüy Ring Methods

Measurement of the pull of a liquid surface directly on a solid object is the basis for two of the standard methods discussed here. In the Wilhelmy plate method, the solid object is a flat, thin plate that the test liquid should wet with a zero contact angle. The plate is suspended vertically from a delicate balance that is zeroed with the plate suspended in air. The test liquid is brought into contact with the bottom of the plate, causing the plate to be pulled down into the liquid by the surface tension force. The force applied to the plate from above is then increased to bring the bottom edge of the plate level with the flat surface

of the liquid (Figure 47.5b). This avoids the necessity to make buoyancy corrections to the measurement. The surface tension is computed from the force measurement, f, using

$$\gamma = \frac{f\cos\theta}{[2(l+t)]} \tag{47.5}$$

where
l is the length of the plate
t is its thickness

The contact angle θ is often assumed to be zero for common liquids on clean glass or platinum plates, but one should be aware of the error caused by a nonzero contact angle. No other correction factors are necessary for this method, and the fluid density does not need to be known.

The du Noüy ring method is known as a maximum pull method, of which there are several variations. The technique is to contact the liquid surface with a ring and then measure the force continuously as the surface is lowered until a maximum force, f_{max}, is recorded. The maximum force typically occurs just before the ring detaches from the surface. The surface tension is obtained from the formula

$$\gamma = \left(\frac{f_{max}}{4\pi R}\right)\left[F\left(\frac{R^3}{V}, \frac{R}{r}\right)\right] \tag{47.6}$$

where
R and r are the radii of the ring and wire, respectively, as indicated in Figure 47.5c
V is the volume of liquid raised by the ring
F is a correction factor (F is tabulated in Table 5, p. 132 of [4])

The du Noüy ring method requires knowledge of the liquid density, ρ_L, in order to determine V from $V = f_{max}/\rho_L$. This method requires the liquid to wet the ring with zero contact angles, and it is not suitable for solutions that attain surface equilibrium slowly.

A single instrument is normally capable of performing either Wilhelmy plate or du Noüy ring measurements. Some commercially available instruments can perform the complete measurement procedure automatically. Computer interfacing with a Wilhelmy plate instrument allows automatic data logging that can be used to follow changes in surface tension with time in surfactant solutions.

47.2.3 Maximum Bubble Pressure Method

The *maximum bubble pressure method* (MBPM) involves direct measurement of the pressure in a bubble to determine the surface tension. A tube is lowered to a depth t in the test liquid, and gas is injected to form a bubble of height Z_B at the tip of the tube as shown in Figure 47.5d. The increase in bubble pressure, P_b, over ambient pressure, P_a, arising from the interface is given by the sum of a hydrostatic pressure and Laplace pressure:

$$\delta p = P_b - P_a - \Delta\rho g t = \Delta\rho g Z_B + \frac{2\gamma}{b} \tag{47.7}$$

As a new bubble begins to form, Z_B increases while b, the radius of curvature at the bubble apex, decreases, resulting in an increase in pressure in the bubble. Ultimately, b increases as the bubble

grows larger, thus reducing the pressure. The pressure in the bubble thus reaches a maximum when δp reaches a maximum, which in turn can be theoretically related to the surface tension. For $\delta p = \delta p_{max}$, Equation 47.7 can be rewritten in dimensionless form as follows:

$$\frac{r}{X} = \frac{r}{b} + \frac{r}{a}\frac{Z_B}{b}\left(\frac{\beta}{2}\right)^{1/2} \tag{47.8}$$

where
- r is the tube radius
- X is a length defined as $X = 2\gamma/\delta p_{max}$
- a is the capillary constant
- β is the Bond number

The dimensionless quantity rlX depends only on r/a, the relationship being determined by Equation 47.8 combined with numerical solutions of Equation 47.3. Tabulations of this relationship are used to calculate the surface tension by an iterative procedure (p. 18 of [1]). The standard MBPM requires knowledge of the fluid densities, tube radius, and depth of immersion of the tube.

A differential MBPM uses two tubes of different diameters immersed to the same depth. The difference in the maximum bubble pressure for the two tubes, ΔP, is measured, eliminating the need to know the immersion depth and making the method less sensitive to errors in the knowledge of the liquid density. For the differential MBPM, surface tension is computed from the following (see [5]):

$$\gamma = A\Delta P\left[1 + \left(\frac{0.69 r_2 \rho_L}{\Delta P}\right)\right] \tag{47.9}$$

where
- r_2 is the radius of the larger tube
- ρ_L is the liquid density
- A is an apparatus-dependent constant that is determined by calibration with several standard liquids [6]

Automated MBPM units are commercially available (see Table 47.1). Sensa Dyne manufactures differential MBPM units that allow for online process measurements under conditions of varying temperature and pressure.

47.2.4 Pendant Drop and Sessile Drop Methods

The shape of an axisymmetric pendant or sessile drop (Figure 47.4) depends on only a single parameter, the Bond number, as discussed previously. The Bond number is a measure of the relative importance of gravity to surface tension in determining the shape of the drop. For Bond numbers near zero, surface tension dominates and the drop is nearly spherical. For larger Bond numbers, the drop becomes significantly deformed by gravity. In principle, the method involves obtaining an image of the drop and comparing its shape and size to theoretical profiles obtained by integrating Equation 47.3 for various values of β and b. Once β and b have been determined from shape and size comparison, the surface tension is calculated from

$$\gamma = \frac{\Delta\rho g b^2}{\beta} \tag{47.10}$$

In practice, the drop shape and size have traditionally been determined by the manual measurement of several characteristic dimensions (see Figure 47.4) of the drop from a photographic print. For pendant drops, the ratio d_s/d_e is correlated to a shape factor H from which surface tension is calculated ([3, p. 27; 7]) according to

$$\gamma = \frac{\Delta\rho g d_e^2}{H} \tag{47.11}$$

For sessile drops, various analytical formulae are available for computation of surface tension directly from the characteristic dimensions (p. 36 of [3]). Drop shape methods based on characteristic dimensions require very accurate measurement of the dimensions for good results. For more accurate results, methods that fit the entire shape of the edge of the drop to the Laplace equation are recommended.

In recent years, the entire procedure has been automated using digital imaging and computer image analysis [8,9]. Typically, several hundred coordinates on the edge of the drop are located with subpixel resolution by computer analysis of the digital image. The size, shape, and horizontal and vertical offsets of the theoretical profile given by Equation 47.3 are varied by varying four parameters: b, β, and the pixel coordinates of the drop apex, x_0 and z_0. A best fit of the theoretical profile to the measured edge coordinates is obtained by minimizing an objective function. A digital image of a pendant drop can be analyzed for surface tension on a desktop computer in 1 or 2 s [10]. The speed of algorithms for pendant drop analysis on modern desktop computers has allowed this method to be used to track changes in surface tension for surfactant-covered surfaces by analyzing a sequence of images. The algorithms can simultaneously track the surface area and volume of the drop or bubble. Both soluble and insoluble surfactants have been studied using the pendant drop, sessile drop, pendant bubble, and captive bubble configurations [11–13]. Table 47.1 lists several manufacturers that can provide software for automated analysis of surface tension from drops or bubbles in pendant or sessile configurations. The increased accuracy and simplicity of the automated pendant drop procedure makes it a very flexible method that has been applied to measure ultralow interfacial tensions, pressure, temperature, and time dependence of interfacial tension, relaxation of adsorption layers, measurement of line tensions, and film-balance measurements [14].

47.2.5 Drop Weight or Volume Method

A pendant drop will become unstable and detach from its support if it grows too large. The weight of the detached portion is related to the surface tension of the fluid by

$$\gamma = \left(\frac{mg}{r}\right)\left[F\left(\frac{r}{V^{1/3}}\right)\right] \tag{47.12}$$

where

mg is the weight of the detached drop
r is the radius of the tip from which the drop hangs
V is the volume of the detached drop

An empirical correction factor, F, is tabulated as a function of $r/V^{1/3}$ (p. 50 of [3]). For Equation 47.12 to apply, drops must be formed slowly. Measurements typically involve weighing the accumulated liquid from a large number of drops to determine the average weight per drop. The density of the fluid must be known in order to determine the drop volume and then obtain the factor F. Another method involves

measuring the volumetric flow rate of liquid to the tip while counting the drops. The density of the fluid must be known in order to determine the drop weight. The latter method allows for automation of measurements [15].

47.2.6 Spinning Drop Method

The spinning drop method is a shape-measurement method similar to the pendant and sessile drop methods. However, the deformation of the drop in this case is caused by radial pressure gradients in a rapidly spinning tube. This method is normally used for measuring interfacial tensions between immiscible liquids. A horizontal glass tube with sealed ends is filled with the more dense liquid through a filling port. The tube is then spun about its axis, while a drop of the lower-density liquid is injected into the tube. The pressure in the outer liquid increases from the center of the tube toward the walls as a result of the spinning motion. The pressure gradient forces the drop to move to the center of the tube and causes it to elongate, while surface tension opposes elongation. Measurement of the maximum drop diameter, $2r_{max}$, and length, $2h_{max}$, together with the angular velocity of rotation, Ω, allows for calculation of the surface tension according to

$$\gamma = \frac{1}{2}\left(\frac{r_{max}}{r^{\star}_{max}}\right)^3 \Delta\rho\Omega^2 \tag{47.13}$$

where $r^{\star}_{max}$ is correlated to the aspect ratio r_{max}/h_{max} [16]. The spinning drop method is particularly suited for measuring ultralow interfacial tensions (10^{-2} to 10^{-4} mN m^{-1}).

47.3 Specialized Methods

47.3.1 Dynamic Surface Tension

In an aqueous solution of soluble surfactant, the surface tension decreases following creation of new surface area because of adsorption of surfactant molecules. Surfactant adsorption kinetics can be studied by measuring the change in surface tension with time. For a dilute solution, the rate of change of surface tension is often slow enough that automated versions of static methods such as the Wilhelmy plate or pendant drop [17] methods can be used to follow the changes in surface tension. In concentrated solutions in which large changes in surface tension can occur within a fraction of a second following surface creation, a dynamic method must be used. A liquid jet emerging from an elliptical orifice has stationary waves on its surface, the wavelengths of which are related to the surface tension. The oscillating jet method has been used to measure surface tension for surface ages as low as 0.6 ms [18]. A dynamic version of the MBPM has been used to measure dynamic surface tension at surface ages down to 0.1 ms [19].

47.3.2 Surface Viscoelasticity

A liquid surface covered by a monolayer of surfactant exhibits viscoelastic behavior. In addition to surface tension, the surface rheology is characterized in terms of dilatational and shear elasticities as well as dilatational and shear viscosities. The dilatational properties in particular are important in a variety of situations from foam stability to the functioning of the human lung. The surface dilatational modulus is proportional to the change in surface tension for a given change in surface area. This modulus depends on the rate of change of surface area for both soluble and insoluble surfactant monolayers, which indicates that relaxation processes are active. These relaxation processes give rise to the surface dilatational viscosity. For the case of soluble surfactants, one of the relaxation processes

is the adsorption or desorption of molecules at the surface. The equilibrium dilatational elasticity of an insoluble monolayer can be measured by slowly expanding or compressing the monolayer in a Langmuir trough while monitoring the surface tension with a Wilhelmy plate apparatus [20]. Studies of surface rheology at high rates of surface expansion or compression are of interest for both soluble and insoluble surfactants.

Surface tension relaxation following sudden expansion or compression of the surface for a soluble surfactant has been studied by the automated pendant drop method [21]. A method known as oscillating bubble tensiometry has been applied to measure the kinetics of adsorption and desorption for soluble surfactants [22,23]. Other methods for studying dynamic dilatational viscoelastic properties are reviewed in [24], including transverse and longitudinal capillary wave methods, a modified MBPM, and an oscillating bubble method.

47.3.3 Measurements at Extremes of Temperature and Pressure

Several of the standard methods described in this chapter can be adapted to make surface or interfacial tension measurements at extreme temperatures and/or pressures. The most common methods used to measure the surface tension of high-temperature molten metals, alloys, and semiconductors are the MBPM [25] and the pendant or sessile drop method [26–28]. Measurement of the interfacial tension between oil and a second immiscible phase at high pressure and elevated temperature is of interest for understanding aspects of enhanced oil recovery. The pendant drop method has been applied under pressures of 82 MPa at 449 K [29], while a capillary wave method has been applied at 136 MPa and 473 K [30]. A pendant drop apparatus capable of measurements to 10,000 psi (69 MPa) and 350 °F (450 K) is commercially available from Temco Inc. (Table 47.2).

47.3.4 Interfacial Tension

Measurement of the interfacial tension between two immiscible liquids can present special difficulties. Measurement by the capillary rise, du Noüy ring, or Wilhelmy plate method is problematic in that the contact angle is often nonzero. The pendant drop [7] and drop weight [31] methods can both be applied, provided the densities of the two liquids are sufficiently different. The pendant drop method, in particular, is widely used for interfacial tension measurement. Interfacial tension can be measured by a modified MBPM in which one measures the maximum pressure in a liquid drop injected into a second immiscible liquid [32]. The modified MBPM [32] and a liquid bridge method [33] have been used to measure interfacial tension between two liquids of equal density. Ultralow values of interfacial tension can be measured by the spinning drop [34], pendant drop [35], and capillary wave methods [34].

Defining Terms

Surface tension: A force per unit length that acts tangential to a liquid surface and perpendicular to any line that lies within the surface.

Surface energy: The excess free energy per unit area associated with a surface between two phases. For a liquid–fluid surface, the surface energy is numerically equivalent to the surface tension.

Acknowledgments

During the preparation of this chapter, one of us (DBT) was supported in part by the National Aeronautics and Space Administration (NASA) and by the Office of Naval Research. The work by one of us (KFM) was performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA.

References

1. A. W. Adamson, *Physical Chemistry of Surfaces*, 5th edn., New York: John Wiley & Sons, 1990.
2. A. E. Alexander and J. B. Hayter, Determination of surface and interfacial tension. In: A. Weissberger and B. W. Rossiter (Eds.), *Physical Methods of Chemistry*, Part V, 4th edn., New York: John Wiley & Sons, 1971.
3. A. Couper, Surface tension and its measurement, In: B. W. Rossiter and R. C. Baetzold (Eds.), *Physical Methods of Chemistry*, Vol. 9A, 2nd edn., New York: John Wiley & Sons, 1993.
4. J. F. Padday, Surface tension. II. The measurement of surface tension. In: E. Matijevic (Ed.), *Surface and Colloid Science*, Vol. 1, New York: John Wiley & Sons, 1969.
5. S. Sugden, The determination of surface tension from the maximum pressure in bubbles. Part II, *J. Chem. Soc*., 125, 27–31, 1924.
6. ASTM Standard D3825-90, Standard test method for dynamic surface tension by the fast-bubble technique, *1996 Annual Book of ASTM Standards*, Vol. 05.02, West Conshohocken, PA: ASTM, 1996, pp. 575–579.
7. D. S. Ambwani and T. Fort, Jr., Pendant drop technique for measuring liquid boundary tensions. In: R. J. Good and R. R. Stromberg (Eds.), *Surface and Colloid Science*, Vol. 11, New York: Plenum Press, 1979.
8. Y. Rotenberg, L. Boruvka, and A. W. Neumann, Determination of surface tension and contact angle from the shapes of axisymmetric fluid interfaces, *J. Colloid Interface Sci*., 93, 169–183, 1983.
9. P. Cheng, D. Li, L. Boruvka, Y. Rotenberg, and A. W. Neumann, Automation of axisymmetric drop shape analysis for measurements of interfacial tensions and contact angles, *Colloids Surf*., 43, 151–167, 1990.
10. D. B. Thiessen, D. J. Chione, C. B. McCreary, and W. B. Krantz, Robust digital image analysis of pendant drop shapes, *J. Colloid Interface Sci*., 177, 658–665, 1996.
11. S. Lin, K. McKeigue, and C. Maldarelli, Diffusion-controlled surfactant adsorption studied by pendant drop digitization, *AIChE J*., 36, 1785–1795, 1990.
12. D. Y. Kwok, D. Vollhardt, R. Miller, D. Li, and A. W. Neumann, Axisymmetric drop shape analysis as a film balance, *Colloids Surf. A*, 88, 51–58, 1994.
13. W. M. Schoel, S. Schurch, and J. Goerke, The captive bubble method for the evaluation of pulmonary surfactant: Surface tension, area, and volume calculations, *Biochim. Biophys. Acta*, 1200, 281–290, 1994.
14. S. Lahooti, O. I. Del Rio, P. Cheng, and A. W. Neumann, Axisymmetric drop shape analysis (ADSA). In: A. W. Neumann and J. K. Spelt (Eds.), *Applied Surface Thermodynamics*, New York: Marcel Dekker, 1996.
15. M. L. Alexander and M. J. Matteson, The automation of an interfacial tensiometer, *Colloids Surf*., 27, 201–217, 1987.
16. J. C. Slattery and J. Chen, Alternative solution for spinning drop interfacial tensiometer, *J. Colloid Interface Sci*., 64, 371–373, 1978.
17. D. Y. Kwok, M. A. Cabrerizo-Vilchez, Y. Gomez, S. S. Susnar, O. Del Rio, D. Vollhardt, R. Miller, and A. W. Neumann, Axisymmetric drop shape analysis as a method to study dynamic interfacial tensions. In: V. Pillai and D. O. Shah (Eds.), *Dynamic Properties of Interfaces and Association Structures*, Champaign, IL: AOCS Press, 1996.
18. W. D. E. Thomas and L. Potter, Solution/air interfaces. I. An oscillating jet relative method for determining dynamic surface tensions, *J. Colloid Interface Sci*., 50, 397–412, 1975.
19. V. B. Fainerman and R. Miller, Dynamic surface tension measurements in the sub-millisecond range, *J. Colloid Interface Sci*., 175, 118–121, 1995.
20. G. L. Gaines Jr., *Insoluble Monolayers at Liquid-Gas Interfaces*, New York: John Wiley & Sons, 1966, 44.
21. R. Miller, R. Sedev, K.-H. Schano, C. Ng, and A. W. Neumann, Relaxation of adsorption layers at solution/air interfaces using axisymmetric drop-shape analysis, *Colloids Surf*., 69, 209–216, 1993.

22. D. O. Johnson and K. J. Stebe, Oscillating bubble tensiometry: A method for measuring the surfactant adsorptive-desorptive kinetics and the surface dilatational viscosity, *J. Colloid Interface Sci*., 168, 21–31, 1994.
23. D. O. Johnson and K. J. Stebe, Experimental confirmation of the oscillating bubble technique with comparison to the pendant bubble method: The adsorption dynamics of 1-decanol, *J. Colloid Interface Sci*., 182, 526–538, 1996.
24. D. A. Edwards, H. Brenner, and D. T. Wasan, *Interfacial Transport Processes and Rheology*, Boston, MA: Butter worth-Heinemann, 1991.
25. C. Garcia-Cordovilla, E. Louis, and A. Pamies, The surface tension of liquid pure aluminium and aluminium-magnesium alloy, *J. Mater. Sci*., 21, 2787–2792, 1986.
26. B. C. Allen, The surface tension of liquid transition metals at their melting points, *Trans. Metall. Soc. AIME*, 227, 1175–1183, 1963.
27. D. B. Thiessen and K. F. Man, A quasi-containerless pendant drop method for surface tension measurements on molten metals and alloys, *Int. J. Thermophys*., 16, 245–255, 1995.
28. S. C. Hardy, The surface tension of liquid silicon, *J. Cryst. Growth*, 69, 456–460, 1984.
29. V. Schoettle and H. Y. Jennings, Jr., High-pressure high-temperature visual cell for interfacial tension measurement, *Rev. Sci. Instrum*., 39, 386–388, 1968.
30. R. Simon and R. L. Schmidt, A system for determining fluid properties up to 136 MPa and 473 K, *Fluid Phase Equilib*., 10, 233–248, 1983.
31. K. Hool and B. Schuchardt, A new instrument for the measurement of liquid-liquid interfacial tension and the dynamics of interfacial tension reduction, *Meas. Sci. Technol*., 3, 451–457, 1992.
32. A. Passerone, L. Liggieri, N. Rando, F. Ravera, and E. Ricci, A new experimental method for the measurement of the interfacial tension between immiscible fluids at zero Bond number, *J. Colloid Interface Sci*., 146, 152–162, 1991.
33. G. Pétré and G. Wozniak, Measurement of the variation of interfacial tension with temperature between immiscible liquids of equal density, *Acta Astronaut*., 13, 669–672, 1986.
34. D. Chatenay, D. Langevin, J. Meunier, D. Bourbon, P. Lalanne, and A. M. Bellocq, Measurement of low interfacial tension, comparison between a light scattering technique and the spinning drop technique, *J. Dispersion Sci. Technol*., 3, 245–260, 1982.
35. D. Y. Kwok, P. Chiefalo, B. Khorshiddoust, S. Lahooti, M. A. Cabrerizo-Vilchez, O. Del Rio, and A. W. Neumann, Determination of ultralow interfacial tension by axisymmetric drop shape analysis. In: R. Sharma (Ed.), *Surfactant Adsorption and Surface Solubilization* (*ACS Symp. Ser. 615*), Washington, DC: ACS, 1995, pp. 374–386.

Further Readings

Hartland, S., *Surface and Interfacial Tension: Measurement Theory and Applications*, New York: Marcel Dekker, 2005.

Miller, R., Pison, U., and Kazakov, V. N., *Dynamic Surface Tensiometry in Medicine*, New York: Elsevier Science Ltd., 2000.

Physical and chemistry Resources, Surface tension. http://physicalchemistryresources.com/Book2_sections_graphicfiles_08142009/htm_PPGLS_Surface%20Tension.htm (accessed on August 8, 2012).

Research and Articles, Surface tension. http://www.bookrags.com/research/surface-tension-wop/ (accessed August 1, 2012).

V

Acoustics

48 Acoustic Measurement *Per Rasmussen* **48**-1
Wave Equation • Plane Sound Waves • Spherical Waves • Acoustic Measurements • Sound Pressure Level Measurements • Frequency Analyzers • Pressure-Based Measurements • Sound Intensity Measurements • Near-Field Acoustic Holography Measurements • Calibration • References • Further Readings

49 Ultrasound Measurement *Peder C. Pedersen* **49**-1
Applications of Ultrasound • Basic Ultrasound Parameters • Conceptual Description of Ultrasound Imaging and Measurements • Single-Element and Array Transducers • Selection Criteria for Ultrasound Frequencies • Basic Parameters in Ultrasound Measurements • Ultrasound Theory and Applications • Review of Common Applications of Ultrasound and Their Instrumentation • Advanced Topics in Ultrasound • References • Further Information



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

48

Acoustic Measurement

48.1 Wave Equation **48**-2
48.2 Plane Sound Waves **48**-3
48.3 Spherical Waves **48**-3
48.4 Acoustic Measurements **48**-4
Condenser Microphones
48.5 Sound Pressure Level Measurements **48**-9
48.6 Frequency Analyzers **48**-12
48.7 Pressure-Based Measurements **48**-12
48.8 Sound Intensity Measurements **48**-14
48.9 Near-Field Acoustic Holography Measurements **48**-16
48.10 Calibration **48**-20
References **48**-22
Further Readings **48**-23

Per Rasmussen
G.R.A.S. Sound and Vibration

Sound is normally defined as vibration of a solid, liquid, or gaseous medium in the frequency range of the human ear, that is, between about 20 Hz and 20 kHz. Here, the definition is further limited and only vibrations in liquid and gaseous media are considered. In contrast to solid media, a liquid or gaseous medium cannot transmit shear forces, so sound waves are always longitudinal waves, in which the particles move in the direction of propagation of the wave. The wave propagation in gaseous and liquid media can be described by the three variables: the pressure p, the particle velocity u, and the density ρ. The relation between these is described by the wave equation [1], and this can be derived from three basic equations: the Euler equation (this is essentially Newton's second law applied to a fluid), the continuity equation, and the state equation. Although the wave equation, in principle, can be used to describe and calculate all sound waves in all situations, it will in practice often be impossible to perform the necessary calculations. In some special cases, it is possible to get analytical results directly from the wave equation, and these cases are therefore of special interest. The cases most often encountered in acoustics are the *free field*, the *diffuse* (or reverberant) *field*, and the *closed coupler*. The free field is, in principle, an infinite, empty (except for the medium and the source) space, with no reflections. Here, the waves are allowed to radiate freely in all directions without reflections. In practice, the free field is implemented in anechoic chambers, where all walls have been made nearly 100% absorptive. The diffuse field is obtained in a reverberation room where all walls have been made, in principle, 100% reflective. At the same time, the walls are made nonparallel and the result is a sound field with sound waves in all directions. The closed coupler is a small chamber, with dimensions small compared to the wavelength of the sound. A special case of this is the standing wave tube. This is a tube with a diameter smaller than the wavelength and with a sound source in one end. With a suitable loudspeaker as a source, the wave propagation in the tube can be assumed to be 1-D. This simplifies the mathematical description so that it is possible to calculate the sound field.

In practice, almost the only parameter measured directly in acoustics is the sound pressure, and all other parameters like sound power, particle velocity, reverberation time, and directivity are derived from pressure measurements. These are performed with measurement microphones in gaseous media and hydrophones in liquid media. The measurement microphones are all of the condenser type to ensure precision, long-term stability, and sensitivity. Hydrophones are usually made with a rubber coating over a sensitive element of piezoelectric material.

The traditional frequency range from 20 Hz to 20 kHz for acoustic measurements is selected because this is the range audible to the human ear. Sound waves exist outside this range in the form of infrasound (below 20 Hz) and ultrasound (above 20 kHz). As the basic equations (the wave equation) and measurement principles are the same for both infrasound and ultrasound, many of the principles from the frequency range from 20 Hz to 20 kHz can be extended to these ranges.

48.1 Wave Equation

Sound wave propagation cannot take place in a vacuum, but is always associated with some kind of medium. For simplicity, assume that this medium is air, although the same equations are valid also for all gaseous and fluid media. In this medium, the concept of an air particle can be introduced. An *air particle* is a small volume of air in which the acoustic parameters like pressure and density can be considered constant. On the other hand, the air volume must be large enough to include a very large number of air molecules, so that the air volume can be considered to be a continuous medium and not a collection of molecules. The *Euler equation* for such an air particle is given by

$$-\text{grad}\, p = \rho_0 \left(\frac{\partial \bar{v}}{\partial t} \right) \tag{48.1}$$

where

p is the pressure
ρ_0 is the density
$\bar{v}$ is the particle velocity

This equation can be considered as Newton's second law ($F = ma$) applied to a fluid. Here, the gradient of the pressure equals the force F acting on the air particle, the density equals the mass m, and the time differentiated particle velocity $\partial v/\partial t$ equals the acceleration.

The second equation necessary to derive the wave equation is the *continuity equation*. This simply states that if you have a small volume of air and you bring in some extra air, the density (or the mass) will increase. Mathematically, this can be formulated as

$$\text{div}\, \bar{v} = -\frac{1}{\rho_0} \cdot \frac{\partial \rho}{\partial t} \tag{48.2}$$

where c is the sound velocity. The sound velocity depends on the composition of the air and the temperature. For normal air at 0 °C, the velocity is 314 m s^{-1}, while at 20 °C, the velocity is 340 m s^{-1}.

The third equation is the *state equation*, which relates pressure changes to changes in the density, that is, if a small volume of air is compressed, the density will increase:

$$\frac{\partial \rho}{\partial t} = \frac{1}{c^2} \times \frac{\partial p}{\partial t} \tag{48.3}$$

Now we have three equations relating the three variables: pressure, particle velocity, and density. By eliminating the particle velocity and the density, we obtain one differential equation for the sound pressure:

$$\nabla^2 p = \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} \tag{48.4}$$

This is the *wave equation* for acoustic waves in gaseous and fluid media. In principle, this allows one to calculate the sound pressure anywhere in a sound field, if some suitable boundary conditions are given. In practice, however, it is only possible to find solutions in a few simple cases.

48.2 Plane Sound Waves

In a free space, at great distance from the sound source, sound waves are approximately plane waves. This means that the wave equation only depends on one of the coordinates in the wave equation. If the direction of propagation of the wave fronts is in the x-direction, the solution to the wave equation reduces to

$$p = A\cos[(\omega t - k)(x + \varphi_a)] \tag{48.5}$$

where ω is the frequency. Similarly, the particle velocity in the free field is given by

$$v = \frac{A}{\rho c}\cos[(\omega t - k)(x + \varphi_a)] = \frac{p}{\rho c} \tag{48.6}$$

This means that in a plane wave, the particle velocity is equal to the pressure divided by the constant ρc and the pressure and particle velocity are inphase. The constant ρc is the *acoustic impedance*, and for air at 20 °C, the density is 1.29 kg m^{-3} and the sound velocity is 340 m s^{-1}, giving an acoustic impedance of 438.6 kg m^{-2} s^{-1}.

The plane wave transmits energy in the direction of propagation. The power transmitted per unit area is the intensity in the direction of the propagation (in many older textbooks, terms like "the intensity of the sound" were mistakenly used for the magnitude of the sound pressure). In general, the intensity is given by the product of the sound pressure and the sound velocity; thus, in the case of the plane wave, the intensity (I) can be calculated from Equations 48.5 and 48.6:

$$I = vp = \frac{p^2}{\rho c} \tag{48.7}$$

Thus, in the plane wave, the intensity transmitted by the wave can be calculated from the sound pressure, and, as the sound power is the intensity per unit area, the sound power can be calculated by multiplying the intensity by the area.

48.3 Spherical Waves

Another simple solution to the wave equation can be found for the radiation from a point source into free space. The point source is an infinitely small sphere whose surface is pulsating radially. In practice, for small sound sources (i.e., where the dimensions of the sound source are small compared to the wavelength of the sound), the point source is a good approximation for the real physical source that makes the spherical wave solution of special interest.

The wave equation in Equation 48.5 is transformed into the spherical coordinates r, θ, and ψ. As the point source radiates equally in all directions, the solution depends only on the distances r from the center of the point source:

$$p = \frac{P_0}{r}\cos(\omega t - kr) \tag{48.8}$$

It can be seen that the sound pressure is inversely proportional to the distance from the sound source. The particle velocity can be divided into a near-field contribution v_n and a far-field contribution v_f:

$$v_f = \frac{P_0}{r\rho c}\cos(\omega t - kr) \tag{48.9}$$

$$v_n = \frac{P_0}{\omega\rho r^2}\sin(\omega t - kr) \tag{48.10}$$

The far-field contribution in Equation 48.9 can be seen to be inphase with the pressure in Equation 48.8, and also the particle velocity is inversely proportional to the distance r from the point source. The near-field contribution is inversely proportional to the square of the distance to the source and therefore dies away rapidly as the distance to the source increases.

As in the case of the plane wave, the intensity in the spherical wave is the product of the pressure and particle velocity. For the near-field contribution, one obtains

$$I = v_n p = \frac{P_0}{\omega\rho r^2}\sin(\omega t - kr)\frac{P_0}{r}\cos(\omega t - kr) = 0 \tag{48.11}$$

That is, the near-field part of the particle velocity does not contribute to the radiated power as the particle velocity is 90° out of phase with the pressure.

The far-field contribution is given by

$$I = v_f p = \frac{P_0}{r\rho c}\cos(\omega t - kr)\frac{P_0}{r}\cos(\omega t - kr) = \frac{P_0^2}{r^2\rho c}\cos^2(\omega t - kr) \tag{48.12}$$

It can be seen that the intensity decreases with the square of the distance to the source and by combining Equations 48.8 and 48.12, one obtains

$$I = \frac{P_0^2}{r^2\rho c}\cos^2(\omega t - kr) = \frac{p^2}{\rho c} \tag{48.13}$$

which is identical to Equation 48.7 for the plane wave. Thus, as for the plane progressive wave, the intensity in the spherical wave can be calculated from the pressure.

48.4 Acoustic Measurements

As can be seen from the wave equation, the full acoustic field can in principle be described from only pressure measurements. This means that all other acoustic parameters can be derived from pressure measurements, and, in practice, pressure is often the only parameter measured. There have been a few attempts to make transducers for particle velocity measurements based on, for example, transmission of ultrasonic waves, but the absolute dominating transducers for acoustic measurements are the condenser-type microphones, as shown in Figure 48.1. These have proven to be superior with respect to temperature stability, long-term stability, and insensitivity to rough handling. While measurement microphones are designed and produced to ensure well-defined and accurate measurements, a wide range of other microphones are available for other purposes. These can, for example, be for incorporation in telephones, where price is a very decisive factor, or for studio recordings, where a subjective evaluation is more important than the objective performance.

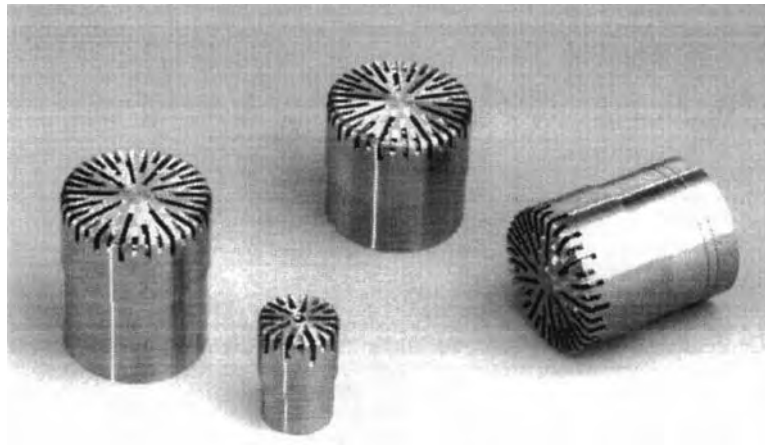


FIGURE 48.1 Measurement microphones: ½ and ¼ in.

48.4.1 Condenser Microphones

The condenser microphone consists basically of five elements: protection grid, microphone casing, diaphragm, backplate, and insulator; see Figure 48.2. The diaphragm and the backplate form the parallel plates of an air capacitor. This capacitor is polarized with a charge from an external voltage supply (externally polarized type) or by an electric charge injected directly into an insulating material on the backplate (prepolarized type). When the sound pressure in the sound field fluctuates, the distances between the diaphragm and the backplate will change and consequently change the capacitance of the diaphragm/backplate capacitor. As the charge on the capacitor is kept constant, the change in capacitance will generate an output voltage on the output terminal of the microphone. The acoustic performance of a microphone is determined by the physical dimensions such as diaphragm area, the distance between the diaphragm and the backplate, the stiffness and mass of the suspended diaphragm, and the internal volume of the microphone casing. These factors will determine the frequency range of the microphone, the sensitivity, and the dynamic range. The sensitivity of the microphone is described as the output voltage of the microphone for a given sound pressure excitation, and is in itself of little interest for the operation of the microphone, except for calibration purposes. However, the sensitivity of the

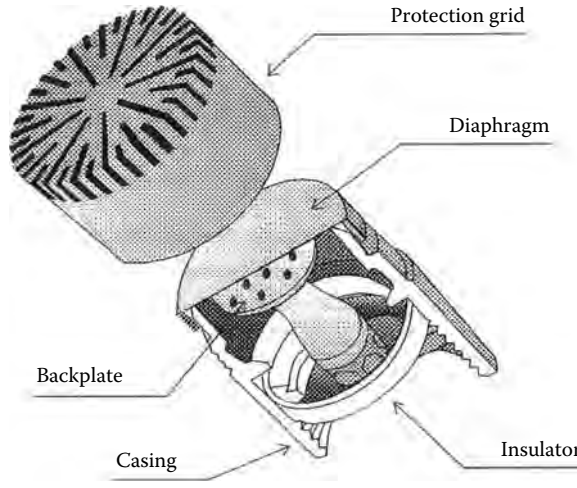


FIGURE 48.2 Basic elements of a measurement microphone.

microphone (together with the electric impedance of the cartridge) also determines the lowest sound pressure level that can be measured with the microphone. For example, with a microphone with a sensitivity of 2.5 mV Pa^{-1}, the lowest level that can be measured is around 40 dB (re. 20 μPa), while a microphone with a sensitivity of 50 mV Pa^{-1} can measure levels down to approximately 15 dB (re. 20 μPa).

The size of the microphone is the first parameter determining the sensitivity of the microphone. In general, the larger the diaphragm diameter, the more sensitive the microphone will be. There are, however, limits to how sensitive the microphone can be made by simply making it larger. The polarization voltage between the diaphragm and the backplate will attract the diaphragm and deflect this toward the backplate. As the size of the microphone is increased, the deflection will increase and eventually the diaphragm will be deflected so much that it will touch the backplate. To avoid this, the distance between the diaphragm and the backplate can be increased or the polarization voltage can be decreased. Both of these actions will, however, decrease the sensitivity, so that the optimum size of a practical measurement microphone for use up to 20 kHz is very close to ½ in. (12.6 mm).

As the size of the microphone is decreased, the useful frequency range of the microphone is increased. The frequency range, which can be obtained, is determined in part by the size of the microphone. At high frequencies, when the wavelength of the sound waves becomes much smaller than the diameter of the diaphragm, the diaphragm will stop behaving like a rigid piston (the diaphragm "breaks up"—this is not a destructive phenomenon). Different parts of the diaphragm will start to move with different magnitude and phase, and the frequency response of the microphone will change. To avoid this, the upper limiting frequency is placed so that the sensitivity of the microphone drops off before the diaphragm starts to break up. This gives, for a typical 0.5 in. microphone, an upper limiting frequency in the range from 20 to 40 kHz, depending on the diaphragm tension. If the diaphragm is tensioned so that it becomes stiffer, the resonance frequency of the diaphragm will be higher; on the other hand, the sensitivity of the microphone will be reduced as the diaphragm deflection by a certain sound pressure level decreases.

The frequency response of the microphone is determined by the diaphragm tension, the diaphragm mass, and the acoustic damping in the air gap between the diaphragm and the backplate. This system can be represented by the mechanical analogy of a simple mass–spring–damper system as in Figure 48.3. The mass in the analogy represents the mass of the diaphragm, and the spring represents the tension in the diaphragm. Thus, if the diaphragm is tensioned to become stiffer, the corresponding spring will become stiffer. The damping element in the analogy represents the acoustic damping between the diaphragm and the backplate. This can be adjusted by, for example, drilling holes in the backplate. This will make it easier for the air to move away from the air gap when the diaphragm is deflected, and therefore decrease damping.

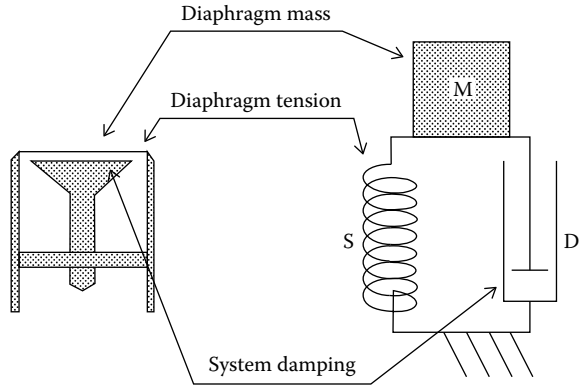


FIGURE 48.3 Mechanical analogy of a microphone.

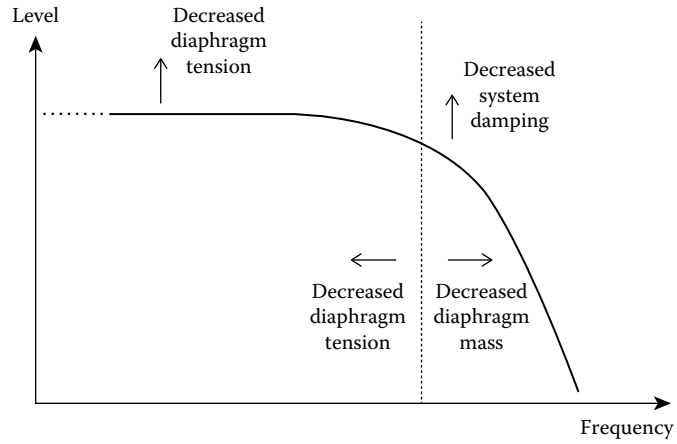


FIGURE 48.4 Influence of microphone parameters on frequency response.

The frequency response of the simple mechanical model of the microphone is given in Figure 48.4, together with the influence of the different parameters. At low frequencies (below the resonant frequency), the response of the microphone is determined by the diaphragm tension, and as described previously, the sensitivity will increase if the tension is decreased. The resonant frequency is determined by the diaphragm tension and the diaphragm mass, with an increased tension giving an increased resonant frequency and an increased mass giving a decreased resonant frequency. The response around the resonant frequency is determined by the acoustic damping, where an increase in the damping will decrease the response.

Although the material selection and assembling techniques have changed during the last few years, the basic types of microphones remain unchanged. The basic types are the free-field microphones, the pressure microphones, and the random incidence microphones. They have been constructed with different frequency characteristics, corresponding to the different requirements.

The *pressure microphone* is meant to measure the actual sound pressure as it exists on the diaphragm. A typical application could be the measurement of the sound pressure in a closed coupler or as in Figure 48.5,

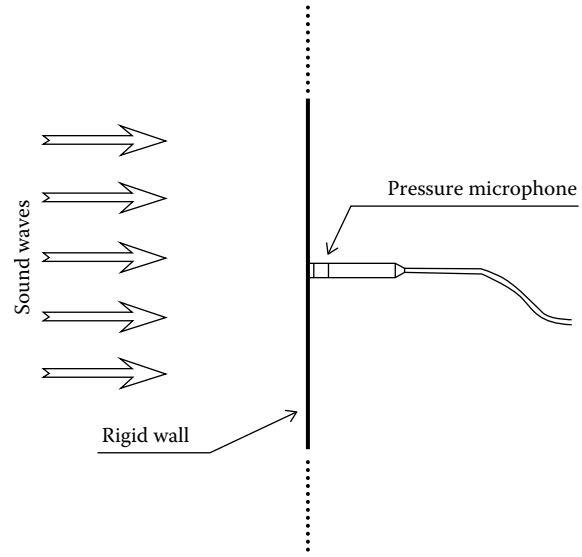


FIGURE 48.5 Application of pressure microphones.

the measurement of the sound pressure at a boundary. In this case, the microphone forms part of the wall and measures the sound pressure on the wall itself. The frequency response of this microphone should be flat in a frequency range as wide as possible, taking into account that the sensitivity will decrease as the frequency range is increased. The acoustic damping in the air gap between the diaphragm and the backplate is adjusted so that the frequency response is flat up to and a little beyond the resonant frequency.

The *free-field microphone* is designed to essentially measure the sound pressure as it existed before the microphone was introduced into the sound field. At higher frequencies, the presence of the microphone itself in the sound field will change the sound pressure. In general, the sound pressure around the microphone cartridge will increase due to reflections and diffraction. The free-field microphone is designed so that the frequency characteristics compensate for this pressure increase. The resulting output of the free-field microphone is a signal proportional to the sound pressure as it existed before the microphone was introduced into the sound field. The free-field microphone should always be pointed toward the sound source (0° incidence), as in Figure 48.6. In this situation, the presence of the microphone diaphragm in the sound field will result in a pressure increase in front of the diaphragm (see Figure 48.7a,) depending on the wavelength of the sound waves and the microphone diameter. For a typical ½ in. microphone, the maximum pressure increase will occur at 26.9 kHz, where the wavelength of the sound ($\lambda = 342$ m s^{-1}/26.9 kHz ≈ 12.7 mm ≈ 0.5 in.) coincides with the diameter of the microphone. The microphone is then designed so that the sensitivity of the microphone decreases by the same amount as the acoustic pressure increases in front of the diaphragm. This is obtained by increasing the internal acoustic damping in the microphone cartridge, to obtain a frequency response as in Figure 48.7b. The result is an output from the microphone, Figure 48.7c, which is proportional to the sound pressure as it existed before the microphone was introduced into the sound field. The curve in Figure 48.7a is also called the "free-field correction curve" for the microphone, as this is the curve that must be added to the frequency response of the microphone cartridge to obtain the acoustic characteristic of the microphone in the free field.

The free-field microphone is required in principle, to be pointed toward the sound source and that the sound waves travel in essentially one direction. In some cases, (e.g., when measuring in a reverberation room or other highly reflecting surroundings), the sound waves will not have a well-defined propagation direction, but will arrive at the microphone from all directions simultaneously. The sound waves arriving at the microphone from the front will cause a pressure increase, as described for the free-field microphone, while the waves arriving from the back of the microphone will be decreased to a certain extent due to the shadowing effects of the microphone cartridge. The combined influence of the waves coming from different directions therefore depends on the distribution of sound waves from different directions. For measurement microphones, a standard distribution has been defined, based on statistical considerations, resulting in a standardized random incidence microphone.

As mentioned previously, measurement microphones can be of either the externally polarized type or the prepolarized type. The externally polarized types are by far the most stable and accurate

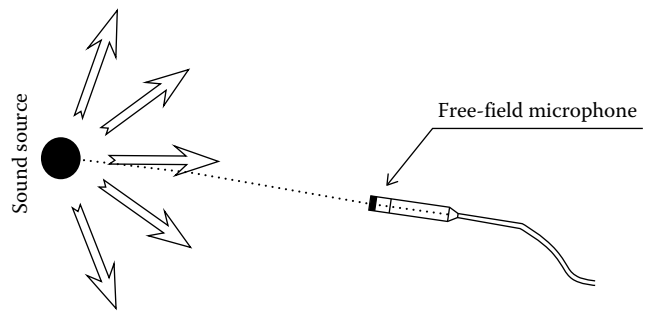


FIGURE 48.6 Application of a free-field microphone.

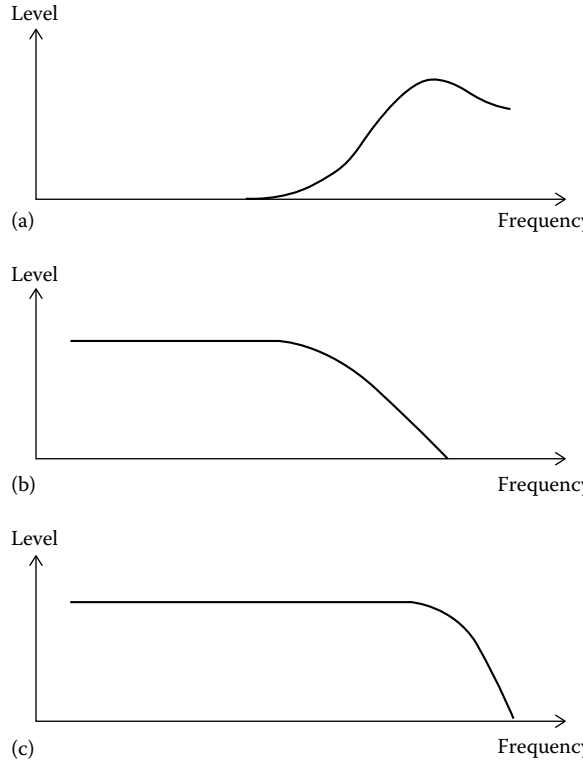


FIGURE 48.7 Frequency response of free-field microphone: (a) pressure increase in front of diaphragm, (b) microphone pressure response, and (c) resulting microphone output.

microphones and should be preferred for precision measurements. The prepolarized microphones are, however, preferred in some cases, in that they do not require the external polarization voltage source. This is typically the case when the microphone will be used on small handheld devices like sound level meters, where a power supply for polarization voltage would add excessively to cost, weight, and battery consumption. Still, it should be realized that prepolarized microphones in general are much less stable to environmental changes than externally polarized microphones.

48.5 Sound Pressure Level Measurements

The human ear basically hears the sound pressure, but the sensitivity varies with the frequency. The human ear is most sensitive to sound in the frequency range from 1 to 5 kHz, while the sensitivity drops at higher and lower frequencies. This has led to the development of several frequency weighting functions, which attempt to replicate the sensitivity of the human ear. Also, the response of the human ear to time-varying signals and impulses has led to the development of instruments with well-defined time weighting functions. The resulting measurement instrumentation is the *sound level meter*, as defined in for example by the IEC International Standard 651, "Sound Level Meters" [2]. The standard defines four classes of sound level meters for different accuracies (Table 48.1). Type 0 is the most accurate, intended for precision laboratory measurements, while type 1 is most widely used for general-purpose measurements; see Figure 48.8. Type 2 is used where low price is of importance, while type 3 is not used in practice because of the wide tolerances, making the results too unreliable. The output of the sound level meter is, in principle, assumed to be an approximate measure of the impression perceived by the human ear.

TABLE 48.1 IEC 651 Sound Level Meter Requirements

Frequency (Hz)	Type 0 (dB)	Type 1 (dB)	Type 2 (dB)	Type 3 (dB)
10	+2; −∞	+3; −∞	+5; −∞	+5; −∞
12.5	+2; −∞	+3; −∞	+5; −∞	+5; −∞
16	+2; −∞	+3; −∞	+5; −∞	+5; −∞
20	±2	±3	±3	+5; −∞
25	±1.5	±2	±3	+5; −∞
31.5	±1	±1.5	±3	±4
40	±1	±1.5	±2	±4
50	±1	±1.5	±2	±3
63	±1	±1.5	±2	±3
80	±1	±1.5	±2	±3
100	±0.7	±1	±1.5	±3
125	±0.7	±1	±1.5	±2
160	±0.7	±1	±1.5	±2
200	±0.7	±1	±1.5	±2
250	±0.7	±1	±1.5	±2
315	±0.7	±1	±1.5	±2
400	±0.7	±1	±1.5	±2
500	±0.7	±1	±1.5	±2
630	±0.7	±1	±1.5	±2
800	±0.7	±1	±1.5	±2
1,000	±0.7	±1	±1.5	±2
1,250	±0.7	±1	±1.5	±2.5
1,600	±0.7	±1	±2	±3
2,000	±0.7	±1	±2	±3
2,500	±0.7	±1	±2.5	±4
3,125	±0.7	±1	±2.5	±4.5
4,000	±0.7	±1	±3	±5
5,000	±1	±1.5	±3.5	±6
6,300	+1; −1.5	+1.5; −2	±4.5	±6
8,000	+1; −2	+1.5; −3	±5	±6
10,000	+2; −3	+2; −4	+5; −∞	+6; −∞
12,500	+2; −3	+3; −6	+5; −∞	+6; −∞
16,000	+2; −3	+3; −∞	+5; −∞	+6; −∞
16,000	+2; −3	+3; −∞	+5; −∞	+6; −∞

The sound level meter can be functionally divided into four parts: microphone and preamplifier, A-weighting filter, rms detector, and display (Figure 48.9). The microphone should ensure the correct measurement of the sound pressure within the frequency range for the given class. Also, the standard gives requirements for the directionality of the microphone. The frequency response of the instrument, including the weighting filter, is given for sound waves arriving at the microphone along the reference direction. For sound waves arriving from other directions, the standard allows wider tolerances at higher frequencies, taking into account the inevitable reflections and diffraction occurring at higher frequencies.

The preamplifier converts the high-impedance output signal from the microphone to a low-impedance signal, but has in itself no or even negative voltage amplification. The signal from the preamplifier is then passed through an A-weighting filter. This is a standardized filter that, in principle, resembles the sensitivity of the human ear, so that a measurement utilizing this filter will give a result that correlates with



FIGURE 48.8 Modern type 1 sound level meter with built-in frequency analyzer.

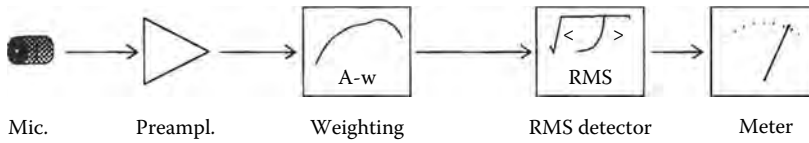


FIGURE 48.9 Functional parts of a sound level meter.

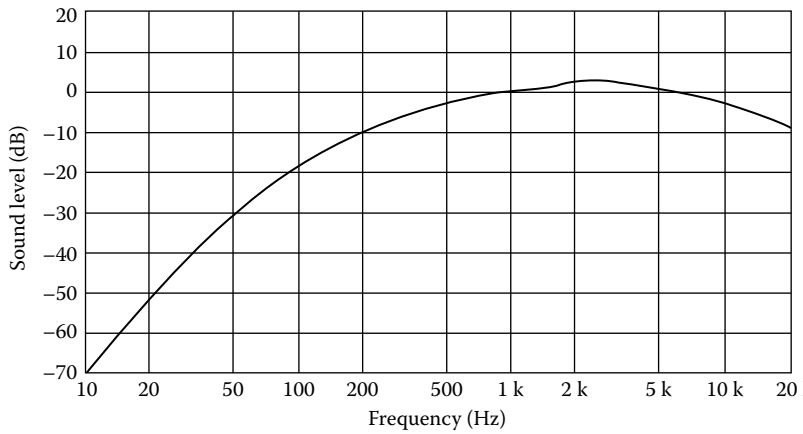


FIGURE 48.10 A-weighting curve in dB vs. Hz.

the subjective response of an average listener. The filter, with the filter characteristic as in Figure 48.10, attenuates low and high frequencies and slightly amplifies frequencies in the mid-frequency range from 1 to 5 kHz. There are a number of other weighting curves, denoted B-weighting, C-weighting, and D-weighting, which may give better correlation with subjective responses in special cases, such as for very high or very low levels or for aircraft noise.

The signal from the A-weighting filter is subsequently passed through an exponential rms detector, with a time constant of either 125 ms ("fast") or 1 s ("slow"). These time constants simulate the behavior of the human ear when subjected to time-varying signals. Especially when the duration of the sound stimuli to the human ear becomes short (e.g., around 200 ms), the sound is subjectively judged as being lower compared to the same sound heard continuously. The same effect is obtained using the "fast" averaging time. This will however give a higher statistical uncertainty on the level estimate than when using the time constant "slow," so this should be chosen if the sound signal is continuously. Other standards describe special sound level meters such as integrating sound level meters or impulse sound level meters intended for special purposes.

48.6 Frequency Analyzers

While the sound level meter gives a single reading for the sound level in the frequency range from 20 Hz to 20 kHz, it is often desirable to have more detailed information about the frequency content of the signal. Two types of frequency analyzers are commonly used in acoustic measurements: FFT analyzers and real-time filter analyzers. The *FFT analyzers*, with very-high-frequency resolutions, can give a wealth of frequency information and make it possible to separate closely space harmonics (e.g., from a gearbox). In contrast to this, the *real-time filter analyzers* (Figure 48.11) use a much broader frequency resolution, usually in one-third-octave bands. The frequency analysis is performed by a bank of filters (nowadays mostly digital filters) with well-defined frequency and time responses. The filter responses have been internationally standardized, with the center frequencies and passbands as shown in Table 48.2. In the frequency range from 20 Hz to 20 kHz, the one-third-octave filter bank consists of 31 filters, simultaneously measuring the input signal. The resulting one-third-octave spectrum resembles the subjective response of the human ear.

48.7 Pressure-Based Measurements

The result of sound pressure measurements will be influenced by many factors: source, source operating conditions, surroundings, measurement position, etc. Depending on the goal of the measurement, these parameters can be controlled in different manners. If the goal of the measurement is to quantify the noise exposure to an operator's ear in a noisy environment, it is important that the microphone is in the same position as the operator's ear would normally be in and that the environment is equal to the normal



FIGURE 48.11 Real-time frequency analyzer for acoustic measurements.

TABLE 48.2 One-Third-Octave Analysis Frequencies

Nominal Center Frequency (Hz)	Exact Center Frequency (Hz)	Passband (Hz)
20	19.95	17.8–22.4
25	25.12	22.4–28.2
31.5	31.62	28.2–35.5
40	39.81	35.5–44.7
50	50.12	44.7–56.2
63	63.1	56.2–70.8
80	79.43	70.8–89.1
100	100.0	89.1–112
125	125.89	112–141
160	158.49	141–178
200	199.53	178–224
250	251.19	224–282
315	316.23	282–355
400	398.11	355–447
500	501.19	447–562
630	630.96	562–708
800	794.33	708–891
1,000	1,000.0	891–1,120
1,250	1,258.9	1,120–1,410
1,600	1,584.9	1,410–1,780
2,000	1,995.3	1,780–2,240
2,500	2,511.9	2,240–2,820
3,150	3,162.3	2,820–3,550
4,000	3,981.1	3,550–4,470
5,000	5,011.9	4,470–5,620
6,300	6,309.6	5,620–7,080
8,000	7,943.3	7,080–8,910
10,000	10,000.0	8,910–11,200
12,500	12,589.3	11,200–14,100
16,000	15,848.9	14,100–17,800
20,000	19,952.6	17,800–22,400

operating environment. If, on the other hand, the task is to describe the sound source as a noise-emitting machine, it is important to minimize the influence of the environment on the measurement result.

If the aim is to describe the measuring object as a noise source, it is customary to state the radiated sound power for the source. This is a global parameter quantifying the total noise radiation from the source and to a certain extent independent of the environment. The sound power can be measured in a number of different ways: in a free field, in a reverberation room, using a substitution technique, or using sound intensity technique. A free field is a sound field in which the sound is radiated freely in all directions, with no restricting walls or reflections. This is most often obtained in a semi-anechoic chamber, where all walls and ceiling have been covered by nearly 100% absorptive material, with only the floor made of reflecting material. When the sound source is placed in the semi-anechoic chamber, the emitted sound waves will radiate freely away from the source and, in the far field, the waves can be considered to be plane waves or spherical waves. Therefore, the sound intensity can be calculated from pressure measurements using Equation 48.7.

As real sound sources seldom radiate equally in all directions, a number of measurements around the test object are averaged. ISO Standard 3745 "Acoustics—Determination of sound power levels of noise sources—Precision method for anechoic and semi-anechoic rooms" [3] specifies an array of microphone

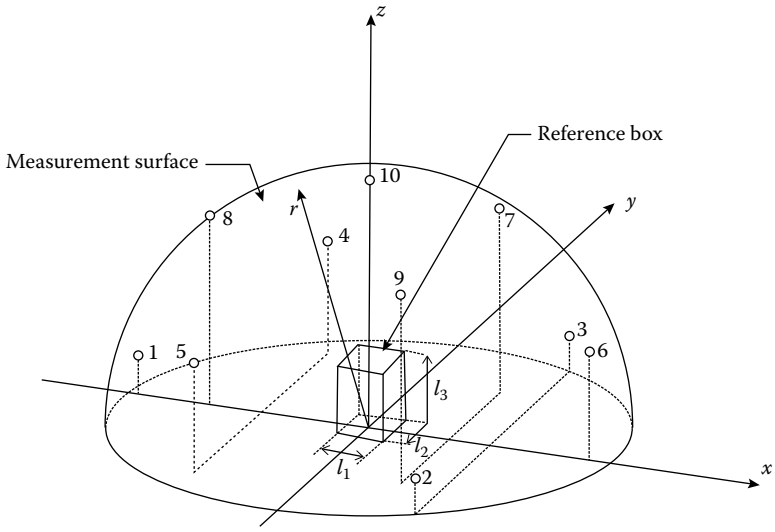


FIGURE 48.12 Measurement points for sound power determination.

TABLE 48.3 Coordinates of Measurement Points for Hemisphere with Radius *r*

Measurement Point No.	*x*/*r*	*y*/*r*	*z*/*r*
1	−0.99	0	0.15
2	0.5	−0.86	0.15
3	0.5	0.86	0.15
4	−0.45	0.77	0.45
5	−0.45	−0.77	0.45
6	0.89	0	0.45
7	0.33	0.57	0.75
8	−0.66	0	0.75
9	0.33	−0.57	0.75
10	0	0	1.0

positions on a hemisphere over the test object, as in Figure 48.12, with the coordinates as in Table 48.3. As all points are associated with the same area and as the sound power is intensity times the area, the total radiated sound power can be calculated as

$$P = A\sum I_n = \frac{2\pi r}{\rho c}\sum p_n^2 \tag{48.14}$$

where

A is the area of the test hemisphere with radius r

p_n is the pressure measured in point number n

48.8 Sound Intensity Measurements

The calculation in Equation 48.14 of the sound power from sound pressure measurements is based on Equation 48.7. This equation, which gives the intensity based on a pressure measurement, is however only valid in a free field, in the direction of propagation. In general, in the presence of background noise



FIGURE 48.13 Two-microphone sound intensity probe.

or with reflections (e.g., from walls), it is not possible to calculate the sound intensity from a single pressure measurement. In these cases, it is, however, possible to measure directly the sound intensity with a two-microphone intensity probe, Figure 48.13.

Sound intensity I is the product of the pressure and the particle velocity:

$$I = pv \tag{48.15}$$

While the pressure p is a scalar and independent of the direction, the particle velocity is a vector quantity and is directionally dependent. When the particle velocity is stated as in Equation 48.15, it is implicit that the velocity is in a certain direction and that the resulting intensity is calculated in the same direction. For example, the particle velocity v in the direction of propagation (Figure 48.14a) gives the intensity radiation away from the point source, while the particle velocity perpendicular to the propagation direction (Figure 48.14b) is zero. The intensity calculated from Equation 48.15 will therefore be zero in the direction perpendicular to the propagation direction even though the sound pressure is the same. This means that the sound energy flows away radially from the point source and no energy is flowing tangentially.

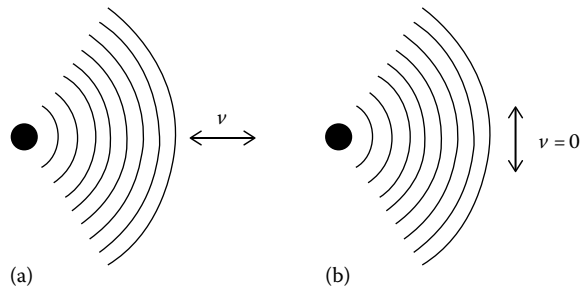


FIGURE 48.14 Particle velocity (a) along direction of propagation and (b) perpendicular to direction of propagation.

The measurement of the sound intensity according to Equation 48.15 requires the measurement of the sound pressure and the particle velocity. With the two-microphone intensity probe, the pressure in a position in between the two microphones is calculated as the mean pressure measured by the two microphones:

$$p = \frac{p_1 + p_2}{2} \tag{48.16}$$

The air particle velocity v, in the direction of the intensity probe, can be calculated from the pressure differences between the two microphone measurements:

$$v = \int \frac{(p_2 - p_1)}{\rho \Delta r} \partial \tau \tag{48.17}$$

where

ρ is the density of the air

Δr is the distance between the microphones

The intensity I is then obtained by multiplying the pressure and the velocity:

$$I = pv = \frac{p_1 + p_2}{2} \int \frac{(p_2 - p_1)}{\rho \Delta r} \partial \tau \tag{48.18}$$

The intensity measurement technique is a powerful tool to localize acoustic noise source and to determine the sound power radiated from a sound source, even in the presence of other strong sound sources.

48.9 Near-Field Acoustic Holography Measurements

The term *acoustic holography* comes from the analogy to optical holographs. It is well known how holography, as opposed to a normal photo, enables one to reconstruct the full image of an object. This is obtained by "recording" information about both the magnitude and the phase of the light, while a normal photo only "records" the magnitude of the light. Similarly, with *acoustic holography*, both the magnitude and the phase of the sound field are measured over a plane surface. These measurements result in a complete description of the sound field where both magnitude and phase are known at all points. It is then possible to calculate acoustic quantities, including sound intensity distribution, particle velocity, sound power, and radiation pattern.

The basic assumption behind *near-field acoustic holography* (NAH) is that the sound field can be decomposed into two simple wave types: plane waves and evanescent waves. The *plane waves* describe the part of the sound field that is propagated away from the near field toward the far field, and the *evanescent waves* describe the complicated sound field existing in the near field. Any sound field can be described as a combination of plane waves and evanescent waves with different magnitude and directions. The magnitude and direction of the individual waves can be described by their spatial frequencies or wave numbers. For a simple plane wave propagating in a certain direction, this can be described in terms of its temporal frequency as well as by its spatial frequency. The temporal frequency (Figure 48.15a), is obtained by looking at the pressure changes with time at a certain point in the sound field. This gives the temporal frequency in hertz or radians per second. Similarly, the spatial frequency (Figure 48.15b) is obtained by looking at the pressure changes at a certain time. At that instant in time, the pressure will be different in different positions in space. If one moves in a certain direction in space, one will see a certain change in the pressure, corresponding to a spatial frequency, measured with the unit cycles per meter or radians per meter. As the temporal frequency gives information about how often the pressure changes with time at a certain point, the spatial frequency gives information about how often the

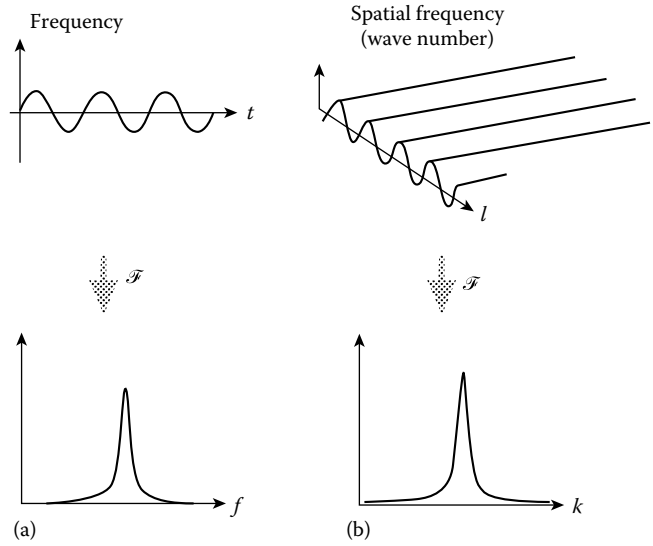


FIGURE 48.15 Spatial and temporal frequency characteristics of plane waves: (a) temporal and (b) spatial.

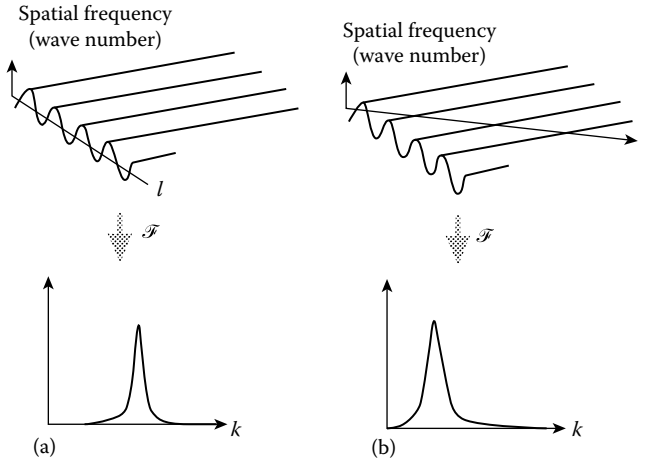


FIGURE 48.16 Spatial frequencies of waves propagating in different directions: (a) measured spatial frequency in the direction of propagation and (b) measured spatial frequency in different direction of propagation.

pressure changes with position at a certain time. In the example of Figure 48.15b, the propagation direction of the plane wave was identical to the direction of the axis along which the spatial frequency was measured. In this case, shown again in Figure 48.16a, the relationship between the spatial frequency k_0 (i.e., the wave number) and the temporal frequency f is given by the speed of sound c:

$$k_0 = \frac{2\pi f}{c} = \frac{\omega}{c} = \frac{2\pi}{\lambda} \tag{48.19}$$

where λ is the wavelength. If, however, the axis along which the spatial frequency is measured is not the same as the propagation direction (see the example in Figure 48.16b), this simple relationship is not valid. In this case, although the temporal frequency is the same as in Figure 48.16a, the spatial frequency is lower. For one particular temporal frequency, the spatial frequencies will thus give information about

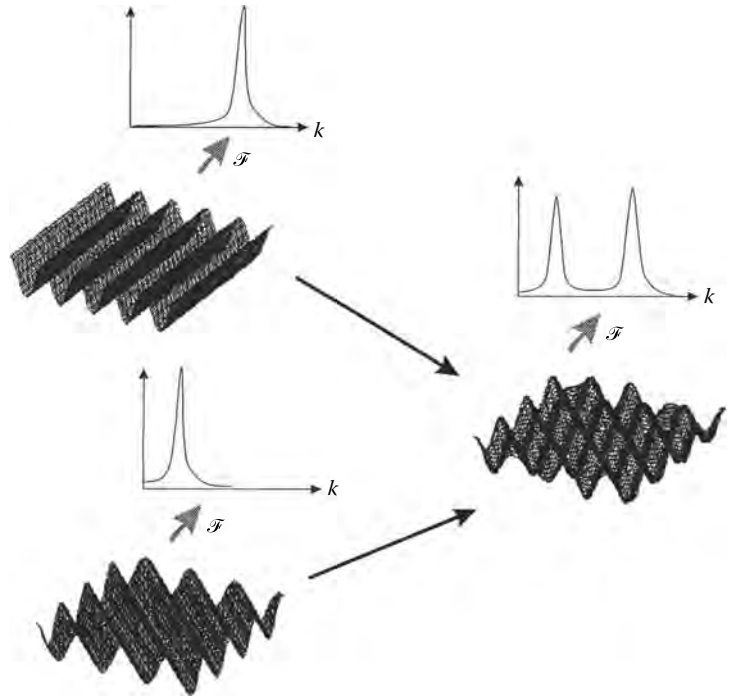


FIGURE 48.17 Spatial spectrum of two waves propagating in different directions.

the propagation directions. Therefore, if the sound field is made up of several plane waves with the same temporal frequency, but with different propagation directions, this will be shown in the spatial spectrum as several spatial frequency components. If, for example, the sound field is a sum of two waves (Figure 48.17) where one wave is traveling along the axis of measurement and the other at an angle of 45° relative to the first wave, the spatial spectrum will contain two spatial frequencies. One spatial frequency will be k, corresponding to a wave in the direction along the axis, and the other frequency will be k^* cos (45°). Thus far, the spatial frequencies have been defined along a single axis corresponding to a 1-D Fourier transformation. In the NAH technique, the sound field is sampled not only along a single axis, but over a plane. Therefore, a 2-D Fourier transformation is used instead. This gives as a result a 2-D spatial frequency spectrum, but otherwise the information is the same as before, namely, information about the direction and magnitude of the simple wave types.

The sound field from a point source cannot be explained by simple plane waves such as those in Figures 48.16 and 48.17, as the amplitude decreases with the distance from the origin. The plane waves retain the same magnitude over the full plane. Thus, to describe the near-field phenomenon, one must introduce evanescent waves. In the 1-D Fourier spectrum, the evanescent waves can be identified as spatial frequencies higher than $k_0 = 2\pi f/c$. Similarly, in the 2-D spatial frequency spectrum, the evanescent waves can be identified as having spatial frequencies or wave numbers higher than k_0.

The individual spatial frequencies in the 2-D spatial frequency spectrum correspond to simple plane waves or evanescent waves in the scan plane (i.e., the measurement plane). For each of these simple wave types, it is easy to calculate the pressure in other planes, see Figure 48.18. For the plane waves, a simple phase shift of the wave is required to calculate the result in a new plane. For the evanescent waves, the changes in amplitude must be taken into account, but in principle, this is also a simple transfer function applied to the 2-D spatial frequency spectrum. In this way, the 2-D spatial frequency spectrum in a new plane can be calculated from the original data by applying simple transfer function operations. The new 2-D spatial frequency spectrum is then an inverse Fourier transform (in two dimensions) to get the sound field in the new plane.

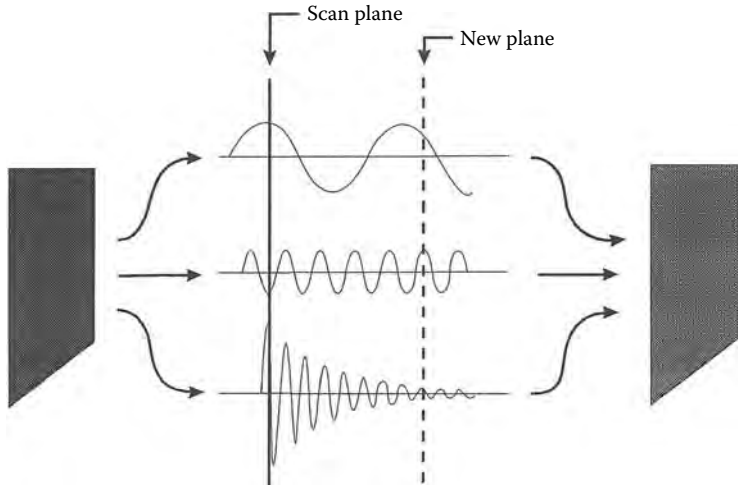


FIGURE 48.18 Transformation of simple wave types in a spatial spectrum, from one measurement plane to another plane.

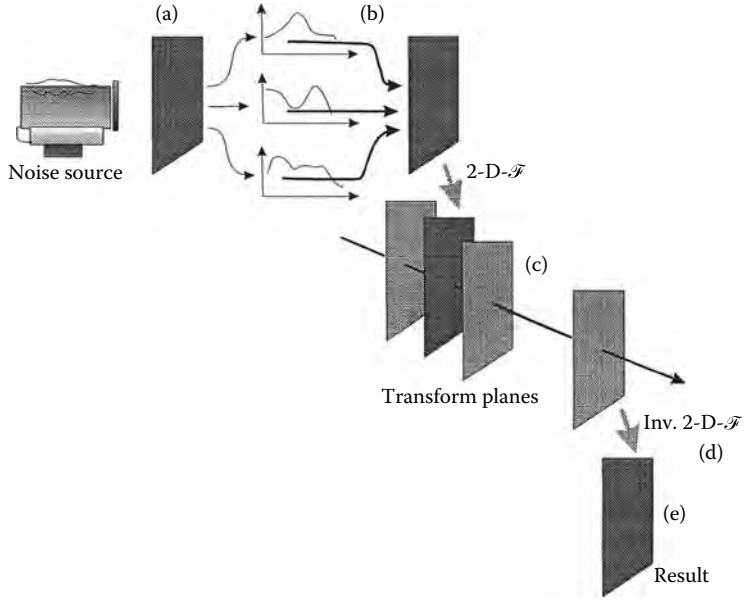


FIGURE 48.19 Overall principle of NAH: (a) measurement of cross spectra in the scan plane, (b) calculation for one temporal frequency at a time; (c) 2-D spatial Fourier transformation, (d) transformation of simple wave types, and (e) inverse 2-D transformation.

The overall principle of NAH can be simplified as in Figure 48.19. The sound field is scanned in a plane close to the measuring object. This gives an array of temporal spectra, one for each scan position. Looking at one temporal frequency at a time, one takes out the information from each of the spectra corresponding to the actual frequency of interest. This generates a new array with information about only one temporal frequency. A Fourier transform (in two dimensions) is then applied to the array to generate a 2-D spatial frequency spectrum. This can then be transformed to new planes using simple transfer function operations. When the 2-D spatial frequency spectrum in the new plane has been calculated, an inverse Fourier transform is used to obtain the new pressure distribution

in the new plane. In principle, the NAH technique requires that all cross spectra between all the scan positions are given; that is, in each of the scan positions, all the cross spectra to all other scan positions must be determined. A simple scan of a sound field with 2,540 scan positions, defining $N = 1{,}000$ scan positions, would result in $\frac{1}{2}N(N + 1) = 500{,}500$ cross spectra. Instead of measuring all these cross spectra, the system uses a set of reference transducers to reduce the amount of cross spectra. The number of necessary reference transducers to give a complete description of the sound field without measuring the full amount of data is determined by the complexity of the sound field. A measurement with, for example, four reference transducers and 2540 scan positions will then be reduced to $4N = 4000$ cross spectrum measurements.

48.10 Calibration

In order to make accurate and reliable measurements, the microphone and connected instruments must be properly calibrated. The calibration of measurement microphones can be divided into two parts: a level calibration and a frequency response calibration. The *level calibration* establishes the output signal of the microphone for a given acoustic input signal at a given frequency, while the *frequency response* gives the output at other frequencies relative to the level calibration frequency. The level calibration can be performed by a number of different methods with different accuracies.

The most accurate method is the *reciprocity calibration method*. This method utilizes the fact that a condenser microphone is a reciprocal transducer; that is, it can be used as a microphone (to convert an acoustic signal to a voltage signal) and as a loudspeaker (to convert a voltage signal into an acoustic signal). By measuring the relationship between three test microphones driven as both transmitters and receivers, one obtains a set of three equations with the three microphone sensitivities as the unknowns. By solving these three equations, one obtains the sensitivity of the three microphones. The reciprocity calibration method is very accurate but rather tedious and requires well-controlled environmental conditions and is therefore seldom used in practical situations.

The comparison or substitution methods are essentially identical in that they are based on measuring the differences between the test microphone and a reference microphone with known sensitivity. In this case, the reference microphone is often calibrated at an accredited national acoustic laboratory like NIST, NPL, or PTB, whereby the traceability is ensured. In the substitution method, the acoustic output of a sound source is measured with the reference microphone. Afterward, the reference microphone is replaced with a test microphone and the output is measured again. Provided that the sound source has been stable, the sensitivity of the test microphone can then be calculated. The comparison method is similar to the substitution method, except that the reference microphone and the test microphone are subjected to the same sound pressure simultaneously and therefore the requirements to the sound source stability are less important.

An often-used method for microphone calibration is the *pistonphone method*. The pistonphone, Figure 48.20, is a very stable sound source, which produces a well-defined sound pressure level inside a closed coupler. It works by volume displacements, Figure 48.21, with a well-defined velocity, usually at 250 Hz. As the piston is moving in and out, the volume of the closed coupler is changed and this will result in pressure variations. The actual pressure level obtained in the pistonphone depends on the volume of the coupler, the volume displacement of the pistons, the barometric pressure, and—to a lesser degree—on other factors such as humidity and heat dissipation. As the pistonphone is based on a relatively simple mechanical system, it is very reliable and easy to use in practice, with an accuracy around 0.1 dB. Also, the pistonphone is often used as the stable sound source for calibrations using comparisons or substitution methods.

A *sound pressure calibrator* is basically a small self-contained comparison calibration device. The test microphone is inserted into a small, closed volume and a small loudspeaker produces a single



FIGURE 48.20 Pistonphone for microphone calibration.

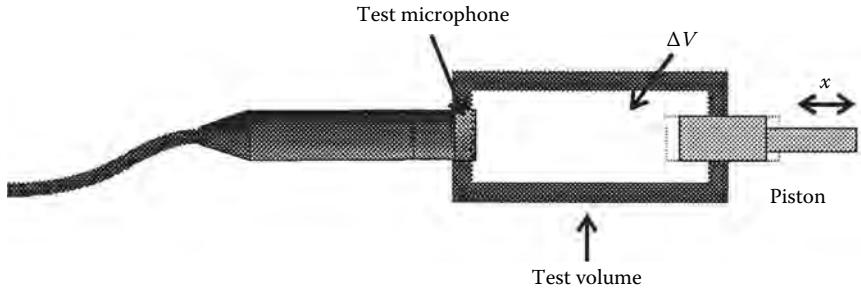


FIGURE 48.21 Principle of a pistonphone.

frequency signal, usually at 1 kHz. The output level of the loudspeaker is controlled in a feedback system with a signal from a reference microphone. Provided that the reference microphone and the feedback gain are stable, the sound level at the test microphone will be well-defined and the sensitivity can be determined. The sound level calibrators are normally not used to make accurate microphone calibrations, but rather to make field checks of the integrity of a complete measurement system.

The frequency response of a microphone is most often determined by the *electrostatic actuator method*. A conducting grid is placed close to and parallel to the microphone diaphragm. An electric field is established between the actuator and the diaphragm by applying 800 V dc to the actuator. A test signal of 50–150 V ac is superimposed on the dc signal, and the electrostatic forces will push and pull the diaphragm, similar to a sound pressure of 1–10 Pa. By sweeping the test signal through the frequencies of interest, the pressure response of the test microphone can be recorded. The electrostatic actuator technique is widely used as a convenient and accurate test method, both during production and final calibration of measurement microphones.

Available instrumentation and manufacturers are given in Tables 48.4 and 48.5.

TABLE 48.4 List of Instrumentation Types, Prices, and Manufacturers

Instrumentation	Types Available	Approximate Price	Manufacturers
Meas. microphones	½ in. free field ½ in. pressure ¼ in. free field ¼ in. pressure	$750–$825	G.R.A.S. Sound & Vibration ACO Pacific B&K The Modal Shop Larson Davies
Preamplifiers	½ and ¼ in.	$600–$850	G.R.A.S. Sound & Vibration ACO Pacific B&K The Modal Shop Larson Davies
Sound level meters	Simple type 1 SLM	$800–$2,000	Rion CEL B&K
	Advanced SLM with freq. analysis and data storage	$2,000–$10,000	Rion CEL Larson Davies B&K
Frequency analyzers	Real-time frequency analyzers/FFT	$5,000–$50,000	Hewlett-Packard Norsonic Data Physics 01 dB
NAH	Complete system with 16–64 channel acquisition and post-processing	$100,000–$200,000	LMS B&K

TABLE 48.5 Companies That Make Acoustic Measurement Instruments

G.R.A.S. Sound & Vibration, Skelstedet 10B, 2950 Vedbaek, Denmark, Tel.: +45 45 66 40 46, http://www.gras.dk/

LMS International, Interleuvenlaan 68, B-3001 Leuven, Belgium, Tel.: +32 16 384 571, http://www.lmsintl.com/

Hewlett-Packard Co., P.O. Box 95052-8059, Santa Clara, CA 95052, Tel.: (206) 335 2000, http://www8.hp.com

Norsonic AS, P.O. Box 24, N-3420 Lierskogen, Norway, Tel.: +47 32 85 20 80, http://www.norsonic.com/

The Modal Shop Inc., 1776 Mentor Avenue, Suite 170, Cincinnati, OH 45212-3521, Tel.: (513) 351 9919, http://www.modalshop.com/

01 dB, 111 rue du 1er Mars, F69100 Villeurbanne, France, Tel.: + 33 4 78 53 96 96

Larson Davies, Inc., 1681 West 820 North, Provo, UT 84601, Tel.: (801) 375 0177, http://www.larsondavis.com/

Brüel & Kjær. Spectris Technologies, Inc., 2364 Park Central Blvd., Decatur, GA 30035-3987, Tel.: (800) 332 2040, http://bksv.com/

Rion, Scantek, Inc., 916 Gist Avenue, Silver Springs, MD 20910, Tel.: (301) 495 7738, http://www.scantekinc.com/

ACO Pacific, Inc., 2604 Read Avenue, Belmont, CA 94002, Tel.: (415) 595 8588, http://acopacific.com/

CEL Instruments, 1 Westchester Drive, Milford, NH 03055, Tel.: (800) 366 2966

References

1. E. Skudrzyk, *The Foundation of Acoustics*, New York: Springer-Verlag, 1971.
2. International Electrotechnical Commission, *Publication 651: Sound Level Meters*, Genève, Switzerland: IEC, 1971.
3. International Organization for Standardization, Standard 3745, Acoustics—Determination of sound power levels of noise sources—Precision method for anechoic and semi-anechoic rooms, Genève, Switzerland, ISO, 1981.

Further Readings

Fraden, J., *Handbook of Modern Sensors*, Berlin, Germany: Springer, 2010.
Iniewski, K., *Optical, Acoustic, Magnetic, and Mechanical Sensor Technologies*, Boca Raton, FL: CRC Press, 2012.
Morgan, D., *Surface Acoustic Wave Filters*, Oxford, U.K.: Elsevier, 2007.
Xiao, Y., *Underwater Acoustic Sensor Networks*, Boca Raton, FL: CRC Press, 2010.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

49

Ultrasound Measurement

49.1 Applications of Ultrasound **49**-1
Medical • Industrial • Underwater
49.2 Basic Ultrasound Parameters **49**-2
49.3 Conceptual Description of Ultrasound Imaging and Measurements **49**-4
49.4 Single-Element and Array Transducers **49**-6
49.5 Selection Criteria for Ultrasound Frequencies **49**-7
49.6 Basic Parameters in Ultrasound Measurements **49**-7
Reflection (or Transmission) Detected or Not • Travel Time • Attenuation • Reflection Coefficient • Parameters Obtained through Signal Processing
49.7 Ultrasound Theory and Applications **49**-8
Sound Velocity • Wave Propagation in Homogeneous Media • Acoustic Intensity and Sound Levels • Wave Propagation across Layers and Boundaries • Attenuation: Its Origin, Measurement, and Applications • CW Fields from Planar Piston Sources • Generation of Ultrasound: Piezoelectric and Magnetostrictive Phenomena • Display Formats for Pulse-Echo Measurements • Flow Measurements by Doppler Signal Processing
49.8 Review of Common Applications of Ultrasound and Their Instrumentation **49**-21
Range Measurements, Air • Thickness Measurement for Testing and Process Control • Detection of Defects, Such as Flaws, Voids, and Debonds • Doppler Flow Measurements • Upstream/Downstream Volume Flow Measurements • Elastic Properties of Solids • Porosity and Grain Size Estimation • Acoustic Microscopy • Medical Ultrasound • Selected Manufacturers of Ultrasound Products
49.9 Advanced Topics in Ultrasound **49**-24
Overview of Diffraction • Fresnel and Fraunhofer Diffraction • Pressure Function at a Given Field Point, Based on Rayleigh Integral
References **49**-26
Further Information **49**-27

Peder C. Pedersen
Worcester Polytechnic Institute

49.1 Applications of Ultrasound

49.1.1 Medical

Ultrasound has a broad range of applications in medicine, where it is referred to as *medical ultrasound*. It is widely used in obstetrics to follow the development of the fetus during pregnancy, in cardiology where images can display the dynamics of blood flow and the motion of tissue structures (referred to as

real-time imaging), and for locating tumors and cysts. 3-D imaging, surgical applications, imaging from within arteries (intravascular ultrasound), and contrast imaging are among the newer developments.

49.1.2 Industrial

In industry, ultrasound is utilized for examining critical structures, such as pipes and aircraft fuselages, for cracks and fatigue. Manufactured parts can likewise be examined for voids, flaws, and inclusions. Ultrasound has also widespread use in process control. The applications are collectively called *nondestructive testing* (NDT) or *nondestructive evaluation* (NDE). In addition, *acoustic microscopy* refers to microscopic examinations of internal structures that cannot be studied with a light microscope, such as an integrated circuit or biological tissue.

49.1.3 Underwater

Ultrasound is likewise an important tool for locating structures in the ocean, such as wrecks, mines, submarines, or schools of fish; the term SONAR (sound navigation and ranging) is applied to these applications.

There are many other usages of ultrasound that lie outside the scope of this handbook: ultrasound welding, ultrasound cleaning, ultrasound hyperthermia, and ultrasound destruction of kidney stones (lithotripsy).

49.2 Basic Ultrasound Parameters

Ultrasound refers to acoustic waves of frequencies higher than 20,000 cycles per second (20 kHz), equal to the assumed upper limit for sound frequencies detectable by the human ear. As acoustic waves fundamentally are mechanical vibrations, a medium (e.g., water, air, or steel) is required for the waves to travel, or propagate, in. Hence, acoustic waves cannot exist in vacuum, such as outer space. If a single-frequency sound wave is produced, also termed a *continuous wave* (CW), the fundamental relationship between frequency, f, in Hz, the sound speed of the medium, c_0, in m s^{-1}, and the wavelength, λ, in meters is given as

$$\lambda = \frac{c_0}{f} \tag{49.1}$$

The wavelength λ describes the length, in the direction of propagation, of one period of the sound wave. The wavelength determines, or influences, the behavior of many acoustic functions: The sound field emitted from an acoustic radiator (e.g., a transducer or loudspeaker) is determined by the radiator's size measured in wavelengths; the ability to differentiate between closely spaced reflectors is a function of the separation measured in wavelengths. Even when a sound pulse, rather than a CW sound, is transmitted, the wavelength concept is still useful, as the pulse typically contains a dominant frequency.

The vibrational activity on the surface of the sound source transfers the acoustic energy into the medium. If one were able to observe a very small volume, referred to as a *particle*, of the medium during transmission of sound energy, one would see the particle moving back and forth around a fixed position. Associated with the particle motion is an acoustic pressure, which refers to the pressure variation around the mean pressure (which is typically the atmospheric pressure). This allows the introduction of two important—and closely related—acoustic quantities: the *particle velocity*, $\vec{\mathbf{u}}(\vec{\mathbf{r}}, \mathrm{t})$, and the *acoustic pressure*, $p(\vec{\mathbf{r}}, t)$. In this notation, the arrow above a symbol in bold indicates a vector. The symbol $\vec{\mathbf{r}}$ represents the position vector, which simply defines a specific location in space. Thus, both particle velocity and pressure are functions of three spatial variables, x, y, and z, and the time variable, t.

To characterize a medium acoustically, the most important parameter is the *specific acoustic impedance, z*. For a lossless medium, z is given as follows:

$$z = \rho_0 c_0 \tag{49.2}$$

In Equation 49.2, ρ_0 is the density of the medium, measured in kg m^{-3}. When a medium absorbs acoustic energy (which all media do to a greater or smaller extent), the expression for acoustic impedance also contains a small imaginary term; this will be ignored in the discussions presented in this chapter. The acoustic impedance relates the particle velocity to the acoustic pressure:

$$z = \frac{p(\vec{\mathbf{r}}, t)}{u(\vec{\mathbf{r}}, t)} \tag{49.3}$$

Note that the relationship in Equation 49.3 uses the scalar value of the particle velocity (a scalar is a quantity, such as temperature, that does not have a direction associated with it). Equation 49.3 is exact for plane wave fields and a very good approximation for arbitrary acoustic fields. In a plane wave, all points in a plane normal to (i.e., which forms a 90° angle with respect to) the direction of propagation have the same pressure and particle velocity.

The acoustic impedances on either side of an interface (boundary between different media) determine the acoustic pressure reflected from the interface. Let a plane wave traveling in a medium with the acoustic impedance z_1 encounter a planar, smooth interface with another medium having the acoustic impedance z_2. Assume that the plane wave propagates directly toward the interface; this is commonly referred to as insonification under normal incidence. In this case, the pressure and the intensity reflection coefficients, R_p and R_i, respectively, are as follows:

$$R_p = \frac{p_r}{p_i} = \frac{z_2 - z_1}{z_2 + z_1} \tag{49.4}$$

$$R_i = \frac{I_r}{I_i} = \left(\frac{z_2 - z_1}{z_2 + z_1}\right)^2$$

Intensity is a measure of the mean power transmitted through unit area and is measured in watts per square meter. The corresponding pressure and intensity transmission coefficients, T_p and T_i, Ti respectively, are

$$T_p = \frac{p_t}{p_i} = \frac{2z_2}{z_2 + z_1} \tag{49.5}$$

$$T_i = \frac{I_t}{I_i} = \frac{4z_1 z_2}{(z_2 + z_1)^2}$$

The subscripts "i," "r," and "t" in Equations 49.4 and 49.5 refer to incident, reflected, and transmitted, respectively. The expressions in these two equations can be considered approximately valid for nonplanar waves under near-normal incidence. However, when the incident angle (angle between direction of propagation and the normal to the surface) becomes large, the reflection and transmission coefficients can change dramatically. In addition, the reflected and transmitted signals will also change if the surface is rough to the extent that the root mean square (rms) height exceeds a few percent of the wavelength.

The wave propagation can take several forms. In fluids and gases, only *longitudinal* (or *compressional*) waves exist, meaning that the direction of wave propagation is equal to the direction of the particle velocity vector. In solids, both longitudinal and *shear* waves exist that propagate in different directions and with different sound speeds. Transverse waves can exist on strings where the particle

motion is normal to the direction of propagation. (Strictly speaking, shear waves can propagate a short distance in liquids [fluids] if the viscosity is sufficiently high.)

49.3 Conceptual Description of Ultrasound Imaging and Measurements

Most ultrasound measurements are based on the generation of a short ultrasound pulse that propagates in a specified direction and is partly reflected wherever there is an abrupt change in the acoustic properties of the medium and detection of the resulting echoes (pulse-echo ultrasound). A change in properties can be due to a cyst in liver tissue, a crack in a high-pressure pipe, or a reflection from layers in the sea bottom. The degree to which a pulse is reflected at an interface is determined by the change in acoustic impedance as described in Equation 49.4. An image is formed by mapping echo strength vs. travel time (proportional to distance) and beam direction, as illustrated in Figure 49.1. This is referred to as *B-mode* imaging (brightness mode). Further signal processing can be applied to compensate for attenuation (the damping out of the pressure pulse as it propagates) of the medium or to control focusing. Signal processing can also be applied to analyze echoes for information about the structure of materials or about the surface characteristics of rough surfaces.

A block diagram of a simplified pulse-echo ultrasound measurement system is shown in Figure 49.2. The pulser circuit can generate a large voltage spike for exciting the transducer in B-mode applications, or the arbitrary function source can produce a short burst for Doppler measurements, or a coded

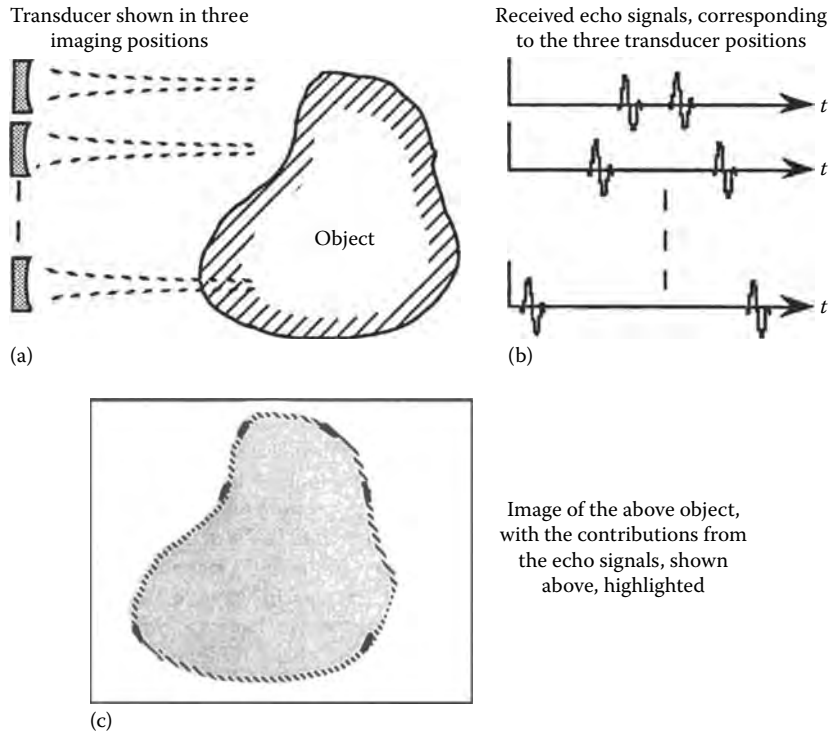


FIGURE 49.1 (a) A focused transducer insonifies the irregular object from different positions out of which only three are shown. The different transducer positions can readily be obtained by the use of an array transducer (to be described later). (b) Received echoes from the front and back of the object are displayed vs. travel time. It is here assumed that the structure is only weakly reflecting and that the attenuation has only a minimal effect. (c) An image is formed, based on the echo strengths and the echo arrival times.

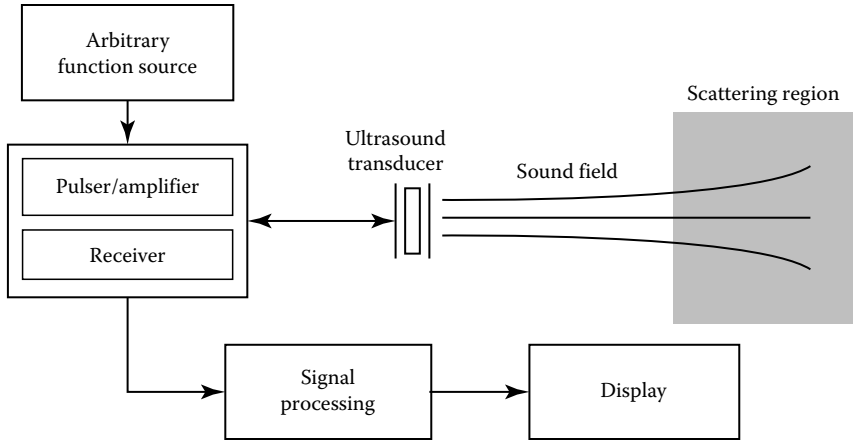


FIGURE 49.2 Simplified block diagram of a pulse-echo ultrasound system.

waveform, such as a linear sweep. The amplifier brings the driving voltage to a level where the transducer can generate an adequate amount of acoustic energy.

The transducer is made from a piezoelectric ceramic that has the property of producing mechanical vibrations in response to an applied voltage pulse and generating a voltage when subjected to mechanical stress. When an image is required, the transducer can be a mechanical sector probe that produces a fan-shaped image by means of a single, mechanically steered, focused transducer element. Alternatively, the transducer can be a linear array transducer (described later) that produces a rectangular image. In the case of an array transducer, the pulser/amplifier must contain a driving circuit for each element in the array, in addition to delay control. To achieve a short pulse and good sensitivity, the transducer is equipped with backing material and matching layers (to be discussed later).

The receiver block contains a low-noise amplifier with time-varying gain to correct for medium attenuation and often a circuit for logarithmic compression. In the case of array transducers, the receiver circuitry is a complex system of amplifiers, time-varying delay elements, and summing circuits. The signal-processing block can be partly analog and partly digital. The echo signals are envelope detected and digitized; the *envelope* of a signal is a curve that follows the amplitude of the received signal. A scan converter changes the signal into a format suitable for display on a grayscale monitor. Information about ultrasound pulser–receiver instrumentation for NDE measurements can be found in [1] and for medical imaging in [2].

Not all ultrasound measurement systems are based on the pulse-echo concept. For material characterization, *transmission measurements* are frequently used, as illustrated in Figure 49.3. The main difference between the pulse-echo system and the transmission system is that two transducers are used

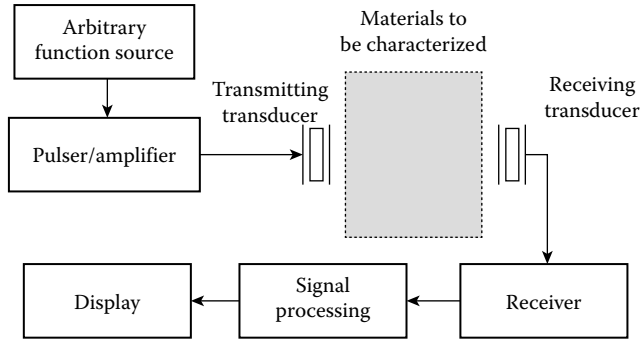


FIGURE 49.3 Simplified block diagram of an ultrasound transmission system.

in the transmission system; the description of the individual blocks for the pulse-echo system applies generally here also. Imaging is generally not possible, although tomographic imaging of either attenuation or velocity has been attempted. If the two transducers are moved together, voids and inclusions located in the transmission path can be detected.

Transmission measurements require, of course, that the structure or the medium of interest is accessible from opposite sides. This system allows accurate characterization of the attenuation and the sound speed of the medium from which a variety of material properties can be derived. Flow in a pipe or a channel can also be determined with a transmission system, by determining the difference between upstream and downstream propagation times.

49.4 Single-Element and Array Transducers

The device that converts electric energy into acoustic energy and vice versa is termed the *transducer* or, more specifically, the ultrasound transducer. Piezoelectric ceramics can be used for all frequencies in the ultrasound range; however, for ultrasound frequencies in the 20–200 kHz range, magnetostrictive devices are occasionally used, which are based on materials that exhibit mechanical deformation in response to an applied magnetic field. Ultrasound transducers are commercially available for many applications over a wide range of frequencies.

Single-element transducers are used for basic measurements and material characterization, while array transducers with many individual transducer elements are used for imaging purposes. The former type costs from a few hundreds of dollars and up, whereas the latter type costs in the thousands of dollars and requires extensive electronics for beam control and signal processing. A third alternative is the mechanical sector scanner where a single transducer is mechanically rotated over a specified angle. A single-element transducer consists of a piezoelectric element, a backing material that enables the transducer to respond to a fairly wide range of frequencies (but with reduced sensitivity), and a matching layer and faceplate that provide improved coupling into the medium and protect the transducer.

Although broadband transducers are very desirable, the energy conversion characteristics of the practical transducer correspond to that of a bandpass filter, that is, the ultrasound pulse has most of its energy distributed around one frequency. This is the frequency referred to when, for example, one orders a 3.5 MHz transducer. The 6 dB bandwidth can be from 50% to 100% of the transducer frequency; the 6 dB bandwidth is the frequency range over which the transducer can produce an acoustic pressure that is at least 50% of the acoustic pressure at the most efficient frequency. The radiation characteristics of transducers are determined by their geometry (square, circular, plane, focused, etc.) and dimensions measured in wavelengths. Hence, a 10 MHz transducer tends to be smaller than a 5 MHz transducer. Focusing can also be achieved by means of an acoustic lens.

Array transducers exist in three main categories: phased arrays, linear arrays, and annular arrays. A fourth category could exist commercially in a few years: the 2-D array or a sparse 2-D array. Common for these array transducers is the fact that *during transmission*, the excitation time and excitation signal amplitude for each transducer element are controlled independently. This allows the beam to be steered in a given direction, as well as focused at a given point in space. *During reception*, an independently controlled time delay—which may be time varying—can be applied to each element before summation with the signal from other elements. The delay control permits the transducer to have maximal sensitivity to an echo from a specified range and, moreover, to shift this point in space away from the transducer as echoes from structures further away are received. A limitation of the phased and linear arrays is that beam steering and focusing can only take place within the image plane. In the direction normal to the image plane, a fixed focus is produced by the physical shape of the array elements.

The phased array transducer consists of a fairly small number of elements, such as 16–32. Both beam steering and focusing is carried out, producing a pie-shaped image. The phased array is most suitable when only a narrow observation window exists, as is the case when imaging the heart from the space between the ribs. The linear array transducer has far more elements, typically between 128 and 256.

It activates only a subset of all the elements for a given measurement and produces focusing, but generally not beam steering, so that the beam direction is normal to the array surface. Consider a linear array transducer with 128 elements, labeled 1–128. If one assumes that the first pulse-echo measurement is made with elements 1–16, the next measurement with elements 2–17, etc., the beam will have been *linearly* translated along the long dimension of the linear array (hence the name), producing a rectangular image format. The annular array transducer consists of a series of concentric rings. As such, it cannot steer the beam, but can focus both in transmit and receive mode, just as the phased and linear arrays. Thus, the annular array transducer is not suitable for imaging unless the transducer is moved mechanically. Its specific advantage is uniform focusing.

49.5 Selection Criteria for Ultrasound Frequencies

The discussion so far has often mentioned *frequency* as an important variable, without providing any guidance as to what ultrasound frequencies should be used. There is generally no technological limitation with respect to choice of ultrasound frequency, as frequencies even in the gigahertz (GHz, 10^9 Hz) range can be produced. The higher the frequency, the better is the control over the direction and the width of the ultrasound beam, leading to improved axial and lateral resolution. However, the attenuation increases with higher frequency. For many media, the attenuation varies with the frequency squared, while for biological soft tissue, the attenuation varies nearly linearly with frequency.

Generally, the ultrasound frequency is chosen as high as possible while still allowing a satisfactory signal-to-noise ratio (SNR) of the received signal. This rule does not always hold; for example, for grain size estimation in metals and for rough surface characterization, the optimal frequency has a specific relationship to the mean grain dimension or the rms roughness.

49.6 Basic Parameters in Ultrasound Measurements

Ultrasound technology today has led to a very wide range of applications, and the following overview mentions only the basic parameters. For an in-depth review, see [3]. A discussion of applications is given at the end of Section 49.7.

49.6.1 Reflection (or Transmission) Detected or Not

The most basic measurement consists of determining whether a reflected or transmitted signal is received or not. This can be used to monitor liquid level where the absence of liquid prevents ultrasound transmission from taking place or to detect the existence of bubbles in a liquid (e.g., for dialysis purposes, where the bubbles are the reflectors) or for counting objects, such as bottles, on a conveyer belt.

49.6.2 Travel Time

Travel time is the elapsed time from transmission of an ultrasound pulse to the detection of a received pulse in either a pulse-echo or a transmission system. If the sound speed is known, the thickness of an object can be found from the time difference between the front surface echo and the back surface echo. Conversely, sound speed can be determined when the thickness is known. By measuring with a broadband pulse (a short pulse containing a broad range of frequencies), velocity dispersion (frequency dependence of velocity) can be found, which has applications for materials characterization. Elastic parameters can also be found from velocity measurements. Imaging applications are generally made based on the assumption of a known, constant velocity; thus, round-trip travel times from the transducer to the reflecting structures and back in a known direction give the basis for the image formation.

49.6.3 Attenuation

The attenuation and its frequency dependence are important materials parameters, as they can be used to differentiate between normal and pathological biological tissue, measure porosity of ceramics, the grain size of metals, and the structure of composite materials. Attenuation can be found by means of transmission measurements by determining the signal amplitude with the test object first absent and then present; here, transmission losses must also be considered, as quantified in Equation 49.5. Attenuation can be obtained from pulse-echo measurements, by comparing the strength of echoes from the front and back surfaces of the specimen under test. Diffraction effects (described later in Section 49.9) might need to be considered. Attenuation can be considered as a bulk parameter for the medium as a whole, or it can be determined for small regions of the medium.

49.6.4 Reflection Coefficient

The strength of a reflection at an interface between two media is determined by the change in acoustic impedance across the interface, as shown in Equation 49.4. This allows, at least in principle, the impedance of one medium to be determined from the measured reflection coefficient and the impedance of the other medium. In fact, the term *impediography* refers to the determination of the impedance profile of a layered medium by means of this concept. In practice, the measurement is not easy to carry out, as the transducer must be aligned very accurately at normal incidence to the medium surface. By measuring the reflection coefficient vs. incident angle at a planar, smooth surface, a velocity estimate of the reflecting medium can be made; in an alternative application, reflection measurements vs. frequency permit the evaluation of rough surface parameters.

49.6.5 Parameters Obtained through Signal Processing

A number of object parameters can be obtained from analysis of the received signals. Material properties can be extracted from speckle analysis (analysis of statistical fluctuations in the image) and other forms of statistical analysis. Recognition of objects with complex shapes can, in some cases, be done by extracting specific features of the received signals. Doppler processing for velocity estimation is another example of information that is only attainable by means of signal processing. By combining several "stacked" images, 3-D reconstruction is possible, from which volume estimation and rate of change of volume are possible, such as for determining the dynamics of the left ventricle of the heart.

49.7 Ultrasound Theory and Applications

49.7.1 Sound Velocity

The propagation speed, c_0, is generally determined by the density and the elastic properties of the medium. For an ideal gas, c_0 may be expressed as [4]

$$c_0 = \sqrt{\gamma \frac{P_0}{\rho_0}} = \sqrt{\gamma r T_K} \tag{49.6}$$

where

$\gamma = C_P/C_V$ is the ratio of specific heats and for air, $\gamma = 1.402$
P_0 is the static pressure of the gas, which at 1 atm is 1.013×10^5 N m^{-2}
ρ_0 is the specific density of the gas, equal to 1.293 kg m^{-3} for air at 1 atm and 0 °C
The constant r is the ratio of the universal gas constant and the molecular weight of the gas
T_K is the temperature in Kelvin

Substituting the values for air into the first equation in Equation 49.6 gives $c_0 = 331.6$ m s^{-1} at 0 °C. As the ratio P_0/ρ_0 is constant for varying pressure but constant temperature, c_0 is likewise pressure independent. (For a real gas, c_0 exhibits, in fact, a small dependence on pressure.)

The temperature dependence of an ideal gas can be obtained by rewriting the second expression in Equation 49.6 as follows:

$$c_0 = c_{\text{ref}}\sqrt{1+\frac{T}{273}} \tag{49.7}$$

where

c_{ref} is the sound speed at 0 °C
T is the temperature in °C

In liquids, the expression for c_0 is given as

$$c_0 = \sqrt{\frac{K_s}{\rho_0}} \tag{49.8}$$

where K_s is the adiabatic bulk modulus. Expressions for actual liquids as a function of temperature and pressure are not easily derived and are often empirically determined. For pure water, the sound speed as a function of temperature and pressure can be given as

$$c_{H_2O} = 1402.4 + 5.01T - 0.055T^2 + 0.00022T^3 + 1.6\times10^{-6}P_0 \tag{49.9}$$

where

T is the temperature in °C
P_0 is the static pressure in N m^{-2}

The expression is valid for a pressure range from 1 to 100 atm.

Finally, for solids, the sound speed for longitudinal waves, c_L, is [5]

$$c_L = \left(\frac{c_{11}}{\rho_0}\right)^{1/2} = \left(\frac{\lambda+2\mu}{\rho_0}\right)^{1/2} \tag{49.10}$$

where c_{ij} is the the elastic stiffness constant, such that c_{11} refers to longitudinal stress over longitudinal strain, and $\lambda = c_{12}$ and $\mu = c_{44}$ are the Lamé elastic constants.

49.7.2 Wave Propagation in Homogeneous Media

In order to effectively use ultrasound for measurement purposes, it is essential to be able to describe the behavior of acoustic fields. This includes the general behavior, and maybe even a detailed understanding, of the radiated field from ultrasound transducers; the transmission, reflection, and refraction of an acoustic field at boundaries and layers; and the diffraction of acoustic field at finite-sized reflectors.

The wave equation is a differential equation that formulates how an acoustic disturbance (acoustic pressure or particle velocity) propagates through a homogeneous medium. For an arbitrary wave field, the wave equation takes the following general form:

$$\nabla^2 p = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{49.11}$$

where ∇^2 is the Laplace operator. Equation 49.11 is also valid when particle velocity is substituted for pressure. If the condition of a plane wave field at a single frequency (also called a harmonic plane wave field) is imposed, the solution to Equation 49.11 is

$$p(\vec{\mathrm{r}}, t) = A\exp\left[j(\omega t - \vec{\mathrm{k}} \times \vec{\mathrm{r}})\right] + B\exp\left[j(\omega t + \vec{\mathrm{k}} \times \vec{\mathrm{r}})\right] \tag{49.12}$$

In Equation 49.12, $p(\vec{\mathbf{r}}, t)$ consists of a plane wave propagating in the direction defined by the propagation vector $\vec{\mathbf{k}}$ and a plane wave in the direction $-\vec{\mathbf{k}}$. As before, $\vec{\mathbf{r}}$ is a position vector that, in a Cartesian coordinate system, can be expressed as $\vec{\mathbf{r}} = x\hat{x} + y\hat{y} + z\hat{z}$ where $\hat{x}$, $\hat{y}$ and $\hat{z}$ are unit vectors. The magnitude of $\vec{\mathbf{k}}$ is called the wave number:

$$k = \frac{\omega}{c_0} = \frac{2\pi}{\lambda} \tag{49.13}$$

Expressing $\vec{\mathbf{k}}$ in the Cartesian coordinate system gives

$$\vec{\mathbf{k}} = k_x\hat{x} + k_y\hat{y} + k_z\hat{z} \tag{49.14}$$

When k, k_x, and k_y are specified, k_z is defined as well:

$$k_z = \sqrt{\left(\left(\frac{\omega}{c_0}\right)^2 - k_x^2 - k_y^2\right)} \tag{49.15}$$

Applying the expressions for $\vec{\mathbf{r}}$ and $\vec{\mathbf{k}}$ to Equation 49.12 and considering only the plane wave in the direction of $+\vec{\mathbf{k}}$ gives

$$p(\vec{\mathbf{r}}, t) = A\ \exp[j(\omega t - (k_x x + k_y y + k_z z))] \tag{49.16}$$

From Equation 49.16, one sees that the amplitude of a plane wave is constant and equal to A but that the phase varies with both time and space. When the direction of $\vec{\mathbf{k}}$ is specified, k_x, k_y, and k_z are found from projection onto the coordinate axes.

Solving Equation 49.11 under the assumption of spherical waves at a single frequency, with the source placed at the origin of the coordinate system, gives

$$p(r, t) = \frac{A}{r}\exp[j(\omega t - kr)] + \frac{B}{r}\exp[j(\omega t + kr)] \tag{49.17}$$

The first term is a diverging spherical wave, and the second term is a converging spherical wave. The reason that the vector dot product does not appear in Equation 49.17 is that $\vec{\mathbf{k}}$ and $\vec{\mathbf{r}}$ always point in the

same direction for spherical waves. While Equation 49.17 gives the complete solution, in most cases only a diverging spherical wave exists. If spherical waves are produced by a spherical source of radius a, then

$$p(r,t) = \frac{A}{r}\exp[j(\omega t - kr)], \quad r > a \tag{49.18}$$

49.7.3 Acoustic Intensity and Sound Levels

The acoustic intensity, I, describes the mean power transported through a unit area normal to the direction of propagation. A general expression for I for a CW pressure function is

$$I = \frac{1}{T}\int_0^T p(t)u(t)dt \tag{49.19}$$

where

T is the one full cycle of the pressure function
$u(t)$ is the particle velocity function

For a plane wave, the intensity is

$$I = \frac{p^2}{2\rho_0 c_0} = \frac{1}{2}PU \tag{49.20}$$

where P and U are the magnitudes of the pressure and the particle velocity functions, respectively. The expression for I in Equation 49.20 can also be used as an approximation for nonplanar waves, such as spherical waves, as long as (kr) is large.

The intensity levels for pulse-echo measurements are often described by the *spatial peak, temporal average* (SPTA) value, which therefore refers to the mean intensity at the point in space where the intensity is the highest. Although the guidelines for exposure levels for medical imaging, set by the FDA, vary for different parts of the body, a general upper limit is 100 mW cm^{-2}. Given that the duty cycle (pulse duration/pulse interval) is typically in the order of 0.0001 for ultrasound imaging, the temporal peak intensity is much higher and is even approaching the level where nonlinear effects can begin to be observed. Intensity levels used for NDE are application dependent but are generally in the same range as in medical ultrasound.

Sound levels are logarithmic expressions (expressions in dB) of the pressure level or the intensity level. The term SPL stands for sound pressure level and is given as

$$\text{SPL} = 20\log\left(\frac{P_e}{P_{\text{ref}}}\right)[\text{dB}] \tag{49.21}$$

where

P_e is the effective pressure of the acoustic wave
P_{ref} is the reference effective pressure

Correspondingly, IL is the intensity level defined as

$$\text{IL} = 10\log\left(\frac{I}{I_{\text{ref}}}\right)[\text{dB}] \tag{49.22}$$

where I_{ref} is the reference acoustic intensity.

Several reference levels are commonly used [6]. In air, $P_{ref} = 20$ μPa and $I_{ref} = 10^{-12}$ W m^{-2} are nearly equivalent reference levels. For water, typical pressure reference levels are $P_{ref} = 1$ μPa, $P_{ref} = 20$ μPa, or $P_{ref} = 1$ μbar = 0.1 Pa, with corresponding intensity reference levels of $I_{ref} = 6.76 \times 10^{-19}$ W m^{-2}, $I_{ref} = 2.70 \times 10^{-16}$ W m^{-2}, and $I_{ref} = 6.76 \times 10^{-9}$ W m^{-2}.

49.7.4 Wave Propagation across Layers and Boundaries

The reflection and transmission coefficients for a plane wave impinging at normal incidence on the interface between two half spaces were given in Equations 49.4 and 49.5. It is also important to consider the transmission and reflection of waves at non-normal incidence and across a layer.

When a longitudinal wave impinges at a liquid–liquid interface, both the transmitted and the reflected waves are longitudinal waves, as illustrated in Figure 49.4a. However, when a longitudinal wave propagating in a liquid medium encounters a liquid–solid interface under oblique incidence, both a longitudinal and a shear transmitted wave and a longitudinal reflected wave will result, as shown in Figure 49.4b. The change of direction of the transmitted waves relative to the incident wave is referred to as *refraction*, which is caused by the difference in sound speed between the two media.

The magnitude and direction of the reflected and transmitted waves are determined by two simple boundary conditions: (1) The *pressure* on both sides of the boundary must be equal at all times, and (2) the *normal particle velocity* on both sides of the boundary must be equal at all times; the normal particle velocity refers to the component of the particle velocity that is normal to the boundary.

The boundary conditions lead to the following two relationships:

$$\sin\theta_i = \sin\theta_r; \quad \frac{\sin\theta_i}{c_1} = \frac{\sin\theta_t}{c_2} \tag{49.23}$$

where

c_1 and c_2 are the sound speeds of the two media

θ_i, θ_r, and θ_t are the incident, reflected, and transmitted angles, respectively

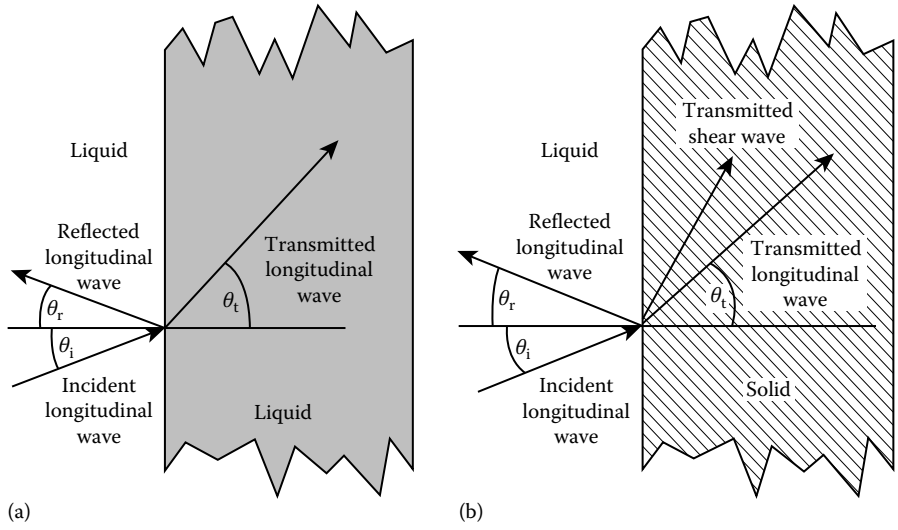


FIGURE 49.4 (a) An incident longitudinal wave at a liquid–liquid interface produces a reflected and a transmitted longitudinal wave. (b) An incident longitudinal wave at a liquid–solid interface produces a reflected longitudinal wave and transmitted longitudinal and shear waves. The incident, reflected, and transmitted angles are indicated.

From Equation 49.23, one sees that

$$\theta_r = \theta_i; \quad \theta_t = \arcsin\left(\frac{c_2}{c_1}\sin\theta_i\right) \tag{49.24}$$

The second expression in Equation 49.24 is referred to as Snell's law and quantifies the degree of refraction. Refraction affects the quality of ultrasound imaging because image formation is based on the assumption that the ultrasound beam travels in a straight path through all layers and inhomogeneities and that the sound speed is constant throughout the medium. When the actual beam travels along a path that deviates to some extent from a straight path and passes through some parts of the medium faster than it does other parts, the resulting image is a distorted depiction of reality. Correcting this distortion is a very complex problem and, in the near future, one should only expect image improvement in the simplest cases.

It can be seen from Equation 49.24 that angle θ_t is not defined if the argument to the arcsine function exceeds unity. This defines a critical incident angle as follows:

$$\sin\theta_c = \frac{c_1}{c_2} \tag{49.25}$$

at or above which the reflection coefficient is 1 and the transmission coefficient is 0. A critical angle only exists when $c_1 < c_2$. The transmitted shear and longitudinal velocities each correspond to a different critical angle of incidence.

From the boundary conditions, the reflection coefficient as a function of incident angle can be determined [7]:

$$R(\theta_i) = \frac{(z_2/z_1) - (\cos\theta_t/\cos\theta_i)}{(z_2/z_1) + (\cos\theta_t/\cos\theta_i)}, \quad \text{where } \cos\theta_t = \sqrt{1-(c_2/c_1)^2\,\sin^2\theta_i} \tag{49.26}$$

The result in Equation 49.26 is referred to as the Rayleigh reflection coefficient.

The transmission of plane waves under normal incidence through a single layer is often of interest. The dimensions and the medium parameters are defined in Figure 49.5.

By applying the boundary conditions to both interfaces of the layer, both reflection and transmission coefficients are obtained [7], as given in Equations 49.27 and 49.28:

$$R = \frac{p_r}{p_i} = \frac{(1 - z_1/z_3)\cos k_2 L + j(z_2/z_3 - z_1/z_2)\sin k_2 L}{(1 + z_1/z_3)\cos k_2 L + j(z_2/z_3 + z_1/z_2)\sin k_2 L} \tag{49.27}$$

$$T = \frac{p_t}{p_i} = \frac{2}{(1 + z_1/z_3)\cos k_2 L + j(z_2/z_3 + z_1/z_2)\sin k_2 L} \tag{49.28}$$

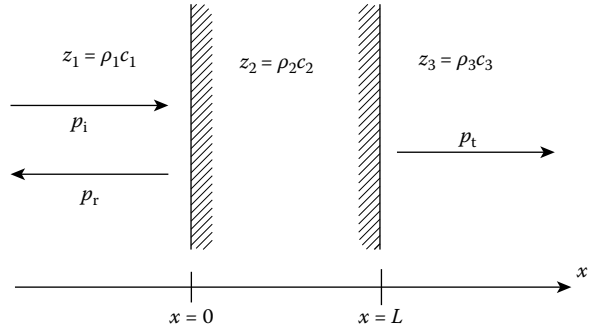


FIGURE 49.5 A pressure wave is incident on a layer of thickness L and with acoustic impedance $z_2 = \rho_2 c_2$.

A number of special conditions for Equation 49.27 can be considered, such as (1) $z_1 = z_3$, which simplifies the numerator in Equation 49.27, and (2) the layer thickness is only a small fraction of a wavelength, that is, $k_2L \ll 1$ which makes $\cos k_2L \approx 1$ and $\sin k_2L \approx k_2L$. One case is of particular interest: the choice of thickness and acoustic impedance for the layer that makes $R = 0$ and therefore gives 100% energy transmission. This is fulfilled for

$$k_2L = \pi/2 + n\pi; \quad z_2 = \sqrt{z_1 z_3} \tag{49.29}$$

The result in Equation 49.29 states that the layer must be a quarter of a wavelength thick (plus an integer number of half wavelengths) and must have acoustic impedance that is the geometric mean of the impedances of the media on either side. Among several applications of the *quarter wavelength impedance matching* is the impedance matching between a transducer and the medium, such as water. A drawback with a single matching layer is that it only works effectively over a narrow frequency range, while the actual acoustic pulse contains a fairly broad spectrum of frequencies. A better matching is achieved by using more than one matching layer, and it has been shown that a matching layer that has continuously varying acoustic impedance across the layer provides broadband impedance matching.

49.7.5 Attenuation: Its Origin, Measurement, and Applications

Attenuation refers to the damping of a signal, here specifically an acoustic signal, with travel time or with travel distance. Attenuation is typically expressed in dB, that is, on a logarithmic scale. Attenuation is an important parameter to measure in many types of materials characterization but also the parameter that sets an upper limit for the ultrasound frequency that can be used for a given measurement. In NDE, attenuation is used for grain size estimation [8,9], for characterization of composite materials, and for determination of porosity [10]. In medical ultrasound, attenuation can be used for tissue characterization [11], such as differentiating between normal and cirrhotic liver tissue, and for classification of malignancies. In flowmeters, attenuation caused by vortices can be used to measure the frequency at which they are shed; this frequency is proportional to the flow velocity.

Attenuation represents the combined effect of *absorption* and *scattering*, where absorption refers to the conversion of acoustic energy into heat due to the viscosity of the medium and scattering refers to the spreading of acoustic energy into other directions due to inhomogeneities in the medium. The absorption can, in part, be due to *classical absorption*, which varies with frequency squared, and *relaxation absorption*, which can result in a complicated frequency dependence of absorption. Gases (except for noble gases), liquids, and biological tissue exhibit mainly relaxation absorption, whereas classical absorption is most prominent in solids, which can also have a significant amount of scattering attenuation.

In general, absorption dominates in homogeneous media (e.g., liquids, gases, fine-grained metals, polymers), whereas scattering dominates in heterogeneous media (e.g., composites, porous ceramics, large-grained materials, bone). The actual attenuation and its frequency dependence can be specified fairly unambiguously for gases and liquids, while for solids it is very dependent on the manufacturing process, which determines the microstructure of the material, such as the grain structure.

Measurement of attenuation can be carried out for at least two purposes: (1) to measure the *bulk attenuation* of a given homogeneous medium and (2) to measure the *spatial distribution of attenuation* over a plane in an inhomogeneous medium. The former approach is most common in materials characterization, whereas the latter approach is found mainly in medical ultrasound. Bulk attenuation can be performed either with transmission measurements or with pulse-echo measurements, as illustrated in Figure 49.6a and b, respectively.

For the measurement of attenuation, in dB/cm, of a medium of thickness d, by transmission measurements, define $v_1(t)$ as the received signal without medium present and $v_2(t)$ as the received signal with

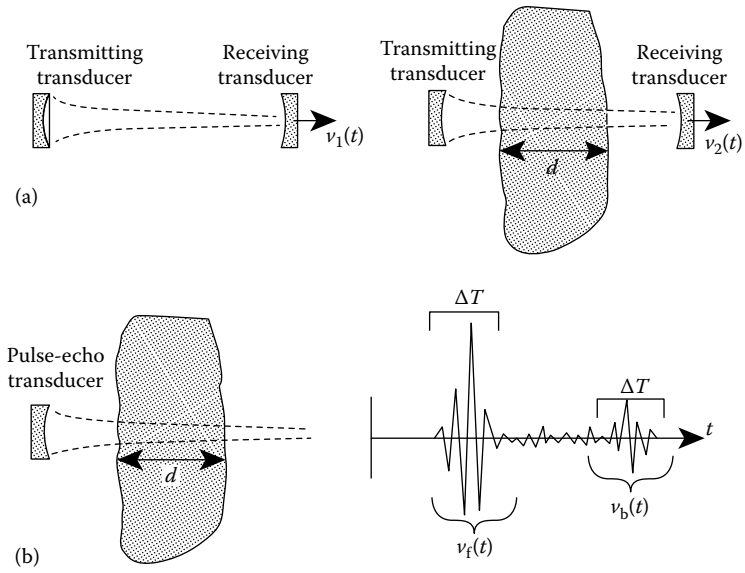


FIGURE 49.6 Measurement of bulk attenuation. (a) Measurement of attenuation by transmission measurement. (b) Measurement of attenuation by pulse-echo measurements.

medium present, as shown in Figure 49.6a. The attenuation is then determined from the ratio of the energies of the two signals, corrected for the transmission losses, as

$$\text{Att [dB/cm]} = \frac{1}{d} 10 \log \left\{ \frac{\int_0^\infty [v_1(t)]^2 dt}{\int_0^\infty [v_2(t)]^2 dt} \right\} - 20 \log \left\{ \frac{4 z_1 z_2}{(z_1 + z_2)^2} \right\} \tag{49.30}$$

Correction for transmission losses (second term) can be avoided by alternatively measuring the incremental attenuation due to an incremental thickness increase.

For the measurement of attenuation by *pulse-echo measurements*, the front and the back wall echoes are termed $v_f(t)$ and $v_b(t)$, respectively. Based on an a priori knowledge of the pulse duration, a time window ΔT is defined. The attenuation is most accurately measured when based on the energies of $v_f(t)$ and $v_b(t)$ but may alternatively be based on the amplitudes of $v_f(t)$ and $v_b(t)$, as stated in Equation 49.31:

$$\text{Att [dB/cm]} = \frac{1}{2d} 10 \log \left\{ \frac{\int_0^{\Delta t} [v_f(t)]^2 dt}{\int_0^{\Delta t} [v_b(t)]^2 dt} \right\} \cong \frac{1}{2d} 20 \log \left\{ \frac{\text{peak ampl., } v_f(t)}{\text{peak ampl., } v_b(t)} \right\} \tag{49.31}$$

Accurate attenuation measurements require attention to several potential pitfalls: (1) diffraction effects (even in the absence of attenuation, echoes from different ranges vary in amplitude, due to beam spreading); (2) misalignment effects (if the reflecting surface is not normal to the transducer axis, there is a reduction in detected signal amplitude, due to phase cancellation at the transducer surface); and (3) transmission losses whose magnitude it is not always easy to determine.

The spatial distribution of attenuation must be measured with pulse-echo measurements, where it is assumed that a backscatter signal of sufficient amplitude can be received from all regions of the medium.

Use is made of the frequency dependence of attenuation, which has the effect that the shift in mean frequency of a given received echo relative to the mean frequency of the transmitted signal varies proportional to the total absorption. The *rate of shift* in mean frequency is thus proportional to the local attenuation.

49.7.6 CW Fields from Planar Piston Sources

In discussing the acoustic fields generated by acoustic radiators (transducers), a clear distinction must be made between fields due to CW excitation and due to pulse excitation. Although the overall field patterns for these two cases are quite similar, the detailed field structure is very different. In this section, only CW fields are discussed.

When a CW excitation voltage is applied to a planar piston transducer, the resulting acoustic field can be divided into a *near-field* region and a *far-field* region. This division is particularly distinct when the transducer has a circular geometry. A piston transducer simply refers to a transducer with the same velocity amplitude at all points on the surface. The length of the near field, NF, is given as

$$\mathrm{NF} = \frac{a^2}{\lambda} \tag{49.32}$$

where

a is the radius of the transducer
λ is the wavelength

As shown in Figure 49.7a, the near field is approximately confined, while the far field is diverging. The angle of divergence, θ, is approximately

$$\theta = \arcsin\left(0.61\frac{\lambda}{a}\right) \tag{49.33}$$

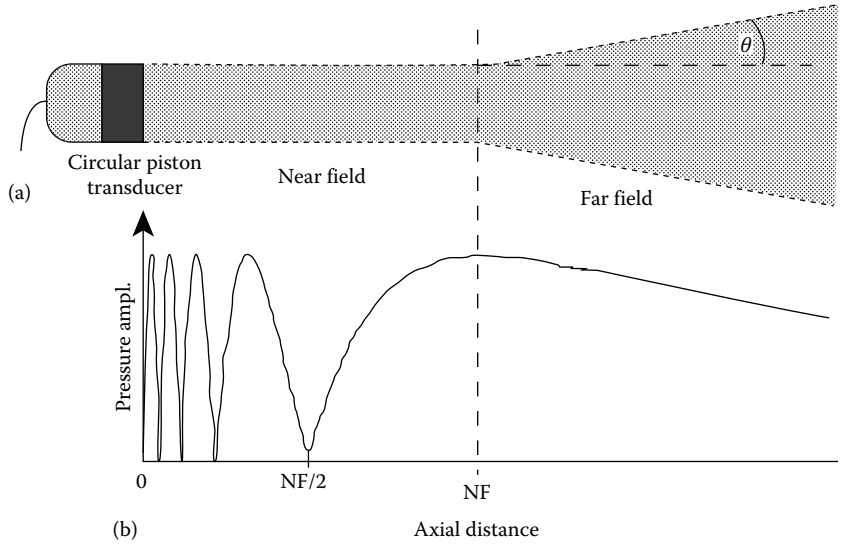


FIGURE 49.7 Pressure fields from a circular planar piston transducer operating with CW excitation. NF = near field. (a) Approximate field distribution in near field and far field. (b) Axial pressure in near field and far field for (ka) = 31.4.

Additional comments to the simplified representation in Figure 49.7a are in order:

1. The actual field has no sharp boundaries, in contrast to what is shown in Figure 49.7.
2. While the beam diameter is roughly constant in the near field, it is not as regular as shown, and at the same time, the near-field structure is very complex.
3. The angle of divergence, θ, is only clearly established well into the far field.
4. The depiction of the far field shows only the main lobe; in addition, there are side lobes, which are directions of significant pressure amplitude, separated by *nulls* (i.e., directions with zero-pressure amplitude).

In general, analytical expressions do not exist for the pressure magnitude at an arbitrary field point in the near field, and the pressure must instead be calculated by numerical techniques. However, an exact expression exists for the axial pressure amplitude, P_{ax}, valid in both near and far field [12].

$$P_{ax}(x) = 2\rho_0 c_0 U_0 \left| \sin\left(0.5ka \left[\sqrt{\frac{x^2}{a^2} + 1} - \left(\frac{x}{a} \right) \right] \right) \right| \tag{49.34}$$

In Equation 49.34, U_0 is the amplitude of the velocity function on the surface of the transducer. The expression is derived for a baffled planar piston transducer where the term "baffled" means that the transducer is mounted in a large rigid surface, called a baffle. Figure 49.7b shows the amplitude of the axial field for a planar piston transducer for which $(ka) = 31.4$. For field points located more than 3–4 near-field distances away from the transducer, a general expression for the pressure amplitude can be derived [12]:

$$P(r, \theta) = \frac{\rho_0 c_0}{2} U_0 ka \left(\frac{a}{r} \right) \left| \left[\frac{2J_1(ka \sin\theta)}{ka \sin\theta} \right] \right| \tag{49.35}$$

where

$J_1(\cdot)$ is the Bessel function of the first order
The variable r is the length of the position vector defining the field point
Angle θ is the angle that the position vector makes with the x-axis

Of interest in evaluating transducers in the far field is the *directivity*, D, defined as follows:

$$D(x) = \frac{I_{ax}(x)[\text{given source}]}{I_{ax}(x)[\text{simple source}]} \tag{49.36}$$

Thus, the directivity gives the factor with which the axial intensity of the given source is increased over that of a simple source (omnidirectional radiator), radiating the same total energy. For a baffled, circular planar piston transducer, the directivity in the far field can be calculated to be [12]

$$D = \frac{(ka)^3}{ka - J_1(2ka)} \tag{49.37}$$

When $(2ka) \gg 1$, Equation 49.37 can be approximated to $D = (ka)^2$. Directivity values in the 1000s are common. For example, a 3.5 MHz transducer with 1 cm diameter radiating into water has a (ka) value of 73.3 and a directivity of 5373.

Often, focused transducers are used where the focusing is either created by the curvature of the piezoelectric element or by an acoustic lens in front of the transducer. The degree of focusing is determined by the (ka) value of the transducer.

49.7.7 Generation of Ultrasound: Piezoelectric and Magnetostrictive Phenomena

Ultrasound transducers today are available over a wide frequency range, in many sizes and shapes, and for a wide variety of applications. The behavior of the ultrasound transducer is determined by several parameters: the transduction material, the backing material, the matching layer(s), and the geometry and dimension of the transducer. A good overview of ultrasound transducers is available in [13].

The transduction material is most commonly a piezoelectric material but can, for some applications, be a magnetostrictive material instead. These materials are inherently narrowband, meaning that they work efficiently only over a narrow frequency range. This is advantageous for CW applications such as ultrasound welding and ultrasound hyperthermia, but is a problem for imaging applications, as impulse excitation will produce a long pulse with poor resolving abilities. To overcome this deficiency, a *backing material* is tightly coupled to the back side of the transducer for the purpose of damping the transducer and shortening the pulse. However, the backing material also reduces the sensitivity of the transducer. Some of this reduced sensitivity can be regained by the use of a *matching layer*, specifically selected for the medium of interest. As seen in Equation 49.29, a quarter wavelength matching layer can provide 100% efficient coupling to a medium, albeit only at one frequency. A combination of several matching layers can provide a more broadband impedance matching. The field is determined by both the geometry (planar, focused, etc.) of the transducer and by the frequency content of the velocity function of the surface of the transducer. In this section, unique aspects of the transduction material itself are described.

49.7.7.1 Piezoelectric Materials

A piezoelectric material exhibits a mechanical strain (relative deformation) due to the presence of an electric field and generates an electric field when subjected to a mechanical stress. A detailed review of piezoelectricity is given in [14]. Piezoelectric materials can either be (1) natural material such as quartz, (2) man-made ceramics (e.g., barium titanate [BaTi], lead zirconate titanate [PZT], or lead meta-niobate), and (3) man-made polymers (e.g., polyvinylidene fluoride [PVDF]). The piezoelectric ceramics are the most commonly used materials for ultrasound transducers. These ceramics are made piezoelectric by poling process in which the material is subjected to a strong electric field while at the same time heating it to above the material's Curie temperature.

Several material constants determine the behavior of a given piezoelectric material, the most important of which are listed in Table 49.1 and defined as follows. An extensive list of parameter values for various piezoelectric materials is available in [15].

TABLE 49.1 List of Most Significant Piezoelectric Parameters for Common Piezoelectric Materials

Parameter	Barium Titanate ($BaTiO_3$)	Lead Zirconate Titanate, PZT-5	Lead Meta-Niobate, $PbNb_2O_6$	Polyvinylidene Fluoride (PVDF)
d_{33}	149 (10^{-12} m/V)	374 (10^{-12} m/V)	75 (10^{-12} m/V)	22 (10^{-12} m/V)
g_{33}	14 (10^{-3} Vm/N)	25 (10^{-3} Vm/N)	35 (10^{-3} Vm/N)	339 (10^{-3} Vm/N)
k_{33}	0.50	0.70	0.38	$k_{31} = 0.12$
k_T	0.38	0.68	0.40	0.11
Q_M	600	75	5	19

d_{33}, Transmission constant; g_{33}, Receiving constant; k_{33}, Piezoelectric coupling coefficient; k_T, Piezoelectric coupling coefficient for a transverse clamped material; Q_M, Mechanical Q.

The transmission constant, d_{33}, gives the mechanical deformation of piezoelectric materials for frequencies well below the resonance frequency. A transducer with a large d_{33} value will therefore become an efficient transmitter. If an electric field, E_3, is applied in the polarized direction of a piezoelectric rod or disk, the strain, S_3, in that direction is approximately

$$S_3 = d_{33}E_3 = d_{33}\frac{V_{\text{appl}}}{l} \tag{49.38}$$

where

V_{appl} is the applied voltage
l is the thickness

The total deformation, Δl, becomes

$$\Delta l = d_{33}V_{\text{appl}} \tag{49.39}$$

The receiving constant, g_{33}, defines the sensitivity of a transducer element as a receiver when the frequency of the applied pressure is well below the resonance frequency of the transducer. If the applied stress (force/area) in the direction of polarization is T_3, the output voltage from a rod or disk is approximately

$$v_{\text{out}} = g_{33}T_3l \tag{49.40}$$

The coupling coefficient describes the power conversion efficiency of a piezoelectric transducer, *operating at or near resonance*. Specifically, k_{33} is the coupling coefficient for an undamped rod; that is, the rod is allowed to deform in the directions orthogonal to the direction of applied force or voltage. In contrast, k_T is the coupling coefficient for a *clamped* disk. Finally, Q_{M} gives the mechanical Q, which is a measure for how narrowband the transducer material inherently is.

Whereas expressions of the type given in Equations 49.38 through 49.40 are adequate for describing static or low-frequency behavior, the behavior near resonance where most transducers operate requires more complex models, which are beyond the scope of this chapter. The Mason model is adequate for narrowband modeling of transducers, whereas the KLM or Redwood models better describe the transducers for broadband applications [13,15].

49.7.7.2 Magnetostrictive Materials

The magnetostrictive phenomenon refers to a magnetically induced contraction or expansion in ferroelectric media, such as in nickel or alfenol, and was discovered by Joule in 1847. Magnetostrictive materials are generally used for ultrasound frequencies below 100 kHz and are therefore relevant mainly for underwater applications. Eddy current losses influence the performance of magnetostrictive materials, but the losses can be reduced by constructing the magnetostrictive transducer from thin laminations.

49.7.7.3 Transducer Specifications

Ultrasound transducers are generally specified by their diameter, center frequency, focal distance, and type of focusing (if applicable). The transducer can be designed as a contact transducer, a submersible transducer, an air transducer, etc.; in addition, the type of connector or cabling can be specified. In many cases, measurement data for the actual transducer can be supplied by the vendor in the form of a measured pressure pulse and the corresponding frequency spectrum, recorded with a hydrophone at a specific field point. Similarly, the beam profile can be measured in the form of pulse amplitude or pulse energy as a function of lateral position.

Detailed information about the acoustic field from a radiating transducer can be obtained with either the Schlieren technique or the optical interferometric technique. Such instruments are quite expensive, falling in the $50–$120 K range.

49.7.8 Display Formats for Pulse-Echo Measurements

The basic description of ultrasound imaging was presented earlier in this chapter. Based on the information presented so far in this section, more specific aspects of pulse-echo ultrasound imaging will now be described. Different display formats are used; the simplest of these is the *A-mode* display.

49.7.8.1 A-Mode Display

When a pulse-echo transducer has emitted a pulse in a given direction and has been switched to receive mode, an output signal from the transducer is produced based on detected echoes from different ranges, as illustrated conceptually in Figure 49.1a. In this signal, distance to the reflecting structure has been converted into time from the start of signal. This signal is often referred to as the RF signal. Demodulating this signal (i.e., generating the envelope of the signal) produces the A-mode (amplitude mode) display or the A-line signal. This signal can be the basis for range measurements, attenuation measurements, and measurement of reflection coefficient.

49.7.8.2 M-Mode Display

If the A-mode signal from a transducer in a fixed position is used to *intensity* modulate a cathode-ray tube (CRT), such as a monitor or oscilloscope, along a straight vertical line, a line of dots with brightness according to echo strengths would appear on the screen. Moving the display electronically across the screen will result in a set of straight horizontal lines. Now consider the case where the reflecting structures are moving in a direction toward or away from the transducer while pulse-echo measurements were being performed. This gives in variations in the arrival time of echoes in the A-line signal, and the resulting lines across the screen are no longer straight, but curved, as determined by their velocity. Such a display is called *M-mode*, or motion mode, display. An application for this would be measurement of the diameter variation of a flexible tube, or blood vessel, due to a varying pressure inside the tube.

49.7.8.3 B-Mode Display

If pulse-echo measurements are repeatedly being performed, while the transducer scans the objects of interest, an image of the object can be generated, as illustrated in Figure 49.1. Specifically, each A-line signal is used to intensity modulate a line on a CRT corresponding to the location of ultrasound beam, which produces an image that maps the reflectivity of the structures in the object. The resulting image is called a *B-mode* (or brightness mode) image. If the transducer is moved linearly, a rectangular image is produced, whereas a rotated transducer generates a pie-shaped or sector image. This motion is typically done electronically or electromechanically, as described earlier under single-element and array transducers. When the scanning is done rapidly (say, 30 scans/s), the result is a real-time image.

Many forms of signal processing and image enhancement can be applied in the process of generating the B-mode image. Echoes from deeper lying structures will generally be of smaller amplitude, due to the attenuation of the overlying layers of the medium. Attenuation correction is made especially in medical imaging, so that echoes are displayed approximately with their true strength. This correction consists of a time-varying gain, called *time-gain control*, or TGC, such that the early arriving echoes experience only a low gain and later echoes from deeper lying structures experience a much higher gain. Various forms for signal compressions or nonlinear signal transfer function can selectively enhance the weaker or the stronger echoes.

49.7.8.4 C-Mode Display

In a C-mode display, only echoes from a specific depth will be imaged. To generate a complete image, the transducer must therefore be moved in a raster scan fashion over a plane. C-scan imaging is slow and cannot be used for real-time imaging but has several applications in NDE for examining a given layer, or interface, in a composite structure, or the inner surface of a pipe.

49.7.9 Flow Measurements by Doppler Signal Processing

Flow velocity can be obtained by various ultrasonic methods, for example, by measuring the Doppler frequency or Doppler spectrum. The Doppler frequency is the difference between the frequency (and pitch) of a moving sound source, as measured by a *stationary* observer, and the actual frequency of the sound source. The change in frequency is determined by the speed and direction of the moving source. The classical example is a moving locomotive with its whistle blowing; the pitch is increased when the train moves toward the observer and vice versa. With respect to ultrasound measurements, only the reflecting (or scattering) gas, fluid, or structure is moving and not the sound source, yet the Doppler phenomenon is present here as well. In order for ultrasound Doppler to function, the gas or fluid must contain scatterers that can reflect some of the ultrasound energy back to the receiver. The Doppler frequency, f_d, is given as follows:

$$f_d = \frac{2v}{c_0}\cos\theta \tag{49.41}$$

where

v is the velocity of the moving scatterers

θ is the angle between the velocity vector and the direction of the ultrasound beam

Doppler flowmeters require only access to the moving gas, fluid, or object from one side. For industrial use, when the fluids or gases often do not contain scatterers, *transmission* methods are preferred; these methods, however, are not based on the Doppler principle.

Two main categories of Doppler systems exist: the CW Doppler system and the PW (pulsed wave) Doppler system. The CW Doppler system transmits a continuous signal and does not detect the distance to the moving structure. It is therefore only applicable when there is just one moving structure or cluster of scatterers in the acoustic field. For example, a CW Doppler is appropriate for assessing the pulsatility and nature of blood flow in an arm or leg. CW Doppler systems are small and relatively inexpensive.

The PW Doppler system transmits a short burst at precise time intervals; it is therefore inherently a sampled system, and, as such, is subject to aliasing. It measures changes from one transmitted pulse to the next in the received signal from moving structures at one or several selected ranges and is able to determine both velocity and range. Assuming that the ultrasound beam is narrow and thus very directional, the PW Doppler can create a *flow image*; that is, it can indicate the magnitude and direction of flow over an image plane. Commonly, color and color saturation are used to indicate direction and speed in a flow image, respectively. In contrast to CW Doppler systems, the PW Doppler systems require much signal processing and are therefore typically expensive and often part of a complete imaging system. When a distribution of velocities, rather than a single velocity, is encountered, as is often the case with fluid flow, a Doppler spectrum rather than a Doppler frequency is determined. A fast Fourier transform (FFT) routine is then used to reveal the Doppler frequencies and thus the velocity components present.

49.8 Review of Common Applications of Ultrasound and Their Instrumentation

49.8.1 Range Measurements, Air

Ultrasound range measurements are used in cameras, in robotics, for determining dimensions of rooms, etc. Measurement frequencies are typically around 50–60 kHz. The measurement concept is pulse echo but with burst excitation rather than pulse excitation. Special electronic circuitry and a thin low-acoustic-impedance air transducer are the most commonly used. Rugged solid or composite piezoelectric-based transducers, however, can also be used, sometimes up to about 500 kHz.

49.8.2 Thickness Measurement for Testing and Process Control

Measurement of thickness is a widely used application of ultrasound. The measurements can be done with direct coupling between the transducer and the object of interest or—if good surface contact is difficult to establish—with a liquid or another coupling agent between the transducer and the object. Ultrasound measurements of thickness have applications in process control, quality control, measuring buildup of ice on an aircraft wing, detecting wall thickness in pipes, as well as medical applications. The instrumentation involves a broadband transducer, pulser–receiver, and display or, alternatively, echo-detecting circuitry and numerical display.

49.8.3 Detection of Defects, Such as Flaws, Voids, and Debonds

The main ultrasound application in NDE is inspection for the localization of voids, crack, flaws, debonding, etc. [3]. Such defects can exist immediately after manufacturing or were formed due to stresses, corrosion, etc. Various types of standard or specialized flaw detection equipment are available from ultrasound NDE vendors.

49.8.4 Doppler Flow Measurements

The flow velocity of a liquid or a moving surface can be determined through Doppler measurements, provided that the liquid or the surface scatters ultrasound back in the direction of the transducer and that the angle between the flow direction and the ultrasound beam is known. Further details are given in the section about Doppler processing. CW and PW Doppler instruments are commercially available, with CW instrumentation being by far the least expensive.

49.8.5 Upstream/Downstream Volume Flow Measurements

When flow velocity is measured in a pipe with access to one or both sides, an ultrasound transmission technique can be used in which transducers are placed on the same or opposite sides of the pipe, with one transducer placed further upstream than the other transducer. From the measured *difference* in travel time between the upstream direction and the downstream direction, and knowledge about the pipe geometry, the volume flow can be determined. Special clamp-on transducers and instrumentation are available. An overview of flow applications in NDE is given in [16].

49.8.6 Elastic Properties of Solids

Since bulk sound speed varies with the elastic stiffness of the object, as given in Equation 49.11, sound speed measurements can be used to estimate elastic properties of solids under different load conditions and during solidification processes. Such measurements can also be used for measurement of product uniformity and for quality assurance. The measurements can be performed on bulk specimens or on thin rods, using either pulse-echo or transmission instrumentation [17]. Alternatively, measurements of the material's own resonance frequencies can be performed for which commercial instruments, such as the *Grindo-sonics*, are available.

49.8.7 Porosity and Grain Size Estimation

Measurement of ultrasound attenuation can reveal several materials parameters. By observing the attenuation in metals as a function of frequency, the grain size and grain size distribution can be estimated. Attenuation has been used for estimating porosity in composites. In medical ultrasound, attenuation is

widely used for tissue characterization, that is, for differentiating between normal and pathological tissues. Pulse-echo instrumentation interfaced with a digitizer and a computer for data analysis is required.

49.8.8 Acoustic Microscopy

The measurement approaches utilized in acoustic microscopy are similar to other ultrasound techniques, in that A-scan, B-scan, and C-scan formats are used. It is in the applications and the frequency ranges where acoustic microscopy differs from conventional pulse-echo techniques. Although acoustic microscopes have been made with transducer frequencies up to 1 GHz, the typical frequency range is 20–100 MHz, giving spatial resolutions in the range from 100 to 25 μm. Acoustic microscopy is used for component failure analysis, electronic component packaging, and internal delaminations and disbonds in materials, and several types of acoustic microscopes are commercially available.

49.8.9 Medical Ultrasound

Medical imaging is a large and diverse application area of ultrasound, especially in obstetrics, cardiology, and vascular studies and for detecting lesions and abnormalities in organs. The display format is either B-mode, using grayscale image, or a combination of Doppler and B-mode, with flow presented in color and stationary structures in grayscale. A wide variety of instruments and scanners for medical ultrasound are available.

49.8.10 Selected Manufacturers of Ultrasound Products

Table 49.2 contains a representative list of ultrasound equipment and manufacturers.

TABLE 49.2 List of Products for and Manufacturers of Ultrasound Measurements

Product Type	Manufacturer
Ultrasound transducers	GE Panametrics, 221 Crescent St., Waltham, MA 02154. (800) 225-8330 http://www.rshydro.co.uk/GE-Panametrics-mid-3.html
Ultrasound transducers	Krautkramer Branson Inc., 50 Industrial Park Rd., Lewistown, PA 17044. (717) 242-0327, http://www.krautkramer.com.au/
Range measurements, air	Polaroid Corporation, Ultrasonics Components Group, 119 Windsor Street, Cambridge, MA 02139. (800) 225-1618, http://www.polaroid.com/en/stream
Pulser–receivers	Panametrics, 221 Crescent St., Waltham, MA 02154. (800) 225-8330
Pulser–receivers	JSR Ultrasonics, 3800 Monroe Ave., Pittsfield, NY 14534. (716) 264-0480, http://jsrultrasonics.com/
Ultrasound power ampl.	Amplifier Research, 160 School House Rd., Souderton, PA 18964. (800) 254-2677, http://www.amp-research.com/
Ultrasound power ampl.	Ritec, 60 Alhambra Rd., Suite 5, Warwick, RI 02886. (401) 738-3660, http://www.ritec.com/
NDE instrumentation	Panametrics, 221 Crescent St., Waltham, MA 02154. (800) 225-8330
NDE instrumentation	Krautkramer Branson Inc., 50 Industrial Park Rd., Lewistown, PA 17044. (717) 242-0327
Acoustic microscopy	Sonoscan, 530 E. Green St., Bensenville, IL 60106. (708) 766-4603, http://www.sonoscan.com/
Medical imaging	Hewlett Packard, Andover, MA; ATL, Bothell, WA; Diasonics, Milpitas, CA; Siemens Ultrasound, Issaquah, WA.
Schlieren-based imaging of acoustic fields	Optison, 568 Weddell Drive, Suite 6, Sunnyvale, CA 94089. (408) 745-0383
Optical-based imaging of acoustic fields	UltraOptec, 27 de Lauzon, Boucherville, Quebec, Canada J4B 1E7. (514) 449-2096

49.9 Advanced Topics in Ultrasound

49.9.1 Overview of Diffraction

The ultrasound theory presented thus far has emphasized basic concepts, and the applications that have been discussed tacitly assume that the field from the transducer is a plane wave field. This simplifying assumption is acceptable for applications such as basic imaging and measurements based on travel time. However, the plane wave assumption introduces errors when materials parameters (e.g., attenuation, surface roughness, and object shape) are sought to be measured with ultrasound. Therefore, to use ultrasound as a quantitative tool, an understanding is needed of the structure of the radiated acoustic field from a given transducer with a given excitation and—equally important—the ability to calculate the actual radiated field. This leads to the topic of diffraction, which is the effect that accounts for the complex structure of both radiated and scattered fields. Not surprisingly, there are direct parallels between optical diffraction and acoustic diffraction. (As a separate issue, it should be noted that the ultrasound theory presented here assumes that the wave amplitudes [pressure, displacement] are small enough so that nonlinear effects can be disregarded.)

Diffraction is basically an edge effect. Whereas a plane wave incident on a large planar interface is reflected in a specific direction, the plane wave incident on an edge results in waves scattered in many directions. Similar considerations hold for the field produced by a transducer: The surface of the transducer produces a so-called *geometric wave* that has the shape of the transducer itself; the edge of the transducer, however, generates an *edge wave* with the shape of an expanding torus. The actual pressure field is a combination of the two wave fields. Very close to a large transducer, the geometric wave dominates, and diffraction effects might not need to be considered. Over small regions far away from the transducer, the field can be approximated by a plane wave field, and diffraction does not need to be considered. However, in many practical cases, diffraction must be considered if detailed information about the pressure field is desired, and numerical methods must be employed for the calculations.

The structure of the axial field, shown in Figure 49.7b, is a direct result of diffraction. Numerical evaluation of the diffracted field from a transducer can be done in several ways: (1) use of the Fresnel or Fraunhofer integrals (not applicable close to the transducer) to calculate the field at a single frequency at a specified plane normal to the transducer axis; (2) calculation of the pressure function at any point of interest in space, based on a specified velocity function, $u(t)$, on the surface of the transducer, using Rayleigh integral; (3) decomposing the velocity field in the plane of the transducer into its plane wave components, using a 2-D Fourier transform technique, followed by a forward propagation of the plane waves to the plane of interest and an inverse Fourier transform to give the diffracted field; or (4) use of finite element methods to calculate the diffracted field at any point or plane of interest. In the following, methods (1) and (2) will be described.

49.9.2 Fresnel and Fraunhofer Diffraction

Let the velocity function on the surface of the transducer, $u(x, y)$, be specified for a particular frequency, ω. Assume that the transducer is located in the $(x, y, 0)$ plane and that one is interested in the pressure field in the (x_0, y_0, z) plane. The Fresnel diffraction formulation [18] assumes that the paraxial approximation is fulfilled; requiring z to be at least five times greater than the transducer radius, in which case the *Fresnel diffraction integral* applies:

$$p(x_0, y_0, z, \omega) = \frac{A_0}{\lambda_z} \iint_s u(x, y, \omega) \exp\left[-jk\frac{x^2 + y^2}{2z}\right] \exp\left[jk\frac{x_0 x + y_0 y}{z}\right] dxdy \tag{49.42}$$

where

s is the surface of the transducer

A_0 is a constant

If one defines the two first terms of the integrand as some complex spatial function, $\Gamma(x, y)$, then Equation 49.42 is a scaled Fourier transform of $\Gamma(x, y)$.

If $k(x^2 + y^2)/2z \ll 1$, or, equivalently, $z > 10\, a^2/\lambda$, the second term in Equation 49.42 can be ignored, and the resulting equation is called the *Fraunhofer diffraction integral:*

$$p(x_0, y_0, z, \omega) = \frac{A_0}{\lambda z} \iint_s u(x, y, \omega) \exp\left[jk \frac{x_0 x + y_0 y}{z} \right] dx dy \tag{49.43}$$

Thus, one can only use the Fraunhofer integral for calculating the far-field diffraction. From Equation 49.43, one can make the interesting observation that the far field of a transducer is a scaled version of the Fourier transform of the source.

49.9.3 Pressure Function at a Given Field Point, Based on Rayleigh Integral

While the Fresnel and Fraunhofer diffraction methods are CW methods, calculation of pressure from the Rayleigh integral is fundamentally an impulse technique and is as such better suited for analysis of pulse-echo measurements. The basis for the calculation is the *velocity potential impulse response*, $h(\vec{\mathbf{r}}, t)$, obtained from the Rayleigh integral:

$$h(\vec{\mathrm{r}}, t) = \frac{1}{2\pi} \iint_s \frac{\delta(t - r'/c)}{r'}\, dS \tag{49.44}$$

In Equation 49.44, r' is the distance from dS on the surface of the transducer to the field point, defined by the position vector $\vec{\mathbf{r}}$, and $\delta(t)$ is the Dirac delta function. As can be seen, $h(\vec{\mathbf{r}}, t)$ is the result of an impulsive velocity excitation on the surface of the transducer and is a function of both time and a spatial location. It is important to note that $h(\vec{\mathbf{r}}, t)$ exists in analytical form for several transducer geometries and, by extension, for annular and linear array transducers [19].

For the case of an arbitrary velocity function, $u(t)$, on the transducer surface, the corresponding velocity potential, $\phi(\vec{\mathbf{r}}, t)$, is obtained as

$$\phi(\vec{\mathbf{r}}, t) = u(t) \otimes h(\vec{\mathbf{r}}, t) \tag{49.45}$$

where $\otimes$ refers to time domain convolution. Both particle velocity, $u(\vec{\mathbf{r}}, t)$, and pressure, $p(\vec{\mathbf{r}}, t)$, can be found from $\phi(\vec{\mathbf{r}}, t)$, as follows:

$$u(\vec{\mathbf{r}}, t) = \nabla \phi(\vec{\mathbf{r}}, t) \tag{49.46}$$

$$p(\vec{\mathbf{r}}, t) = -\rho_0 \frac{\partial}{\partial t} \phi(\vec{\mathbf{r}}, t) \tag{49.47}$$

where ∇ is the gradient operator.

Thus, from the preceding expressions, the pressure can be calculated for any field point, $\vec{\mathbf{r}}$, when $u(t)$ and the transducer geometry are defined. In this calculation, all diffraction effects are included. However, given the high frequency content in $h(\vec{\mathbf{r}}, t)$ and in particular in the time derivative of $h(\vec{\mathbf{r}}, t)$, care must be taken to avoid aliasing errors, as described in [19].

49.9.3.1 Received Signal in Pulse-Echo Ultrasound

The expression in Equation 49.47 allows for quantitative evaluation of the pressure field for an arbitrary point, line, or plane. However, it does not describe the calculation of the received signal in a pulse-echo system.

Consider a small planar reflector, placed in a homogeneous medium, and referred to as dR. The reflector has the area dA. The dimensions of dR must be small with respect to the shortest wavelength in the insonifying pulse. The location and the orientation of the planar reflector is given by $\vec{\mathbf{r}}$ and $\vec{\mathbf{n}}$, respectively, where $\hat{\mathbf{n}}$ is a unit normal vector to the small reflector.

The voltage from the receiving transducer in a pulse-echo system due to dR is termed $dv(\vec{\mathbf{r}}, t)$ and can be determined when $u(t)$ is specified. The electroacoustic transfer function for both the transmitting and the receiving transducer is assumed to be unity for all frequencies. For the case when the acoustic impedance of dR is much higher than that of the medium, $dv(\vec{\mathbf{r}}, t)$ is given as [20]:

$$dv(\vec{\mathbf{r}}, t) = A_0 \rho_0 \frac{\cos[\psi(\vec{\mathbf{r}})]}{c_0} \left[h(\vec{\mathbf{r}}, t) \otimes h(\vec{\mathbf{r}}, t) \otimes \frac{\partial^2}{\partial t^2} u(t) \right] dA$$

$$= A_0 \cos[\psi(\vec{\mathbf{r}})]\ u(t) \otimes \frac{\rho_0}{c_0} \left[\frac{\partial^2}{\partial t^2} (h(\vec{\mathbf{r}}, t) \otimes h(\vec{\mathbf{r}}, t)) \right] dA \tag{49.48}$$

where

A_0 is determined by the reflection coefficient of the reflector
The term $\cos[\psi(\vec{\mathbf{r}})]$ is a correction term (obliquity factor)
$\psi(\vec{\mathbf{r}})$ is the angle between $\hat{\mathbf{n}}$ and the propagation direction of the wave field at $\vec{\mathbf{r}}$

For an extended surface, the received voltage can be found by decomposing the surface into small reflectors and calculating the total received signal as the sum of the contributions from all the small reflectors. An efficient numerical technique for this type of integration has been developed [21].

References

1. C. M. Fortunko and D. W. Fitting, Appropriate ultrasonic system components for NDE of thick polymer-composites, in D. O. Thompson and D. E. Chimenti, *Review of Progress in Quantitative Nondestructive Evaluation, Vol. 10B*. New York: Plenum Press, 1991, pp. 2105–2112.
2. C. R. Hill, Medical imaging and pulse-echo imaging and measurement, in C. R. Hill (ed.), *Physical Principles of Medical Ultrasound*, New York: Halsted Press, 1986, Chapters 7 and 8, pp. 262–304.
3. L. C. Lynnworth, *Ultrasonic Measurements for Process Control*, San Diego, CA: Academic Press, 1989, pp. 53–89.
4. L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, 3rd edn., New York: John Wiley & Sons, 1982, p. 106.
5. J. D. Achenbach, *Wave Propagation in Elastic Solids*, 1st edn., New York: Elsevier Science, 1975, p. 123.
6. L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, 3rd edn., New York: John Wiley & Sons, 1982, pp. 115–117.
7. L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, 3rd edn., New York: John Wiley & Sons, 1982, pp. 127–133.
8. J. Saniie and N. M. Bilgutay, Quantitative grain size evaluation using ultrasonic backscattered echoes, *J. Acoust. Soc. Am.*, 80, 1816–1824, 1986.
9. E. P. Papadakis, Scattering in polycrystalline media, in P. D. Edmonds (ed.), *Ultrasonics*, New York: Academic Press, 1981, pp. 237–298.
10. S. M. Handley, M. S. Hughes, J. G. Miller, and E. I. Madaras, Characterization of porosity in graphite/epoxy laminates with polar backscatter and frequency dependent attenuation, *1987 Ultrasonics Symposium.*, Denver, CO, 1987, pp. 827–830.
11. J. C. Bamber, Attenuation and absorption, in C. R. Hill (ed.), *Physical Principles of Medical Ultrasound*, New York: Halsted Press, 1986, pp. 118–199.

12. L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, 3rd edn., New York: John Wiley & Sons, 1982, pp. 176–185.
13. M. O'Donnell, L. J. Busse, and J. G. Miller, Piezoelectric transducers, in P. D. Edmonds (ed.), *Ultrasonics*, New York: Academic Press, 1981, pp. 29–65.
14. Collaboration, An American National Standard—IEEE standard on piezoelectricity, IEEE Transactions on Sonic and Ultrasonic, v SU-31, n 2, pp. 1–55, March 1984.
15. G. S. Kino, *Acoustic Waves*, Englewood Cliffs, NJ: Prentice-Hall, 1987, pp. 17–83 and 554–557.
16. L. C. Lynnworth, *Ultrasonic Measurements for Process Control*, San Diego, CA: Academic Press, 1989, pp. 245–368.
17. L. C. Lynnworth, *Ultrasonic Measurements for Process Control*, San Diego, CA: Academic Press, 1989, pp. 537–557.
18. V. M. Ristic, *Principles of Acoustic Devices*, New York: John Wiley & Sons, 1983, pp. 316–320.
19. D. P. Orofino and P. C. Pedersen, Multirate digital signal processing algorithm to calculate complex acoustic pressure fields, *J. Acoust. Soc. Am*., 92, 563–582, 1992.
20. A. Lhemery, Impulse-response method to predict echo responses from targets of complex geometry. I. Theory, *J. Acoust. Soc. Am*., 90, 2799–2807, 1991.
21. S. K. Jespersen, P. C. Pedersen, and J. E. Wilhjelm, The diffraction response interpolation method, *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, 45, 1485–1496, November 1998.

Further Information

Edmonds, P. D. (ed.), *Ultrasonics*, (Vol. 19 in the series: *Methods of Experimental Physics*). New York: Academic Press, 1981, in-depth description of ultrasound interaction with many types of materials, along with discussion of ultrasound measurement approaches.

Ensminger, D. and Bond, L. J., *Ultrasonics: Fundamental, Technologies, and Applications*, 3rd edn., Boca Raton, FL: CRC Press, 2012.

Ensminger, D. and Stulen, F. B., *Ultrasonics: Data, Equations and Their Practical Uses*, Boca Raton, FL: CRC Press, 2009.

Jensen, J. A., *Estimation of Blood Velocities Using Ultrasound*, Cambridge, U.K.: Cambridge University Press, 1996, a very up-to-date book about ultrasound Doppler measurement of flow and the associated signal processing.

Kinsler, L. E., Frey, A. R., Coppens, A. B., and Sanders, J. V., *Fundamentals of Acoustics*, 3rd edn., New York: John Wiley & Sons, 1982, a very readable introduction to acoustics.

Kremkau, F. W., *Diagnostic Ultrasound: Principles and Instruments*, 5th edn., Philadelphia, PA: W. B. Saunders Co., 1998, a very readable and up-to-date introduction to medical ultrasound.

Papadakis, E. P. (ed.), *Ultrasonic Instruments and Devices: Reference for Modern Instrumentations, Techniques, and Technology*, *in the series Physical Acoustics*, Vol. 40, New York: Academic Press, 1998.

Rose, J. L., *Ultrasonic Waves in Solid Media*, Cambridge, U.K.: Cambridge University Press, 1999.

Ultrasonic Scientific, Principles of the Ultrasonic Measurement. http://www.ultrasonic-scientific.com/Technology/Technology%20-%20principles.htm (accessed on July 27, 2012).



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

VI

Flow and Spot Velocity

50 Capillary-Type Mass Flow Meter *Reza Pakdaman Zangabad and Manouchehr Bahrami* **50-1**
Principles of Thermal Mass Flow Meters • Capillary-Type Thermal Mass Flow Meter • Electronic Readout Circuit • Accuracy • Summary • References

51 Differential Pressure Flowmeters *Richard Thorn* **51**-1
Important Principles of Fluid Flow in Pipes • Common Differential Pressure Flowmeters • Other Differential Pressure Flowmeters • Performance and Applications • Future Developments • Defining Terms • References • Further Information • Partial List of Vendors and Suppliers

52 Variable Area Flowmeters *Adrian Melling, Herbert Köchner, and Reinhard Haak* **52**-1
General Description of Variable Area Flowmeters • Measuring Principles of Variable Area Flowmeters • Summary • References • Further Readings • Partial List of Vendors and Suppliers

53 Positive Displacement Flowmeters *Zaki D. Husain and Donald J. Wass* **53**-1
Design and Construction • Some Commercially Available PD Meter Designs • Advantages and Disadvantages of PD Meters • Applications • Further Information • Partial List of Vendors and Suppliers

54 Turbine and Vane Flowmeters *David Wadlow* **54**-1
Axial Turbine Flowmeters • Insertion Axial Turbine Flowmeters • Angular Momentum Flowmeters • Multijet Turbine Flowmeters • Cylindrical Rotor Flowmeter • Manufacturers • References • Standards and Recommendations Applicable in the United States

55 Impeller Flowmeters *Harold M. Miller* **55**-1
Introduction • Sensing Principles • Flow Ranges • Installation and Applications • Manufacturers • Bibliography

56 Electromagnetic Flowmeters *Halit Eren* **56**-1
Faraday's Law of Induction • Construction and Operation of Electromagnetic Flowmeters • Types of Electromagnetic Flowmeters • Installations and Practical Applications • Signal Pickup and Demodulation Techniques • Bibliography • Partial List of Manufacturers and Suppliers

57 Ultrasonic Flowmeters *Hans-Peter Vaterlaus, Thomas Hossle, Paolo Giordano, and Christophe Bruttin*.....**57**-1
Introduction • Transit Time Flowmeter • Measurement of Flow in Open Channels • Application: Penstock Leak Detection with Surge Chambers • References • Further Readings

58 Vortex-Shedding Flowmeters *Wade M. Mattar and James H. Vignos*..... **58**-1
Introduction • Principle of Operation • Calculation of Mass Flow and Standard Volume • Flowmeter Construction • Application Considerations • Recent Developments • Further Information • Partial List of Vendors and Manufacturers

59 Thermal Anemometry *Jugal K. Agarwal and John G. Olin*..... **59**-1
General Description • Principle of Operation • Measurements • References

60 Coriolis Effect Mass Flowmeters *Jesse Yoder*..... **60**-1
Theory of Operation • Construction • Advantages • Disadvantages • Applications • A Look Ahead • Bibliography • Partial List of Vendors and Suppliers

61 Drag Force Flowmeters *Rekha Philip-Chandy, Roger Morgan, and Patricia J. Scully*..... **61**-1
Design of the Flow Sensor • Principle of Fluid Flow Measurement • Implementation Using Resistive Strain Gages • Optical Fiber Strain Gage Drag Force Flowmeter • Conclusion • References • Partial List of Vendors and Suppliers

62 Pitot Probe Anemometer *John A. Kleppe*..... **62**-1
Theory • The Pitot Tube in Flow with Variable Density • Volumetric Flow Measurements • A Hybrid System • Commercial Availability • References

63 Thermal Dispersion Mass Flow Meters *John G. Olin*..... **63**-1
Introduction • Description • Principle of Operation • Dry-Sensor Microprocessor-Based System • Flow Meter Selection, Installation, and Flow Calibration • References

64 Laser Anemometry *Rajan K. Menon*..... **64**-1
Principle of Operation • Signal Processing • Seeding and Other Aspects • Data Analysis • Extension to Particle Sizing • Phase Doppler System: Principle • Conclusion • Acknowledgments • References • Further Information

50

Capillary-Type Mass Flow Meter

50.1 Principles of Thermal Mass Flow Meters **50**-2
50.2 Capillary-Type Thermal Mass Flow Meter **50**-2
50.3 Electronic Readout Circuit **50**-8
Power Supply • Current Source • Sensor Assembly • Instrumentation Amplifier (AD522) • Analog-to-Digital Converter and Display
50.4 Accuracy **50**-12
50.5 Summary **50**-13
References **50**-13

Reza Pakdaman Zangabad
Sabanci University

Manouchehr Bahrami
University of Tabriz

In many processes, mass flow rate, rather than volume flow rate, is the desired parameter because volumetric flow measurements are less reliable than are mass flow measurements. Since changes in gas temperature and pressure will affect the density of a fixed volume of gas, thermal mass flow meters and controllers, unlike rotameters, turbine meters, and other volumetric flow devices, are relatively immune to changes in inlet temperature and pressure. Because these instruments directly measure molecular flow, they provide the most reliable, repeatable, and accurate method for delivering material from a supply volume to a process at a desired rate. Features, such as direct electronic readout, fast response, exceptional sensitivity at low flow rate, negligible pressure drop, no moving parts, simple installation, unobstructed straight through flow path, absence of temperature or pressure corrections, and superior retain accuracy over a wide range of flow rates, make thermal mass flow meters a superior alternative to conventional flow meters in many important applications. Some other advantages of this type of flow meter are accuracy in a 2% range, low flow sensitivity, reliability, and long life with advances in electronic technology; in a sense, these flow meters are getting smarter and more capable with microprocessor computing power, thus expanding the range of their usage.

The technology of the thermal mass flow meter was initially inspired by the space program's need for a reliable, low-power device to measure airflow in an astronaut's space suit. In recent years, they have become the most important instruments for gas flow measurement in the process industry, offering considerably lower cost than the Coriolis alternative. The only requirement for use of this type of flow meter is that the fluid being metered should be very clean and free from solid or liquid particles that may block the sensor passage. One of the big companies fabricating these sensors is "Sierra Instrumentation" [1]. Today, thermal mass flow meters are used to measure the flow of gases in a growing range of applications particularly in the electronic and chemical industries. The addition of an electromagnetic control valve turns the mass flow meter into a mass flow controller (MFC). Thermal mass flow meters are also employed for monitoring or controlling mass-related processes such as chemical reactions that depend on the relative masses of unreacted ingredients. They are thus widely used in automotive industry, utility services, petroleum and gas industries, HVAC, R & D, raw material industries, and food processing.

In short, thermal MFCs and mass flow meters for gases are among the standard instruments used in industrial laboratory and production processes.

Heat transfer is defined as the movement of energy due to a difference in temperature. It is characterized by the three mechanisms of conduction, convention, and radiation. Heat conduction takes place through different mechanisms in different media. Theoretically, it takes place in a gas through collisions of the molecules, in a fluid through oscillations of each molecule in a "cage" formed by its nearest neighbors, in metals mainly by electrons carrying heat, and in other solids by molecular motion that in crystals take the form of lattice vibrations. Typical for heat conduction is that the heat flux is proportional to the temperature gradient. Heat convection takes place through the net displacement of a fluid, which transports the heat content in a fluid through the fluid's own velocity. The term convection is also used for the heat dissipation from a solid surface to a fluid, where the heat transfer coefficient and the temperature difference across a fictitious film describe the flux. It is worthwhile to mention that both types measure fluid mass flow rate by means of the heat convected from a heated surface to the flowing fluid. The principles of operation of the two types are both thermal in nature but are so substantially different that two separate standards are required. Additionally, their applications are much different. Thermal dispersion flow meters are commonly used for general industrial gas flow applications in pipes and ducts, whereas capillary types are primarily used for smaller flows of clean gases or liquids in tubes.

50.1 Principles of Thermal Mass Flow Meters

The heat transfer scenario in the sensor tube consists of heat generation in the heating wire or over the tube wall, conduction along the tube wall, convection into the gas, and heat loss to the surroundings by natural convection, all the phenomena occurring simultaneously.

In general, there are three categories of thermal mass flow meters: insertion type, calorimetric sensors, and time-of-flight sensors. Thermal mass flow meters that measure the effect of the flowing fluid on a hot body are usually called insertion-type thermal mass flow meters. Those that measure the passage time of a heat pulse over a known distance are usually identified as time-of-flight sensors. Thermal mass flow meters that measure the displacement of temperature profile around the heater and also are modulated by the fluid flow are called calorimetric sensors.

Figure 50.1 shows the basic schematic of insertion type. As the figure demonstrates, a heater inside the pipeline heats the fluid to specific temperature that depends on the size of the pipeline. After reaching the desired temperature, the heating power is turned off, and consequently, the temperature will start to decrease. According to the speed of temperature decreasing in temperature profile diagram, the velocity of fluid can be realized. By this approach, the more rapid decrease, for the example, is due to the higher fluid velocity; the slower decrease correspondingly is due to lower fluid velocity.

In the time-of-flight type one, the working principle is as follows: by bringing the heater to a desired temperature and keeping it constant, fluid rate can be measured through sensing the temperature at a known distance away from heater. In this case, the high sensed temperature is due to the low fluid velocity, and the low sensed temperature is the result of high fluid velocity.

As Figure 50.2 depicts, by increasing fluid velocity, sensed temperature will be decreased. The applied heat power to heater is determined by characteristics of conduit and fluid. For example, increasing the temperature can result in better accuracy in conduits with bigger diameter.

50.2 Capillary-Type Thermal Mass Flow Meter

This type of mass flow meter can be classified as calorimetric sensors. In the capillary-type thermal mass flow meter, the sensor is a small diameter stainless steel or constantan capillary, heated by an electric current. The heat is supplied either at the central point or over the whole length of the tube. In the absence of gas flow, the temperature profile is symmetric about the midpoint. Gas flow through the capillary tube cools the entry section while heating the exit. Cooling and heating introduce an asymmetry

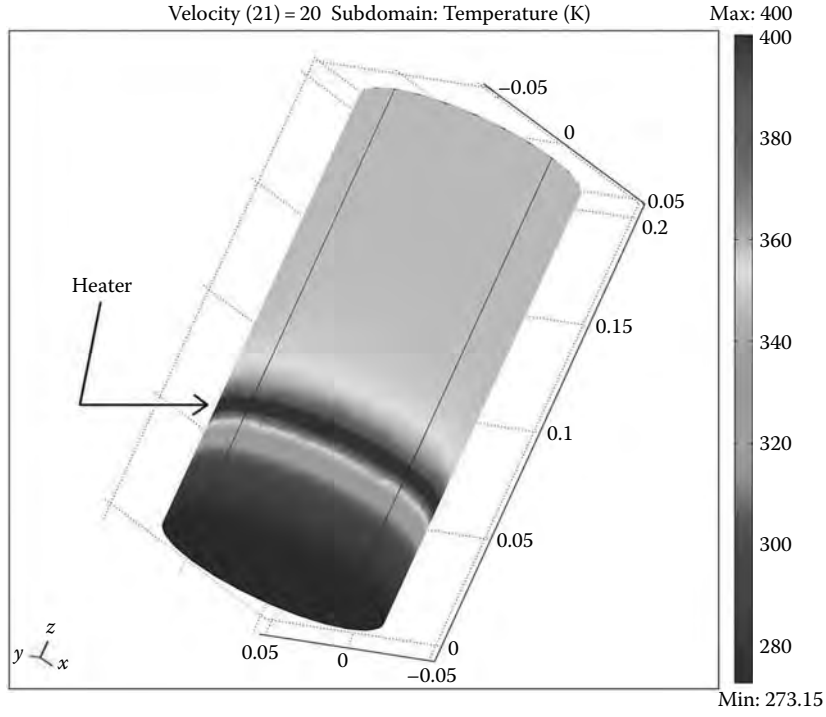


FIGURE 50.1 Basic schematic of thermal mass flow meter. Evenly heat distribution makes a linear relation between temperature and fluid velocity.

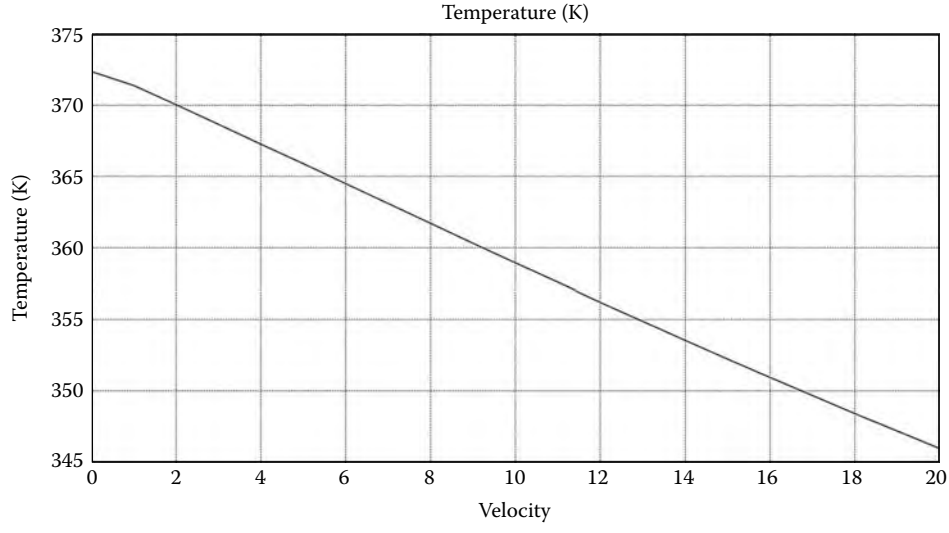


FIGURE 50.2 Linearly varying between sensed temperature and fluid velocity.

in the temperature profile, which is a measure of the fluid flow rate. The asymmetry is measured in terms of difference of temperature between two points on the capillary tube equidistant from the mid-point. A fluid flow shunt (bypass line) is added in parallel to the sensor tube to increase the range of the meter by several orders of magnitude. Flow through both the capillary and the bypass line is required to be laminar, often limited to the Reynolds number of 50–100 in the sensor (capillary) and 500–1500

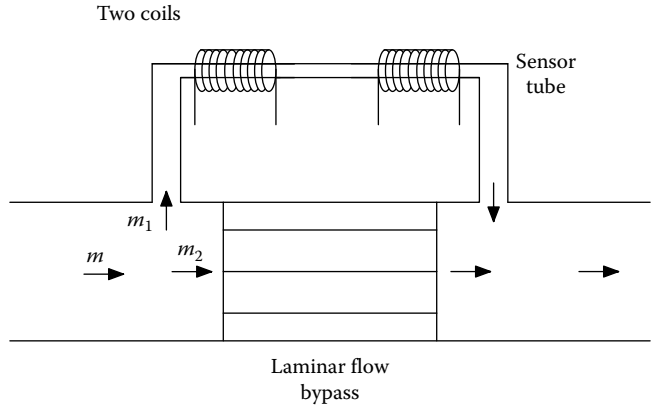


FIGURE 50.3 Structure of basic capillary thermal mass flow meter.

in the bypass line. By maintaining laminar flow within each portion, the split ratio between the sensor and the bypass flow element is maintained independent of the Reynolds number, and the total flow rate can be measured through the sensor path.

As gas from the source adding to the inlet of capillary mass flow meter, it goes through four stages and conditioning that eliminates all effects of flow disturbances and nonuniformities. Accurately achieving laminar flow is the key to assuring either linearity or accuracy. The laminar flow element (LFE) generates a small pressure drop, forcing a very small portion of gas flow through the stainless steel capillary tube where measurement occurs. Figure 50.3 shows a basic schematic of a capillary tube. The LFE plays an important role in capillary flow meter. It must be designed so that the process gas maintains a smooth, uniform flow profile called laminar flow (in other words, the gas has a Reynolds number below 2000). As a result, the small fraction of the total flow measured in the sensor tube (m_1) remains in direct proportion to the gas flowing through the LFE bypass (m_2). If we can measure the flow in the sensor tube, we can calculate the total mass flow rate m. Both m and m_1 are in mass flow rate units of g/s or kg/s:

$$m = m_1 + m_2 \tag{50.1}$$

$$m = m_1\left(1 + \frac{m_2}{m_1}\right) = km_1 \tag{50.2}$$

The sensor tube is precisely wrapped around with two platinum wire coils and heated to form pair resistance temperature detectors known as RTDs, and because they are on the outside of the tube, this sensing technology is noninvasive. When no gas flows, the capillary-tube temperature profile is symmetric around midpoint. It is hotter in the middle and cooler in the ends (Figure 50.4). As gas flows through the heated capillary tube, the gas molecules are also heated and carry the heat downstream; consequently, the temperature profile shifts downstream. This variation in the temperature profile affects the coil resistance. The two identical RTD coils are equally heated by connecting onto an electronic circuit. This method provides an easy method to measure any change in temperature by measuring the change in voltage of the circuit created by the RTD coils. The relative cooling of the upstream coil and the relative heating of downstream coil create a difference in the resistance of two RTDs. This difference in resistance is measured by electronics circuits and converted to the mass flow signal. This signal is directly proportional to the total mass flow.

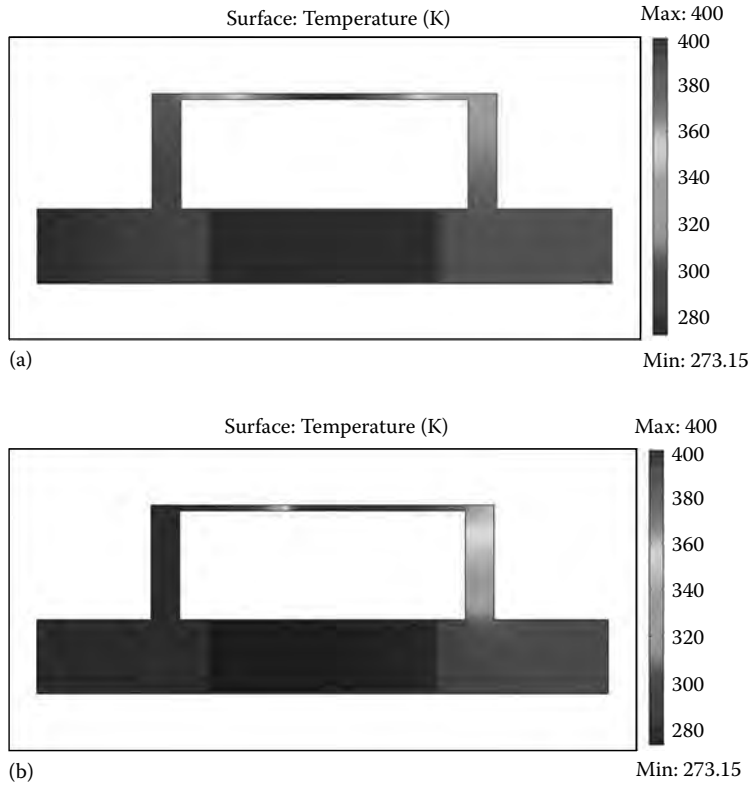


FIGURE 50.4 (a) Temperature profile of capillary zone while there is no fluid. (b) Temperature profile of capillary zone while a fluid passes through it.

Figure 50.4a offers a temperature profile when no fluid passes through the capillary tube and it demonstrates the midpoint's symmetry; Figure 50.4b depicts an exaggeration temperature profile, while a fluid passes through capillary and eventuates in shifting in the profile and creates asymmetric profiles [2].

As we know, heat transfer follows the first law of thermodynamics, commonly referred to as the principle of conservation of energy. Since the internal energy (***U***) is not a common quantity to measure, the basic law is usually rewritten in the term of temperature (***T***). So, the resulting heat equation used for measuring and simulating is written as

$$\rho C_p\left(\frac{\delta T}{\delta t}+(\mathbf{u}\cdot\nabla)T\right)=-(\nabla\cdot\mathbf{q})+\tau:\mathbf{S}-\frac{T}{\rho}\frac{\delta\rho}{\delta T}\bigg|_p\left(\frac{\delta p}{\delta t}+(\mathbf{u}\cdot\nabla)p\right)+Q \tag{50.3}$$

where

ρ is the density (kg/m^3)
C_p is the specific heat capacity at constant pressure (J/(kg·K))
T is absolute temperature (K)
u is the velocity vector (m/s)
q is the heat flux by conduction (W/m^2)
p is pressure (Pa)

S is the strain rate tensor (1/s) that defines as $\mathbf{S} = \frac{1}{2}\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^{\mathrm{T}}\right)$, τ is the viscous stress tensor (Pa) and represents as $\tau = \eta(2\mathbf{s} - (2/3)(\nabla \cdot \mathbf{u})I)$ with viscosity η (Pa·s), and I denotes the identity tensor. In Equation 50.3, Q contains heat sources other than viscous heating (W/m^3).

It is often appropriate to assume that the fluid density is almost constant, i.e., the given fluid is incompressible.

$$\rho\frac{\delta u}{\delta t} - \nabla\cdot\left[\eta\left(\nabla u + (\nabla u)^{\mathrm{T}}\right)\right] + \rho(u\cdot\nabla)u + \nabla p = F \tag{50.4}$$

$$\nabla\cdot\mathbf{u} = 0 \tag{50.5}$$

Equations 50.4 and 50.5 depict momentum equation and continuity condition for incompressible fluids, respectively. In the aforementioned equations, F is volume force and other variables are the same as defined previously.

Another important part of sensor that must be designed carefully in order to fulfill our expectations about achieving laminar flow and all the aforementioned characteristics is LFE. The flow sensor is made of a long, thin capillary tube; therefore, the flow through this sensor will be laminar. If the flow through the LFE is also laminar, then the ratio between the two flows can be considered constant for all flow rates. Therefore, the flow measured by the flow sensor is a ratio of the flow through the flow meter. Changing the capacity of the flow element will change the capacity of a flow meter. An example of such a flow element is given in Figure 50.5. The flow is forced through numerous tiny channels, which have a typical capacity of 10 mL$_n$/min (of air) at 35 mbar. The capacity of the flow element (and consequently of the flow meter) can be set by choosing the number of flow elements, each containing a specified number of channels. In this way, flow instruments can be manufactured from 10 mL$_n$/min up to 1000 L$_n$/min and even higher using single types of sensors.

Making the sensor tube very small will eventuate in restricting the mass flow rate through the tube (m_1); so by carefully designing of the LFE bypass, then we limit the solution of these equations to a

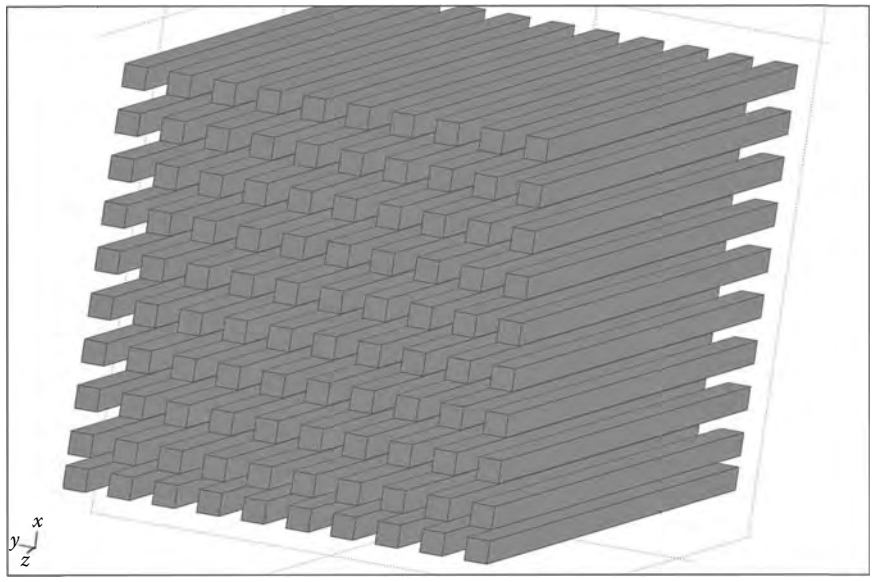


FIGURE 50.5 Basic structure of LFE in capillary-type thermal mass flow meter.

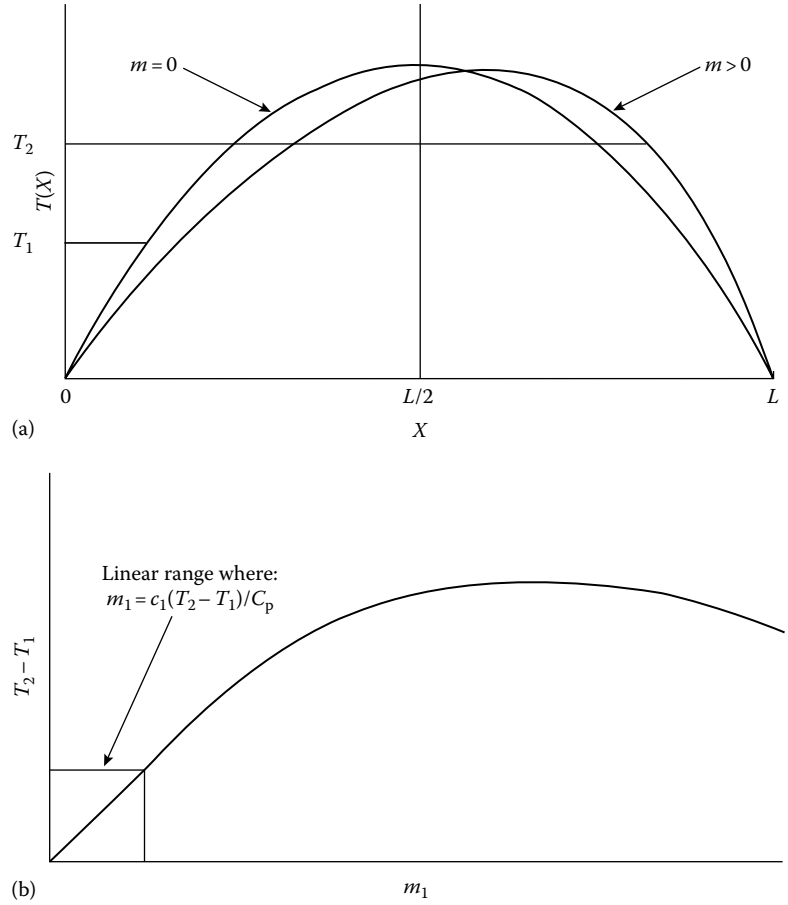


FIGURE 50.6 (a) Capillary-tube zone temperature distribution. (b) Capillary-tube sensor linear range for a given gas.

narrow range over which m_1 is directly proportional to $T_2 - T_1$, the difference in the average temperatures of the downstream coil (T_2) and the upstream coil (T_1).

In the linear range shown earlier in Figure 50.6a, we have for any given gas

$$m_1 C_p = C_1(T_2 - T_1) \tag{50.6}$$

where C_p and C_1 are the coefficient of specific heat for the given gas and a constant for the sensor applicable to all gases (in the linear range only), respectively.

From Figure 50.3, we have derived $m = km_1$. Substituting this into Equation 50.6 earlier, we get

$$mC_p = C(T_2 - T_1) \tag{50.7}$$

where $C = kC_1$, that is again a constant for sensor but applicable to all gases.

Equation 50.7 illustrates how the only parameter required in order to enable us to measure flow is temperature. Consequently, by measuring the upstream and downstream temperature, we will reach to our goal.

Based on the aforementioned discussion, the sensor was designed and simulated using COMSOL Multiphysics in order to verify the theory. Figure 50.7a shows the temperature distribution of sensor in capillary zone and Figure 50.7b shows the temperature profile.

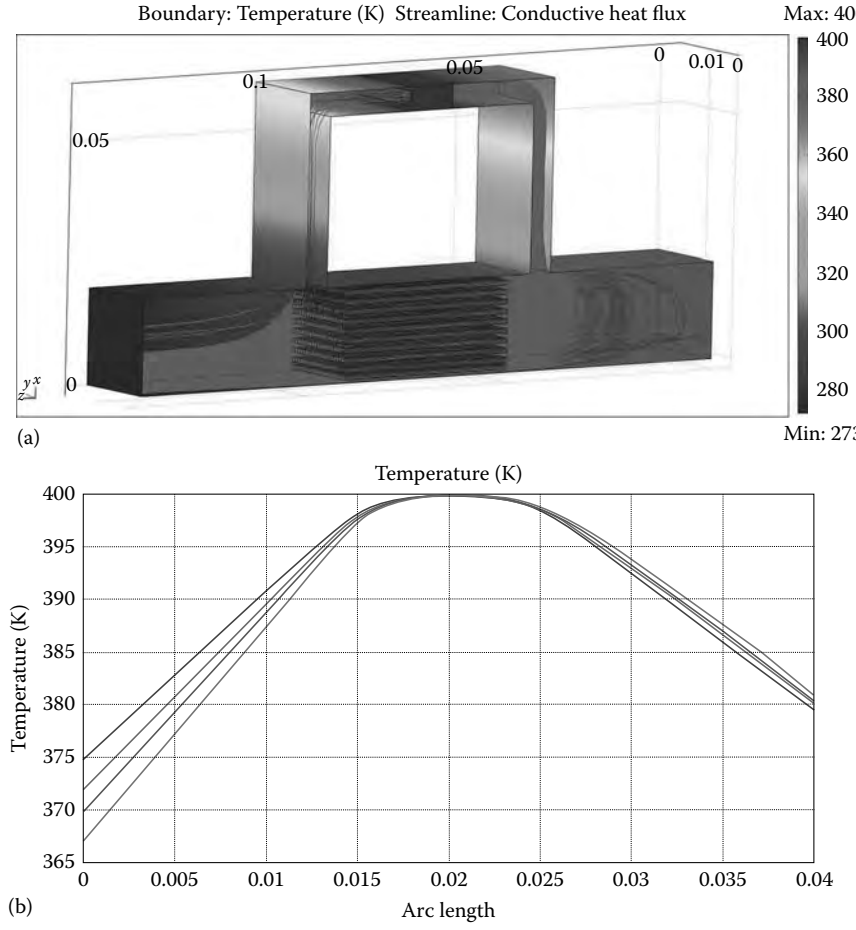


FIGURE 50.7 (a) Simulation result of temperature distribution over capillary zone in the sensor. (b) Simulated profile temperature for capillary-type thermal mass flow meter.

Figure 50.4a demonstrates the effect of LFE in order to have both laminar flows through the capillary and temperature distribution of capillary tube. As it appears, the temperature profile was shifted downstream (right-hand side), and consequently, the downstream temperature is a little higher than that of upstream. The temperature sensor that is located at those two points thus can measure the temperature. The basic and simple circuit for such sensing principle is the Wheatstone bridge, where in this sensing process, the temperature values are changed to resistor and then voltage values.

50.3 Electronic Readout Circuit

A major positive feature of the thermal flow meter is the electrical voltage output and direct digital readout proportional to the mass flow rate through the sensor that is ensured by providing an appropriate electronic circuitry. All the electronic components can be mounted on a PCB that is attached to the body of the flow meter. The main functions of the electronic circuitry are

a. To provide heating current to the sensor
b. To sense, amplify, and display the thermocouple output voltage

Figure 50.8 shows a block diagram of the electric circuit for sensing part.

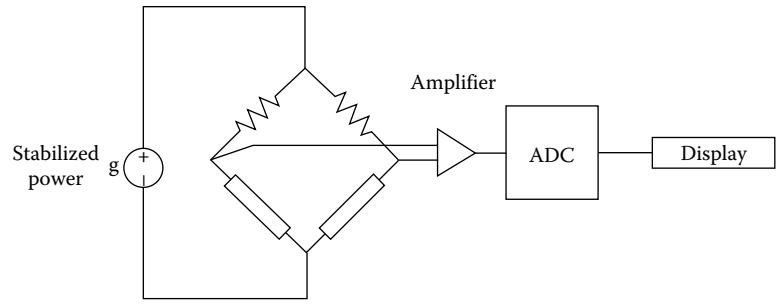


FIGURE 50.8 Basic schematic of electronic readout circuit.

The main sections of the electronic circuit are as follows:

1. DC power supply
2. Current source
3. Sensor assembly
4. Instrumentation amplifier (AD522)
5. Analog-to-digital converter (ADC) and display

50.3.1 Power Supply

For example, the power supply circuit (Figure 50.9) can be designed for an input voltage of 24 ± 10% V that is stabilized to 18 V DC. It can be built using an IC 7818–type voltage regulator, a few capacitors, and the diodes. The IC LM7818 is a three-terminal voltage regulator, which holds the load voltage constant even if the load current and the source voltage change. It needs a minimum input-to-output difference of 2 V, i.e., $(V_i - V_o) \geq 2$ V. For obtaining a regulated output of 18 V, the IC needs an input voltage between 20 and 37 V. A diode is used at the input for protection against accidental polarity reversal.

We have used a 1N4007 diode with a forward bias voltage drop of 0.7 V. Thus, 20.7 V is the minimum input voltage at which the regulator maintains 18 V at the output. We have specified the input voltage to be 24 ± 10% V, i.e., 21.6 to 26.4 V. The maximum output current of the IC is 500 mA. Actual dissipation in the proposed circuits has been estimated to be less than 200 mA. A small heat sink appropriate to the current requirement has been fitted with the IC. The power supply unit is fabricated on the same PCB.

50.3.2 Current Source

A high-frequency current source is used for heating the sensor tube. The heating element (sensor tube) is of very low resistance (50 mΩ) (taking into consideration that the two halves of the capillary tube come in parallel across the heating current source). An equivalent electric circuit of the sensor tube is shown in Figure 50.10.

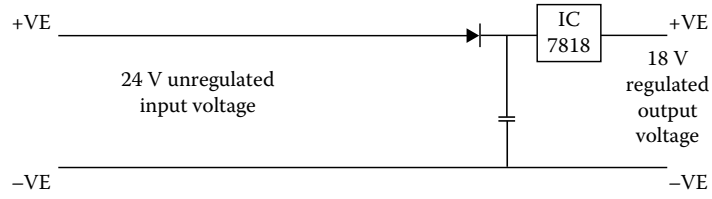


FIGURE 50.9 Circuit diagram of a regulated power supply.

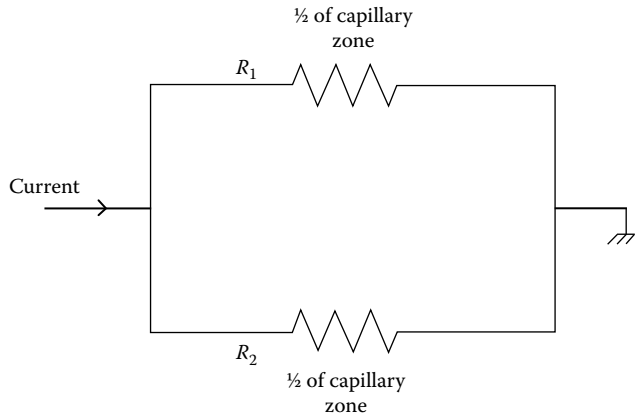


FIGURE 50.10 Equivalent electric circuit of the sensor tube.

For delivering a steady power of approximately 0.25 W, either alternating current (AC) or direct current (DC) could be used. But if a DC voltage is applied, it will be indistinguishable from the thermocouple voltage generated and therefore impossible to measure the temperature asymmetry over the sensor tube. Therefore, the solution to this problem is application of an AC voltage as the heating current so that its remnants can easily filter out from the thermal *emf* during measurement. The applied frequency depends on the transformer size and frequency-handling capability of the active devices (IC, transistor, etc.). In this case, one device can be PWM IC SG3525 as the oscillator with 10% dead time between commutations, as shown in Figure 50.11. A ferrite-ring transformer can be used to convert the power into low voltage and high current (e.g., 0.15 V RMS, 3A). This transformer also provides isolated voltage supply for other sections (amplifier, ADC, and display) through auxiliary windings and rectifiers. IC SG3525 is a pulse-width modulator (PWM) designed for use with switched-mode power supply (SMPS). Another useful feature of this chip is its control over the dead time. When a high frequency is used, self-impedance of the coil is high; hence, the device draws a rather small amount of current from the power supply.

By using high frequency, the audible noise effects from the transformer due to magnetostriction are also eliminated. At higher frequencies, the transformer size becomes smaller for the same power output, but at the same time, the electronic devices in the circuit consume more power. The two-phase output of the IC is buffered (isolated from the following circuits) through transistors (BC 547, BEL 187, BEL 188) to drive the primary winding of the toroidal transformer (ferrite core). The prime output of the transformer in the thick secondary winding is 0.15 V RMS and can provide 3 A of current. This output is used to heat the sensor. Two more auxiliary windings provide isolated voltages of ±5 and 9 V for the amplifier and meter sections.

50.3.3 Sensor Assembly

The heating current source is connected to the center of the capillary tube, and the return path of the current is through the brass body of the sensor assembly (see Figure 50.12). Thus, the sensor tube is split into two equal parts, and the current through each half of the tube becomes half of the current supplied by the source. The potential difference between the points A and B (which are located at equal distance from the midpoint C) vanishes all the time, and the thermal emf generated by temperature asymmetry can be comfortably measured. The small alternating potential difference caused by fabrication errors and resulting current asymmetry can be easily filtered.

Constantan wires are brazed at optimized locations (A and B) to make the constantan-SS constantan differential thermocouple. Connections (*Va*, *Vb*, and *Vc*) from the thermocouples and the midpoint are terminated on a small PCB at the base of the sensor body.

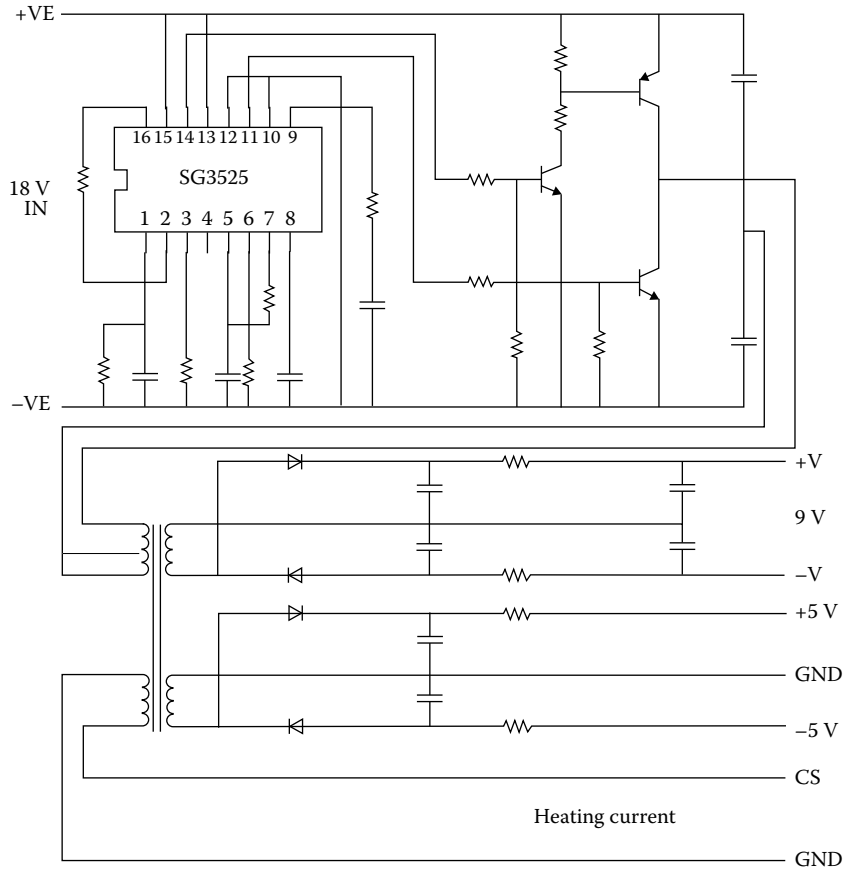


FIGURE 50.11 Proposed current source and power supply.

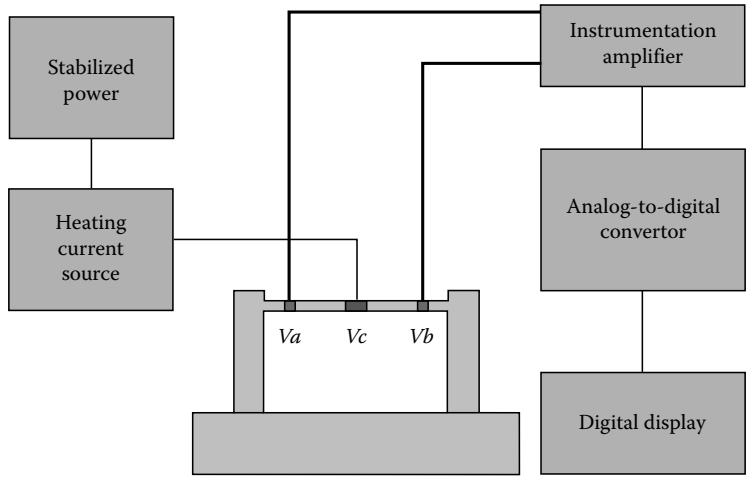


FIGURE 50.12 Basic topology of sensor assembling.

50.3.4 Instrumentation Amplifier (AD522)

The AD522 is a precision IC instrumentation amplifier designed for data acquisition applications requiring high accuracy under worst-case operating conditions. An outstanding combination of high linearity, high common-mode rejection, low voltage drift, and low noise makes the AD522 suitable for use in many 12-bit data acquisition systems. An instrumentation amplifier is usually employed as a bridge amplifier for resistance transducers (thermistors, strain gages, etc.) found in process control, instrumentation, data processing, and medical testing. The operating environment is frequently characterized by low signal-to-noise levels, fluctuating temperatures, imbalanced input impedances, and remote location, all of which hinder recalibration.

50.3.5 Analog-to-Digital Converter and Display

An ADC must be used to convert the analog output of the differential thermocouple into a more sophisticated and easily readable digital display. An inbuilt three-and-half-digit LCD panel meter is one of the devices to cover both ADC and display.

The input sensitivity of the meter is 200 mV DC. The output display gives a direct readout for the flow rate. The digital display cannot exceed the three-and-half digits 1999. Hence, it can read a maximum flow rate of 19.99 slpm. If the flow rate exceeds 19.99 slpm, the rightmost three digits will be blank and only the left hand "1" will appear on the display, indicating an "over-range" condition. After the over-range condition has been removed, it may take several seconds for the flow meter to recover and resume normal operation. This will not harm the instrument.

50.4 Accuracy

The following method is designed to determine the precision and accuracy of the prototype thermal mass flow meter. The experimental setup is employed in series with a known mass flow meter. After the installation of the experimental unit and covering the electronic warm-up period, test loop is purged with clean, dry nitrogen gas for a minimum of 5 min. The flow rate is maintained at the full-scale reading of the prototype flow meter. Once the setup is ready, the flow rate is varied over full range of the instrument in discrete steps in both increasing and decreasing orders. The readings on the test and standard instruments are noted down and analyzed for precision and accuracy.

The bias in a set of readings is defined as the average difference between the readings of the prototype flow meter and a standard flow meter:

$$B = \frac{\sum_{i=1}^{n} (m_i - m_{s,i})}{n} \tag{50.8}$$

where

m_i is flow rate reading in the prototype flow meter
$m_{s,i}$ is flow rate reading in the standard flow meter

The precision of a reading is usually defined in terms of the standard deviation of a set of readings corresponding to the same true value. Extending the concept a little further, the precision is defined in terms of the standard deviation of the difference in readings between the prototype and the standard flow meter. In other words,

$$P = \sqrt{\frac{\sum_{i=1}^{n} \left[(m_i - m_{s,i}) - B\right]^2}{n}} \tag{50.9}$$

where n is total number of readings at flow rates close to the specified value.

Accuracy is defined in terms of bias B and precision P as

$$A = \left(P + |B|\right)\frac{B}{|B|} \tag{50.10}$$

Accuracy has the same sign as bias B and is larger than B by the value of precision P.

50.5 Summary

The capillary-type thermal mass flow meter has been studied in the beginning of this chapter in the discussion of the basic and various parts of a sensor and, more importantly, how these parts affect the sensor characteristics. Then a very basic readout circuit is introduced in order to measure the mass flow; at the end, measuring the accuracy of prototype sensor is introduced.

According to Equation 50.2, by measuring the amount of flow that passes through the capillary zone, the entire flow can be measured. According to the temperature profile of a capillary tube that depicts the shift in the profile according to flow velocity variation, the temperature is measured from two points of upstream and downstream; consequently, Equation 50.6 will supply the amount of total mass flow rate. Here, the most critical factor is measuring the temperature, which is discussed earlier. As mentioned previously, the constants of k, C_1, and C are related to sensors and can be found while measuring the accuracy of prototype sensor.

After precisely measuring the mass flow rate, the accurately control of the mass flow rate is required in most applications. A unit known as mass flow controller consists of a valve in which the data come from the capillary mass flow meter and control the valve in order to achieve a desired mass flow rate.

References

1. Products, SIERRA Instruments, http://www.sierrainstruments.com/ (accessed on March 15, 2013).
2. Pakdaman Zangabad, R., Bahrami, M., Chamanian, S., et al., Design and simulation of a novel capillary-type thermal mass flow meter, *Proceedings of 2012 IEEE Sensors Applications Symposium, SAS 2012*, Brescia, Italy, pp. 6–9, 2012.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

51

Differential Pressure Flowmeters

51.1 Important Principles of Fluid Flow in Pipes **51**-2
Bernoulli's Equation
51.2 Common Differential Pressure Flowmeters **51**-4
Orifice Plate • Venturi Tube • Nozzle
51.3 Other Differential Pressure Flowmeters **51**-7
51.4 Performance and Applications **51**-7
Installation • Differential Pressure Measurement • Standards
51.5 Future Developments **51**-11
Defining Terms **51**-11
References **51**-11
Further Information **51**-12
Partial List of Vendors and Suppliers **51**-12

Richard Thorn
University of Derby

Flow measurement is an everyday event. Whether you are filling up a car with petrol (gasoline) or wanting to know how much water the garden sprinkler is consuming, a flowmeter is required. Similarly, it is also difficult to think of a sector of industry in which a flowmeter of one type or another does not play a part. The world market in flowmeters is expected to grow steadily for the foreseeable future. The value of product being measured by these meters is also very large. For example, in the United Kingdom alone, it was estimated that in 1994, the value of crude oil produced was worth $15 billion.

Given the size of the flowmeter market, and the value of product being measured, it is somewhat surprising that both the accuracy and the capability of many flowmeters are poor in comparison to those instruments used for measurement of other common process variables such as pressure and temperature. For example, the orifice plate flowmeter, which was first used commercially in the early 1900s and has a typical accuracy of ±2% of reading, is still the only flowmeter approved by most countries for the fiscal measurement of natural gas. Although newer techniques such as Coriolis flowmeters have become increasingly popular in recent years, the flow measurement industry is by nature conservative and still dominated by traditional measurement techniques. For a review of recent flowmeter developments, refer to [1].

Over 40% of all liquid, gas, and steam measurements made in industry are still accomplished using common types of *differential pressure flowmeters*, such as the orifice plate, venturi tube, and nozzle. The operation of these flowmeters is based on the observation made by Bernoulli that if an annular restriction is placed in a pipeline, then the velocity of the fluid through the restriction is increased. The increase in velocity at the restriction causes the static pressure to decrease at this section, and a pressure difference is created across the element. The difference between the pressure upstream and the pressure downstream of this obstruction is related to the rate of fluid flowing through the restriction and therefore through the pipe. A differential pressure flowmeter consists of two basic elements: an

obstruction to cause a pressure drop in the flow (a *differential producer*) and a method of measuring the pressure drop across this obstruction (a *differential pressure transducer*).

One of the major advantages of the orifice plate, venturi tube, or nozzle is that the measurement uncertainty can be predicted without the need for calibration, if it is manufactured and installed in accordance with one of the international standards covering these devices. In addition, this type of differential pressure flowmeter is simple, has no moving parts, and is therefore reliable. The main disadvantages of these devices are their limited range (typically 3:1), the permanent pressure drop they produce in the pipeline (which can result in higher pumping costs), and their sensitivity to installation effects (which can be minimized using straight lengths of pipe before and after the flowmeter). The combined advantages of this type of flowmeter are still quite hard to beat, and although it has limitations, these have been well investigated and can be compensated for in most circumstances. Unless very high accuracy is required, or unless the application makes a nonintrusive device essential, the differential flowmeter should be considered. Despite the predictions of its demise, there is a little doubt that the differential pressure flowmeter will remain a common method of flow measurement for many years to come.

51.1 Important Principles of Fluid Flow in Pipes

There are a number of important principles relating to the flow of fluid in a pipe that should be understood before a differential pressure flowmeter can be used with confidence. The principles are as follows: the difference between laminar and turbulent flows, the meaning of Reynolds number, and the importance of the flow's velocity profile.

Fluid motion in a pipe can be characterized as one of the three types: laminar, transitional, or turbulent. In *laminar flow*, the fluid travels as parallel layers (known as streamlines) that do not mix as they move in the direction of the flow. If the flow is turbulent, the fluid does not travel in parallel layers, but moves in a haphazard manner with only the average motion of the fluid being parallel to the axis of the pipe. If the flow is *transitional*, then both types may be present at different points along the pipeline, or the flow may switch between the two.

In 1883, Osborne Reynolds performed a classic set of experiments at the University of Manchester that showed that the flow characteristic can be predicted using a dimensionless number, now known as the Reynolds number. The Reynolds number Re is the ratio of the inertia forces in the flow ($\rho\bar{v}D$) to the viscous forces in the flow (η) and can be calculated using

$$\mathrm{Re} = \frac{\rho\bar{v}D}{\eta} \tag{51.1}$$

where

ρ is the density of the fluid
$\bar{v}$ is the mean velocity of the fluid
D is the pipe diameter
η is the dynamic viscosity of the fluid

If Re is less than 2000, viscous forces in the flow dominate and the flow will be laminar. If Re is greater than 4000, inertia forces in the flow dominate and the flow will be turbulent. If Re is between 2000 and 4000, the flow is transitional and either mode can be present. The Reynolds number is calculated using mainly properties of the fluid and does not take into account factors such as pipe roughness, bends, and valves that also affect the flow characteristic. Nevertheless, the Reynolds number is a good guide to the type of flow that can be expected in most situations.

The fluid velocity across a pipe cross section is not constant and depends on the type of flow present. In laminar flow, the velocity profile is parabolic since the large viscous forces present cause the fluid to move more slowly near the pipe walls. Under these conditions, the velocity at the center of the pipe is twice the

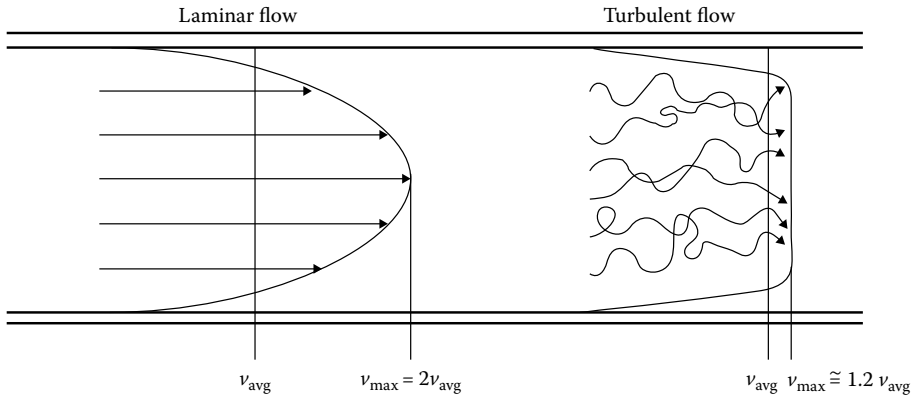


FIGURE 51.1 Velocity profiles in laminar and turbulent flow.

average velocity across the pipe cross section. The laminar-flow profile is unaffected by the roughness of the pipe wall. In turbulent flow, inertia forces dominate, pipe wall effects are less, and flow's velocity profile is flatter, with the velocity at the center being about 1.2 times the mean velocity. The exact flow profile in a turbulent flow depends on pipe wall roughness and Reynolds number. Figure 51.1 shows the "fully developed" flow profiles for laminar and turbulent flow. These are the flow profiles that would be obtained at the end of a very long pipe, thus ensuring that any changes to the flow profile due to pipe bends and fittings are no longer present. To have confidence in the performance of a differential pressure flowmeter, both the characteristic and the velocity profile of the flow passing through the flowmeter should be stable and known.

51.1.1 Bernoulli's Equation

The Bernoulli's equation defines the relationship between fluid velocity (v), fluid pressure (p), and height (h) above some fixed point for a fluid flowing through a pipe of varying cross section and is the starting point for understanding the principle of the differential pressure flowmeter. For the inclined, tapered pipe shown in Figure 51.2, Bernoulli's equation states that

$$\frac{p_1}{\rho g} + \frac{v_1^2}{2g} + h_1 = \frac{p_2}{\rho g} + \frac{v_2^2}{2g} + h_2 \tag{51.2}$$

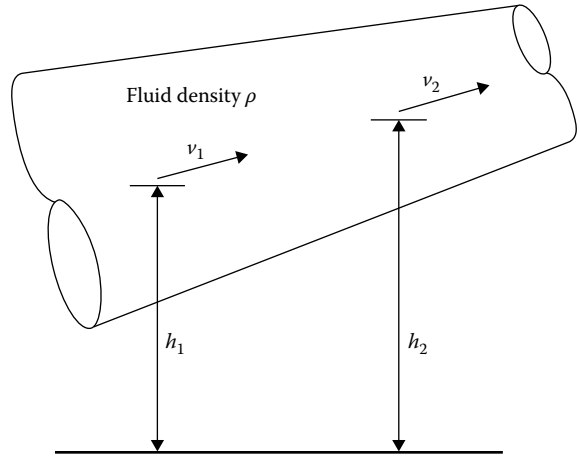


FIGURE 51.2 Flow through an inclined, tapered pipe.

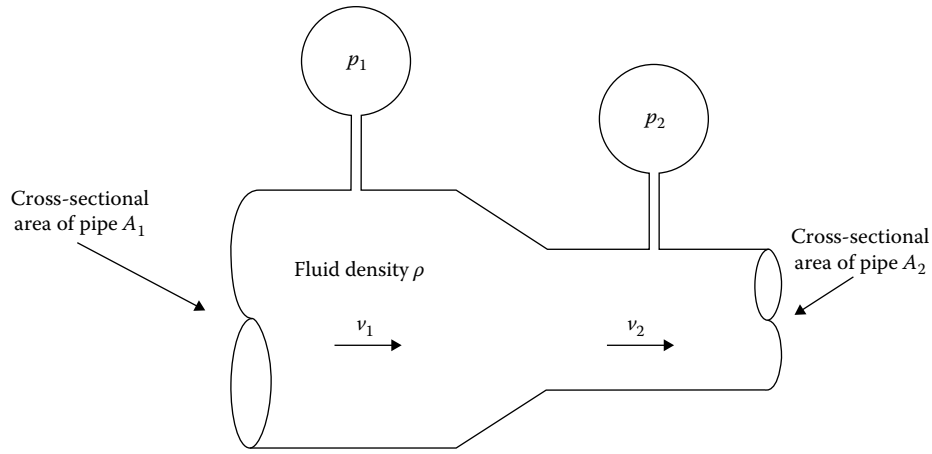


FIGURE 51.3 Using a restriction in a pipe to measure fluid flow rate.

Thus, the sum of the pressure head ($p/\rho g$), the velocity head ($v/2g$), and the potential head (h) is constant along a flow streamline. The term "head" is commonly used because each of these terms has the unit of meters. Equation 51.2 assumes that the fluid is frictionless (zero viscosity) and of constant density (incompressible). Further details on the derivation and significance of Bernoulli's equation can be found in most undergraduate fluid dynamics textbooks (e.g., [2]).

Bernoulli's equation can be used to show how a restriction in a pipe can be used to measure flow rate. Consider the pipe section shown in Figure 51.3. Since the pipe is horizontal, $h_1 = h_2$, and Equation 51.2 reduces to

$$\frac{p_1 - p_2}{\rho} = \frac{v_1^2 - v_2^2}{2} \tag{51.3}$$

The conservation of mass principle requires that

$$v_1 A_1 \rho = v_2 A_2 \rho \tag{51.4}$$

Rearranging Equation 51.4 and substituting for v_2 in Equation 51.3 gives

$$Q = v_1 A_1 = \frac{A_2}{\sqrt{1 - (A_2/A_1)^2}} \sqrt{\frac{2(p_1 - p_2)}{\rho}} \tag{51.5}$$

This shows that the volumetric flow rate of fluid Q can be determined by measuring the drop in pressure ($p_1 - p_2$) across the restriction in the pipeline—the basic principle of all differential pressure flowmeters. Equation 51.5 has limitations, the main ones being that it is assumed that the fluid is incompressible (a reasonable assumption for most liquids) and that the fluid has no viscosity (resulting in a flat velocity profile). These assumptions need to be compensated for when Equation 51.5 is used for practical flow measurement.

51.2 Common Differential Pressure Flowmeters

51.2.1 Orifice Plate

The orifice plate is the simplest and cheapest type of differential pressure flowmeter. It is simply a plate with a hole of specified size and position cut in it, which can then be clamped between flanges in a pipeline (Figure 51.4). The increase that occurs in the velocity of a fluid as it passes through the hole in

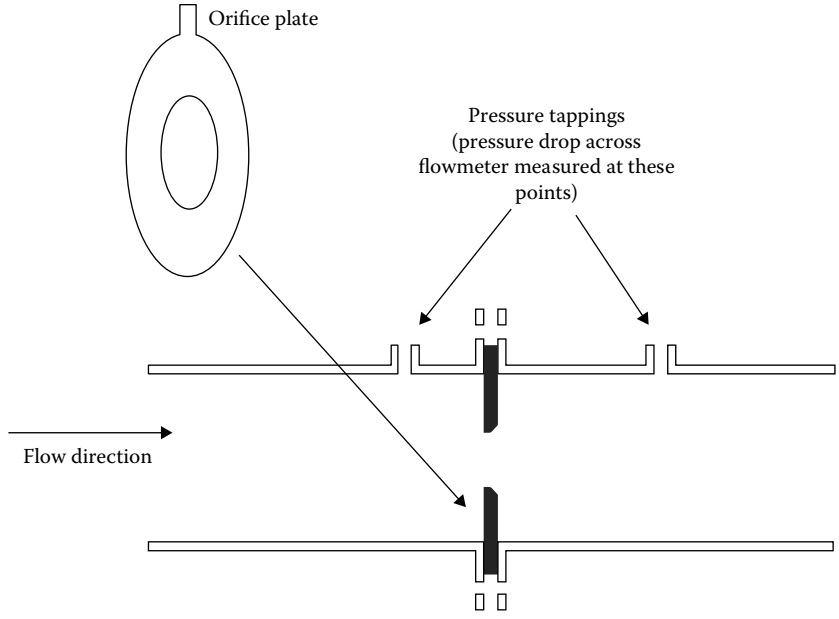


FIGURE 51.4 Square-edged orifice plate flowmeter.

the plate results in a pressure drop being developed across the plate. After passing through this restriction, the fluid flow jet continues to contract until a minimum diameter known as the vena contracta is reached. If Equation 51.5 is used to calculate volumetric flow rate from a measurement of the pressure drop across the orifice plate, then an error would result. This is because A_2 should strictly be the area of the vena contracta, which of course is unknown. In addition, turbulence between the vena contracta and the pipe wall results in an energy loss that is not accounted for in this equation.

To overcome the problems caused by the practical application of Equation 51.5, two empirically determined correction factors are added. After some reorganization, Equation 51.5 can be written as

$$Q = \frac{C}{\sqrt{1-\beta^4}} \varepsilon \frac{\pi}{4} d^2 \sqrt{\frac{2(p_1 - p_2)}{\rho}} \tag{51.6}$$

where

ρ is the density of the fluid upstream of the orifice plate
d is the diameter of the hole in the orifice plate
β = is the diameter ratio d/D, where D is the upstream internal pipe diameter

The two empirically determined correction factors are C the discharge coefficient and ε the expansibility factor. C is affected by the changes in the diameter ratio, the Reynolds number, the pipe roughness, the sharpness of the leading edge of the orifice, and the points at which the differential pressure across the plate is measured. However, for a fixed geometry, it has been shown that C is only dependent on Reynolds number, and so this coefficient can be determined for a particular application; ε is used to account for the compressibility of the fluid being monitored. Both C and ε can be determined from equations and tables in a number of internationally recognized documents known as standards. These standards specify not only C and ε but also the geometry and installation conditions for the square-edged orifice plate, venturi tube, and nozzle and are essentially a design guide for the use of the most commonly used types of differential pressure flowmeter. Installation recommendations are intended

to ensure that fully developed turbulent flow conditions exist within the measurement section of the flowmeter. The most commonly used standard in Europe is ISO 5167-1 [3], while in the United States, API 2530 is the most popular [4]. There are differences between some of the recommendations in these two standards (e.g., the minimum recommended length of straight pipe upstream of the flowmeter), but work is under way to resolve these.

Equation 51.6 illustrates perhaps the greatest strength of the orifice plate, which is that measurement performance can be confidently predicted without the need for calibration if the device is manufactured, installed, and operated in accordance with one of the international standards. In addition, the device is cheap to manufacture, has no moving parts, is reliable, and can be used for metering most clean gases, liquids, and steam.

The major disadvantages of the orifice plate are its limited range and sensitivity to flow disturbances. The fact that fluid flow rate is proportional to the square root of the measured differential pressure limits the range of a one-plate/one-differential-pressure transmitter combination to about 3:1. The required diameter ratio (also known as beta ratio) of the plate depends on the maximum flow rate to be measured and the range of the differential pressure transducer available. Sizing of the orifice plate is covered in most of the books in the further reading list, and nowadays, computer programs are also available to help perform this task. The flow measurement range can be increased by switching; the ways in which this may be achieved are described in [5]. Equation 51.6 assumes a fully developed and stable flow profile, and so installation of the device is critical, particularly the need for sufficient straight pipework upstream of the meter. Wear of the leading edge of the orifice plate can severely alter measurement accuracy; thus, this device is normally only used with clean fluids.

Only one type of orifice plate, the square-edged concentric, is covered by the standards. However, other types exist, having been designed for specific applications. One example is the eccentric orifice plate, which is suited for use with dirty fluids. The details of these other types of orifice plate can be found in [6].

51.2.2 Venturi Tube

The classical or Herschel venturi tube is the oldest type of differential pressure flowmeter, having first been used in 1887. As Figure 51.5 shows, a restriction is introduced into the flow in a more gradual way than for the orifice plate. The resulting flow through a venturi tube is closer to that predicted in theory by Equation 51.5, and so the discharge coefficient C is much nearer unity, being typically 0.95. In addition, the permanent pressure loss caused by the venturi tube is lower, but the differential pressure is also lower than for an orifice plate of the same diameter ratio. The smooth design of the venturi tube means that it is less sensitive to erosion than the orifice plate and thus more suitable for use with dirty gases or liquids. The venturi tube is also less sensitive to upstream disturbances and therefore needs shorter lengths of straight pipework upstream of the meter than the equivalent orifice plate or nozzle. Like the orifice plate and nozzle, the design, installation, and use of the venturi tube is covered by a number of international standards.

The major disadvantages of the venturi tube flowmeter are its size and cost. It is more difficult and therefore more expensive to manufacture than the orifice plate. Since a venturi tube can be typically 6

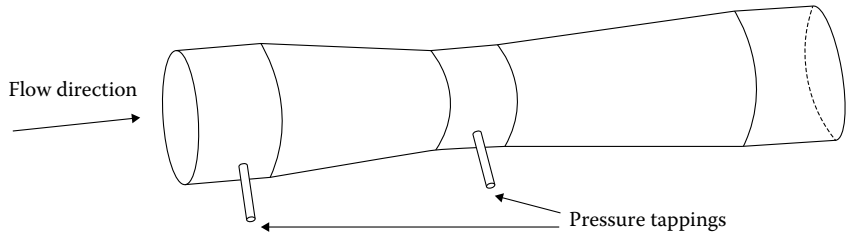


FIGURE 51.5 Venturi tube flowmeter.

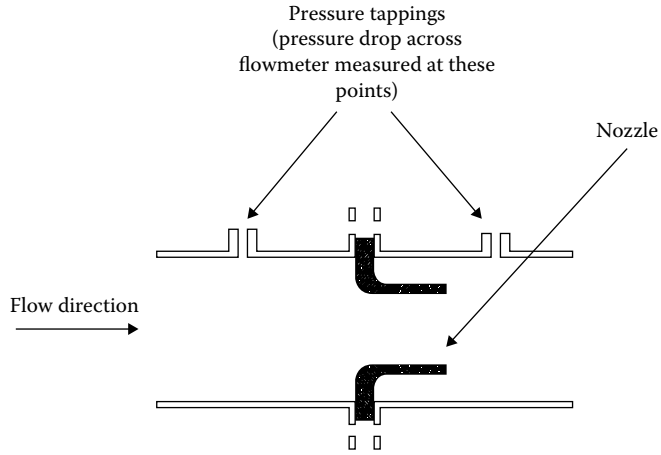


FIGURE 51.6 Nozzle flowmeter.

diameters long, it can become cumbersome to use with larger pipe sizes, with associated maintenance of upstream and downstream pipe lengths also becoming a problem.

51.2.3 Nozzle

The nozzle (Figure 51.6) combines some of the best features of the orifice plate and venturi tube. It is compact and yet, because of its curved inlet, has a discharge coefficient close to unity. There are a number of designs of nozzle, but one of the most commonly used in Europe is the ISA-1932 nozzle, while in the United States, the ASME long-radius nozzle is more popular. Both of these nozzles are covered by international standards.

The smooth inlet of the nozzle means that it is more expensive to manufacture than the orifice plate as the curvature of the inlet changes with diameter ratio, although it is cheaper than the venturi tube. The device has no sharp edges to erode and causes changes in calibration and thus is well suited for use with dirty and abrasive fluids. The nozzle is also commonly used for high-velocity, high-temperature applications such as steam metering.

A variation of the nozzle is the sonic (or critical flow venturi) nozzle, which has been used both as a calibration standard for testing gas meters and as a transfer standard in interlaboratory comparisons [7].

51.3 Other Differential Pressure Flowmeters

There are many other types of differential pressure flowmeter, including the segmental wedge, V-cone, elbow, and Dall tube. Each of these has advantages over the orifice plate, venturi tube, and nozzle for specific applications. For example, the segmental wedge can be used with flows having a low Reynolds number, and a Dall tube has a lower permanent pressure loss than a venturi tube. However, none of these instruments are yet covered by international standards, and thus, calibration is needed to determine their accuracy. Further information on these, and other less-common types of differential pressure flowmeter, can be found in [8].

51.4 Performance and Applications

Table 51.1 shows the performance characteristics and main application areas of the square-edged orifice plate, venturi tube, and nozzle flowmeters. Compared to other types of flowmeters on the market, these differential pressure flowmeters only have moderate accuracy, typically ±2% of reading; but of course,

TABLE 51.1 Performance and Application Areas of Common Differential Pressure Flowmeters

	Performance					Applications				
	Typical Uncalibrated Accuracy (%)	Typical Range	Typical Pipe Diameter (mm)	Permanent Pressure Loss	Comparative Cost	Clean Gas	Dirty Gas	Clean Liquid	Slurry	Steam
Orifice plate	±2	3:1	10–1000	High	Low	Yes	No	Yes	No	Yes
Venturi tube	±2	3:1	25–500	Low	High	Yes	Maybe	Yes	Maybe	Maybe
Nozzle	±2	3:1	25–250	High	Medium	Yes	Maybe	Yes	No	Yes

this can be improved if the device is calibrated after installation. Although in some circumstances, these flowmeters can be used with dirty gases or liquids, usually only small amounts of a second component can be tolerated before large measurement errors occur. When calculating the cost and performance of a differential flowmeter, both the primary element and the differential pressure transducer should be taken into account. Although the orifice plate is the cheapest of the primary elements, the cost of the fitting needed to mount it in the pipeline, particularly if online removal is required, can be significant.

Choosing which flowmeter is best for a particular application can be very difficult. The main factors that influence this choice are the required performance, the properties of the fluid to be metered, the installation requirements, the environment in which the instrument is to be used, and, of course, the cost. There are two standards that can be used to help select a flowmeter: BS 1042 (Section 1.4), which is a guide to the use of the standard differential pressure flowmeters [9], and BS 7405, which is concerned with the wider principles of flowmeter selection [10].

Because all three flowmeters have similar accuracy, one strategy for selecting the most appropriate instrument is to decide if there are any good reasons for not using the cheapest flowmeter that can be used over the widest range of pipe sizes: the orifice plate. Where permanent pressure loss is important, the venturi tube should be considered, although the high cost of this meter can only usually be justified where large quantities of fluid are being metered. For high-temperature or high-velocity applications, the nozzle should be considered because under these conditions, it is more predictable than the orifice plate. For metering dirty fluids, either the venturi tube or the nozzle should be considered in preference to the orifice plate, the choice between the two depending on cost and pressure loss requirements. Table 51.2 lists some suppliers of differential pressure flowmeters.

51.4.1 Installation

Correct installation is essential for successful use of a differential pressure flowmeter because the predicted uncertainty in the flow rate/differential pressure relationship in Equation 51.6 assumes a steady flow, with a fully developed turbulent velocity profile, which is passing through the flowmeter. Standards contain detailed recommendations for the minimum straight lengths of pipe required before and after the flowmeter, in order to ensure a fully developed flow profile. Straight lengths of pipe are required after the flowmeter because disturbances caused by a valve or bend can travel upstream and thus also affect the installed flowmeter. Table 51.3 gives examples of installation requirements taken from ISO 5167-1. If it is not possible to fit the recommended lengths of straight pipe before and after the flowmeter, then the flowmeter must be calibrated once it has been installed.

The other major problem one faces during installation is the presence of a rotating flow or swirl. This condition distorts the flow velocity profile in a very unpredictable way and is obviously not desirable. Situations that create swirl, such as two 90° bends in different planes, should preferably be avoided. However, if this is not possible, then swirl can be removed by placing a flow conditioner (also known as a flow straightener) between the source of the swirl and the flowmeter. There are a wide range of flow conditioner designs, some of which can be used to both remove swirl and correct a distorted velocity

TABLE 51.2 Selection of Companies That Supply Differential Pressure Flowmeters

ABB Kent-Taylor
Oldens Lane, Stonehouse
Gloucestershire, GL10 3TA
England
Tel.: +44 1453 826661
Fax: +44 1453 826358

Daniel Industries Inc.
9720 Katy Road
P.O. Box 19097
Houston, TX 77224
Tel.: (713) 467-6000
Fax: (713) 827-3880

Hartmann & Braun (U.K.) Ltd.
Bush Beach Engineering Division
Stanley Green Trading Estate, Cheadle Hulme
Cheshire SK8 6RN
England
Tel.: +44 161 4858151
Fax: +44 161 4884048

ISA Controls Ltd.
Hackworth Industrial Park
Shildon
County Durham DL4 1LH
England
Tel.: +44 1388 773065
Fax: +44 1388 774888

Perry Equipment Corporation
Wolters Industrial Park
P.O. Box 640
Mineral Wells, TX 76067
Tel.: (817) 325-2575
Fax: (817) 325-4622

TABLE 51.3 Minimum Straight Lengths of Pipe Required between Various Fittings and an Orifice Plate or Venturi Tube (as Recommended in ISO 5167-1) to Ensure That a Fully Developed Flow Profile Exists in the Measurement Section

	Upstream of the Flowmeter				Downstream of the Flowmeter
Diameter Ratio β	Single 90° Bend	Two 90° Bends in the Same Plane	Two 90° Bends in Different Planes	Globe Valve Fully Open	For Any of the Fittings Shown to the Left
0.2	10	14	34	18	4
0.4	14	18	36	20	6
0.6	18	26	48	26	7
0.8	46	50	80	44	8

Note: All lengths are multiples of the pipe diameter.

profile [11]. Because they obstruct the flow, all flow conditioners produce an unrecoverable pressure loss, which in general increases with their capability (and complexity).

51.4.2 Differential Pressure Measurement

Apart from the differential producer, the other main element of a differential pressure flowmeter is the *transducer* needed to measure the pressure drop across the producer. The correct selection and installation of the differential pressure transducer plays an important part in determining the accuracy of the flow rate measurement.

The main factors that should be considered when choosing a differential pressure transducer for a flow measurement application are the differential pressure range to be covered, the accuracy required, the maximum pipeline pressure, and the type and temperature range of the fluid being metered.

Most modern differential pressure transducers consist of a pressure capsule in which either capacitance, strain gage, or resonant wire techniques are used to detect the movement of a diaphragm.

Using these techniques, a typical accuracy of ±0.1% of full scale is possible. The transducer is usually part of a unit known as a transmitter, which converts differential pressure, static pressure, and ambient temperature measurements into a standardized analog or digital output signal. "Smart" transmitters use a local, dedicated microprocessor to condition signals from the individual sensors and compute volumetric or mass flow rate. These devices can be remotely configured, and a wide range of diagnostic and maintenance functions are possible using their built-in "intelligence."

As far as installation is concerned, the transmitter should be located as close to the differential producer as possible. This helps to ensure a fast dynamic response and reduces problems caused by vibration of the connecting tubes. The position of the pressure tappings is also important. If the liquid flow in a horizontal pipe is being measured, then the pressure tappings should be located at the side of the pipe so that they cannot be blocked with dirt or filled with air bubbles. For horizontal gas flows, if the gas is clean, the pressure tappings should be vertical; if steam or dirty gas is being metered, then the tappings should be located at the side of the pipe. These general guidelines show that considerable care must be taken with the installation of the differential pressure transmitter if large measurement errors are to be avoided. For further details on the installation of differential pressure transmitters, see ISO 2186 [12].

51.4.3 Standards

International standards that specify the design, installation, and use of the orifice plate, venturi tube, and nozzle and allow their accuracy to be calculated without the need for calibration are probably the main reason for the continuing use of this type of flowmeter. Table 51.4 gives details of the most common standards related to differential pressure flowmeters. There are still inconsistencies in the various standards. For example, ISO 5167-1 states that any flow conditioner should be preceded by at least 22 diameters of straight pipe and followed by at least 20 diameters of

TABLE 51.4 Standards Related to Differential Pressure Flow Measurement

American National Standards Institute, New York	
ANSI/ASHRAE 41.8	Standard methods of measurement of flow of liquids in pipes using orifice flowmeters
ANSI/ASME MFC-7M	Measurement of gas flow by means of critical flow venturi nozzles
ANSI/ASME MFC-14M	Measurement of fluid flow using small-bore precision orifice meters
American Petroleum Institute, Washington, DC	
API 2530	Manual of Petroleum Measurement Standards, Chapter 14—Natural gas fluids measurement, Section 3—Orifice metering of natural gas and other related hydrocarbon fluids
American Society of Mechanical Engineers, New York	
ASME MFC-3M	Measurement of fluid flow in pipes using orifice, nozzle, and venturi
ASME MFG-8M	Fluid flow in closed conduits—connections for pressure signal transmissions between primary and secondary devices
British Standards Institution, London	
BS 1042	Measurement of fluid flow in closed conduits
International Organization for Standardization, Geneva	
ISO 2186	Fluid flow in closed conduits—connections for pressure transmissions between primary and secondary elements
ISO TR 3313	Measurement of pulsating fluid flow in a pipe by means of orifice plates, nozzles, or venturi tubes
ISO 5167-1	Measurement of fluid flow by means of pressure differential devices
ISO 9300	Measurement of gas flow by means of critical flow venturi nozzles

straight pipe. This would seem to contradict one application of a flow conditioner, which is to reduce the length of straight pipe required upstream of a flowmeter.

Despite the occasional inconsistency and difference between the standards, they are the internationally accepted rules for the installation and use of the square-edged orifice plate, venturi tube, and nozzle.

51.5 Future Developments

In spite of the vast amount of published data available on differential pressure flowmeters, continued research is needed to improve the understanding of the effect of flow conditions, and flowmeter geometry, on the uncalibrated accuracy of these devices. For example, work has been recently undertaken to derive an improved equation for the discharge coefficient of an orifice plate [13].

The metering of multiphase flow is an area of increasing importance. In addition to developing new measurement techniques, many people are investigating ways in which traditional flowmeters can be used to meter multiphase flows. A good review of the use of differential pressure flowmeters for multiphase flow measurement can be found in [14].

The development of "smart" differential pressure transmitters has overcome the limitations of differential pressure flowmeters in some applications. For example, these devices are being used to linearize and extend the range of differential pressure flowmeters.

The aforementioned developments should help to ensure the continued popularity of the differential pressure flowmeter for the foreseeable future, despite increasing competition from newer types of instrument.

Defining Terms

Differential pressure flowmeter: A flowmeter in which the pressure drop across an annular restriction placed in the pipeline is used to measure fluid flow rate. The most common types use an orifice plate, venturi tube, or nozzle as the primary device.

Differential pressure transmitter: Secondary device that measures the differential pressure across the primary device and converts it into an electrical signal.

Nozzle: Primary device consisting of a convergent inlet connected to a cylindrical section.

Orifice plate: Primary device consisting of a thin plate in which a circular aperture has been cut.

Venturi tube: Primary device consisting of a converging inlet, cylindrical mid-section, and diverging outlet.

References

1. R. A. Furness, Flowmetering: evolution or revolution, *Meas. Control*, 27 (8), 15–18, 1994.
2. B. S. Massey, *Mechanics of Fluids*, 6th ed., London, U.K.: Chapman and Hall, 1989.
3. International Organization for Standardization, ISO 5167-1, *Measurement of Fluid Flow by Means of Pressure Differential Devices—Part 1 Orifice Plates, Nozzles and Venturi Tubes Inserted in Circular Cross-Section Conduits Running Full*, Geneva, Switzerland: International Organization for Standardization, 1991.
4. American Petroleum Institute, API 2530, *Manual of Petroleum Measurement Standards Chapter 14—Natural Gas Fluids Measurement, Section 3—Orifice Metering of Natural Gas and Other Related Hydrocarbon Fluids*, Washington, DC: American Petroleum Institute, 1985.
5. E. L. Upp, *Fluid Flow Measurement*, Houston, TX: Gulf Publishing, 1993.
6. H. S. Bean, *Fluid Meters Their Theory and Application*, 6th ed., New York: American Society of Mechanical Engineers, 1983.

7. P. H. Wright, The application of sonic (critical flow) nozzles in the gas industry, *Flow Meas. Instrum*., 4 (2), 67–71, 1993.
8. D. W. Spitzer, *Industrial Flow Measurement*, 2nd ed., Research Triangle Park, NC: ISA, 1990.
9. British Standards Institution, BS 1042, *Measurement of Fluid Flow in Closed Conduits—Part 1 Pressure Differential Devices—Section 1.4 Guide to the Use of Devices specified in Sections 1.1 and 1.2*, London, U.K.: British Standards Institution, 1992.
10. British Standards Institution, BS7405, *Guide to the Selection and Application of Flowmeters for Measurement of Fluid Flow in Closed Conduits*, London, U.K.: British Standards Institution, 1991.
11. E. M. Laws and A. K. Ouazzane, Compact installations for differential pressure flow measurement, *Flow Meas. Instrum*., 5 (2), 79–85, 1994.
12. International Organization for Standardization, ISO 2186, *Fluid Flow in Closed Conduits—Connections for Pressure Signal Transmissions between Primary and Secondary Elements*, Geneva, Switzerland: International Organization for Standardization, 1973.
13. M. J. Reader-Harris, J. A. Slattery, and E. P. Spearman, The orifice plate discharge coefficient equation—further work, *Flow. Meas. Instrum*., 6 (2), 101–114, 1995.
14. F. C. Kinghorn, Two-phase flow measurement using differential pressure meters, *Multi-Phase Flow Measurement Short Course*, London, U.K., June 17–18, 1985.

Further Information

Baker, R. C., *An Introductory Guide to Flow Measurement*, London, U.K.: Mechanical Engineering Publications, 1989, a good pocket sized guide on the choice and use of flowmeters commonly used in industry.

OMEGA Complete Flow and Level Measurement handbook and Encyclopedia, OMEGA Press, 1995. Also available at http://www.omega.com/literature/transactions/volume4/T9904-07-DIFF.html (accessed on June 27, 2012).

Flow Measurement and Instrumentation, Editor-in-Chief: J. Delsing, Oxford, U.K.: Butterworth-Heinemann, a quarterly journal covering all aspects of flowmeters and their applications. A good source of information on current research activity.

Hayward, T. J., *Flowmeters—A Basic Guide and Source-Book for Users*, London, U.K.: Macmillan, 1979, an overview of the important areas of flow measurement which is a joy to read.

Miller, R. W., *Flow Measurement Engineering Handbook*, 3rd ed., New York: McGraw-Hill, 1996, a thorough reference book particularly on differential pressure flowmeters, covers European and U.S. standards with calculations using both U.S. and SI units.

Spitzer, D. W., *Flow Measurement: Practical Guides for Measurement and Control*, Research Triangle Park, NC: ISA, 1991, intended for practicing engineers, this book covers most aspects of industrial flow measurement.

Upp, E. L., *Fluid Flow Measurement*, Houston, TX: Gulf Publishing, 1993, contains a lot of practical advice on the use of differential pressure flowmeters.

Universal Flow Monitors, Differential pressure flowmeter, Hazel Park, MI. http://www.flowmeters.com/ufm/index.cfm?task=differential_pressure (accessed on June 21, 2012).

Partial List of Vendors and Suppliers

Furness Controls, 2020 Younts Rd., Indian Trail, NC 28079, Phone: 704-882-3311, 800-898-5325 (toll free), Fax: 704-882-3811, www.furnesscontrols.com/

Lambda Square, Inc., 71 Deer Park Ave., Babylon, NY 11702, Phone: 631-587-1000, 800-587-5423 (toll free), Fax: 631-587-1011, www.lambdasquare.com

Orange Research, Inc., 140 Cascade Blvd., Milford, CT 06460, Phone: 203-877-5657, 800-989-5657 (toll free), Fax: 203-783-9546, www.orangeresearch.com

Paragon Controls, Inc., 2371 Circadian Way, Santa Rosa, CA 95407, Phone: 707-579-1424, Fax: 707-579-8480, www.paragoncontrols.com

Sensocon, Inc., PO Box 518, Highland City, FL 33846, Phone: 863-248-2800, Fax: 863-248-2798, www.ensocon.com

Sierra Instruments, 5 Harris Ct., Bldg. L, Monterey, CA 93940, Phone: 831-373-0200, 800-866-0200 (toll free), Fax: 831-373-4402, www.sierrainstruments.com



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

52

Variable Area Flowmeters

Adrian Melling
University of Erlangen–Nuernberg

Herbert Köchner
University of Erlangen–Nuernberg

Reinhard Haak
University of Erlangen–Nuernberg

52.1 General Description of Variable Area Flowmeters **52**-1
Rotameter • Movable Vane Meter • Weir and Flume
52.2 Measuring Principles of Variable Area Flowmeters..... **52**-8
Rotameter • Weir and Flume • Rotameter: Internal Flow Analysis
52.3 Summary..... **52**-18
References..... **52**-18
Further Readings..... **52**-19
Partial List of Vendors and Suppliers..... **52**-19

The term *variable area flowmeters* refers to those meters in which the minimum cross-sectional area available to the flow through the meter varies with the flow rate. Meters of this type that are discussed in this section include the rotameter and the movable vane meter used in pipe flows and the weir or flume used in open-channel flows. The measure of the flow rate is a geometrical quantity such as the height of a bob in the rotameter, the angle of the vane, or the change in height of the free surface of the liquid flowing over the weir or through the flume.

Most of the discussion here is devoted to the rotameter, firstly because the number of installed rotameters and movable vane meters is large relative to the number of weirs and flumes and secondly because the movable vane is often used simply as a flow indicator rather than as a meter.

The following section includes basic information describing the main constructional features and applications of each type of meter. In the third section, the principles of measurement and the design of rotameters and open-channel meters are described in some detail; the movable vane meter is not considered further because most design aspects are similar to those of rotameters. Then, the contribution of modern computational and experimental methods of fluid mechanics to flowmeter design is discussed, using the results of a detailed investigation of the internal flow.

The details of rotameters and movable vane meters, together with approximate costs of these meters, are tabulated in Tables 52.1 and Table 52.2. A partial list of manufacturers and suppliers is provided at the end of this chapter.

52.1 General Description of Variable Area Flowmeters

52.1.1 Rotameter

The *rotameter* is a robust and simple flowmeter for gases and liquids and holds a large share of the market for pipe diameters smaller than about 100 mm. In its basic form, the rotameter consists of a conical transparent vertical glass tube containing a "bob" (Figure 52.1), which rises in the tube with increasing flow rate until a balance is reached between gravitational, buoyancy, and drag forces on the bob. Within

TABLE 52.1 Approximate Rotameter Prices

Type	Size	Price (Plastic)	Price (Stainless Steel, Borosilicate Glass)
Glass	<1/2 in. (12.7 mm)	$50	$200
Plastic	1/2 in. (12.7 mm)	$50	$200
	1 in. (25.4 mm)	$70	$150
	2 in. (50.8 mm)	$200	$500
Metal	1/2 in. (12.7 mm)		$500
	1 in. (25.4 mm)		$600
	2 in. (50.8 mm)		$700
	3 in. (76.2 mm)		$1100
	4 in. (101.6 mm)		$1400
Additional costs			
Electric limit switch	$100		
Current transducer	$500		
Pneumatic	$1200		

TABLE 52.2 Approximate Movable Vane Meter Prices

Type	Price
Plastic versions (small scale)	From $10
Direct-coupled pointer versions incl. electric flow switch: metal, screw connections 1/2 in. (12.7 mm) to 2 in. (50.8 mm)	$400 (brass) to $600 (Stainless steel)
Flanged versions	Add $200 (1/2 in., 12.7 mm) to $400 (2 in., 50.8 mm)
Magnetic-coupled indicator, stainless steel	$1200 (1/2 in., 12.7 mm) to $2100 (8 in., 203 mm)

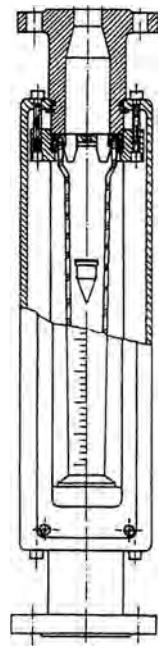


FIGURE 52.1 Cross section of a rotameter. The level of the bob rises linearly with increasing flow rate.

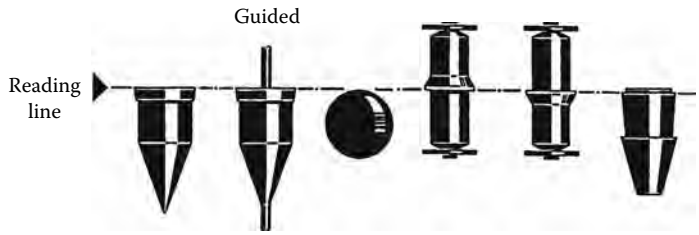


FIGURE 52.2 Typical rotameter bob geometries.

the range of a particular flowmeter (depending on the bob shape and density, the tube shape, and the fluid density and viscosity), the flow rate is linearly proportional to the height of the bob in the tube and is determined simply by reading the level of the upper edge of the bob. The rotameter is also known as "floating element flowmeter," although the buoyancy force on the "floating element" is not sufficient to make it float.

The bob is commonly formed of a combination of cylindrical and conical sections, but a spherical bob is often used in small diameter tubes (see Figures 52.2). More complicated bob geometries can reduce the sensitivity to viscosity of the fluid. Frequently, the bob has several shallow inclined grooves around the upper rim that induce a slow rotation (frequency about 1 Hz) of the bob, which helps to maintain a stable position of the bob. In larger rotameters where the additional friction is acceptable, the bob can be allowed to slide up and down a rod on the tube axis to prevent any sideways motion.

Various tube geometries are in use. The basic requirement for an increase in the cross-sectional area up the height of the tube leads to the conical tube. An alternative form uses three ribs arranged circumferentially around the tube to guide the bob; the flow area between the ribs increases along the tube height. For spherical bobs, a triangular tube in which the bob remains in contact with three surfaces over the whole height of the tube prevents lateral movement of the bob with minimum friction. Commonly, the tube is made of glass to facilitate the reading of the flow rate from a scale engraved on the tube. For general-purpose applications, the scale can be marked in millimeters; the flow rate is then determined by a conversion factor depending on the tube dimensions, the mass of the bob, the pressure and temperature, and the properties of the fluid. For specific application to a single fluid under controlled conditions, it is more convenient to have the flow rate directly marked on the tube.

For laboratory use, medical equipment, and other applications with small flow rates, the glass tube is almost universal. Laboratory meters are often equipped with a flow valve that allows direct-flow setting and reading. For many small rotameters, integrated flow controllers are offered; these hold the pressure drop and the flow setting constant, even with changing upstream or downstream pressure.

For most rotameters, electric limit switches or analog electric signal transducers are available. For glass and plastic meters, inductive or optical switches are used that can be positioned to the desired level of the bob. While some units just detect the bob within their active area, other switches have bistable operation, wherein the switch is toggled by the float passing the switch.

For industrial applications, the metering of corrosive fluids or fluids at high temperature or pressure leads to the use of stainless steel tubes or a wide range of other materials chosen to suit the application requirements (e.g., temperature, pressure, corrosion). Most metal rotameters use a permanent magnetic coupling between the float and the pointer of the indicator, enabling a direct analog flow reading without electric supply. Electric switches can, however, be added to the indicator, and electric and pneumatic transmitters for the flow reading are offered by most producers. Some units are equipped with options such as flow totalizers, mass flow calculation units for gas applications in conjunction with temperature and pressure sensors, or energy calculation units. Due to the frequent application of the rotameter in the chemical and petrochemical industry, electric options are often designed for use in hazardous areas.

The rotameter is characterized by

- Simple and robust construction
- High reliability
- Low pressure drop

Applicable to a wide variety of gases and liquids
Flow range typically 0.04 L h^{-1}–150 m^3 h^{-1} for water
Flow range typically 0.5 L h^{-1}–3000 m^3 h^{-1} for air
10:1 flow range for given bob-tube combination
Uncertainty 0.4%–4% of maximum flow
Insensitivity to nonuniformity in the inflow (no upstream straight piping needed)
Typical maximum temperature 400 °C
Typical maximum pressure 4 MPa (40 bar)

- Low investment cost
- Low installation cost

52.1.2 Movable Vane Meter

The movable vane meter is a robust device suitable for the measurement of high flow rates where only moderate requirements on the measurement accuracy are made. Dirty fluids can also be metered. It contains a flap that at zero flow is held closed by a weight or a spring (Figure 52.3). A flow forces the vane open until the dynamic force of the flow is in balance with the restoring force of the weight or the spring. The angle of the vane is thus a measure of the flow rate, which can be directly indicated by a pointer attached to the shaft of the vane on a calibrated scale. The resistance provided by the vane depends on the vane position and hence on the flow rate or Reynolds number; a recalibration is therefore necessary when the fluid is changed. An important application is the metering of the airflow in automotive engines with fuel injection.

In low-cost flow indicators, a glass or plastic window allows a direct view of the flap (Figure 52.4). Sometimes, optical or reed switches in combination with a permanent magnet attached to the flap are used as electric flow switches. All-metal units use a magnetic coupling between the vane and the pointer of the indicator, thus avoiding most of the friction and material selection problems. Many applications

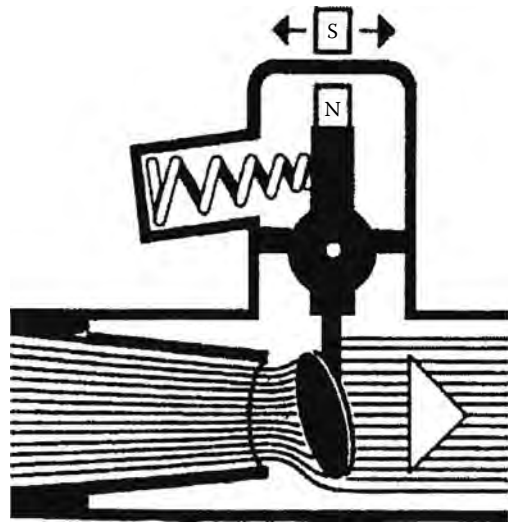


FIGURE 52.3 Movable vane meter. The magnet (N,S) transmits the vane position to an indicator.

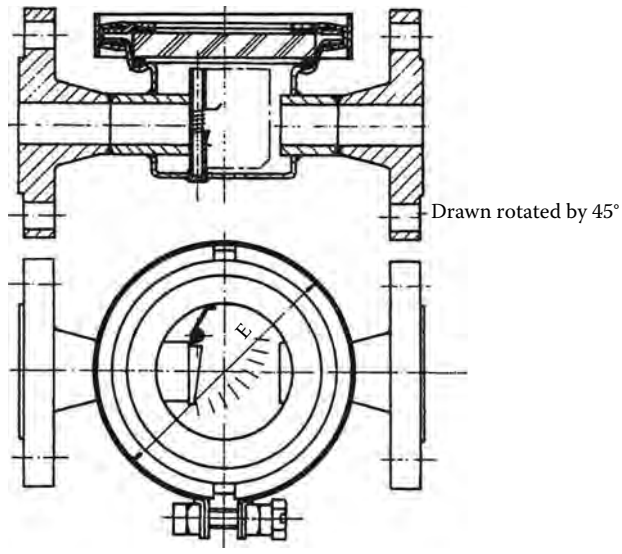


FIGURE 52.4 Flow indicator. The angle of the vane provides a measure of the flow rate.

use the movable vane meter as a flow switch with appropriate mechanical, magnetic, inductive, or optical switches. The setting of the switch is done once for all during the installation. Since a continuous flow indication is not needed, flow reading uncertainties are not of concern.

Most aspects of design and application of rotameters can be applied to movable vane meters, although the latter are characterized by a higher uncertainty of the flow reading. The general features of both types of variable area flowmeter are summarized, for example, in Ref. [1]. Since the basic construction of the two meter types is very similar, the unit costs also lie in the same range.

52.1.3 Weir and Flume

Of the methods available for the metering of a liquid (generally water) in open-channel flow, the weir and the flume fall within the scope of this discussion on variable area flowmeters. In each case, the flowmetering depends on measurement of the difference in height h of the water surface over an obstruction across the channel and the surface sufficiently far upstream. There is a wide variety of geometries in use (see Ref. [2]), but most of these can be described as variants of three basic types. In the sharp-crested weir, also known as the thin-plate weir (Figure 52.5), the sill or crest is only about 1–2 mm thick. The sheet of liquid flowing over the weir, called the nappe or vein, separates from the weir body after passing over the crest. An air-filled zone at atmospheric pressure is formed underneath the outflowing jet, and the streamlines above the weir are strongly curved. When the width of the weir is less than that of the upstream channel (Figures 52.6 and 52.7), the term "notch" is frequently used. A broad-crested weir (Figure 52.8) has a sill that is long enough for straight, parallel streamlines to form above the crest. In this review, the term "weir" is applied generally to both the weir and the notch.

The resistance to flow introduced by the weir causes the water-level upstream to rise. If it is assumed that the flow velocity, some distance upstream of the weir, is zero, then a simple measure of the upstream water level suffices to determine the discharge over a weir of known geometry. As in the case of an orifice in pipe flow, there will be a contraction of the nappe and a frictional resistance at the sides as water flows over a weir. The actual discharge is less than the theoretical discharge according to an empirically determined coefficient of discharge C_d. In the analysis of rectangular notches and weirs, it is frequently

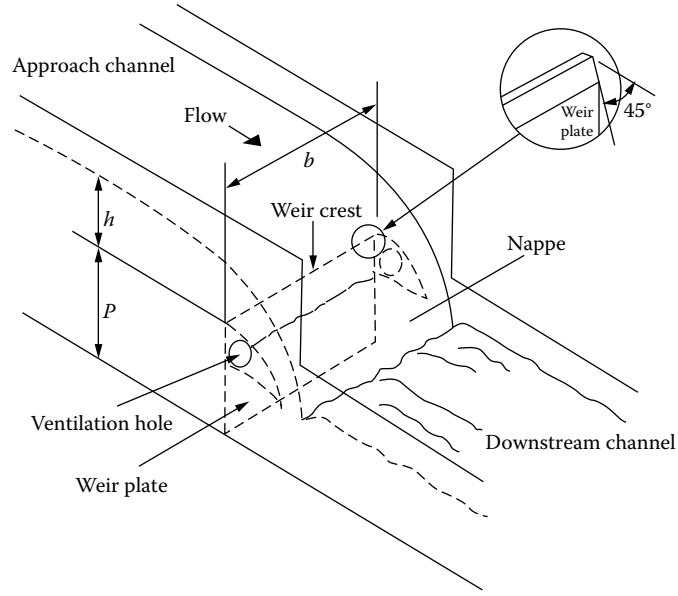


FIGURE 52.5 Full-width thin-plate weir. Flow rate $\dot{Q}$ varies with water level h above the weir crest ($\dot{Q} - h^{3/2}$) and with a flow-dependent discharge coefficient C_d.

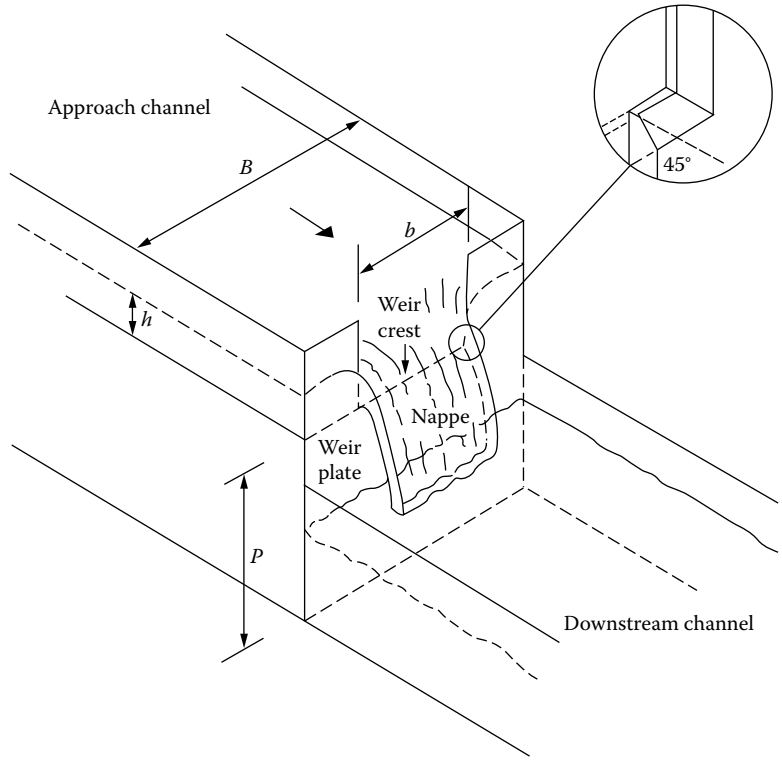


FIGURE 52.6 Thin-plate rectangular notch weir: discharge coefficient C_d is flow dependent.

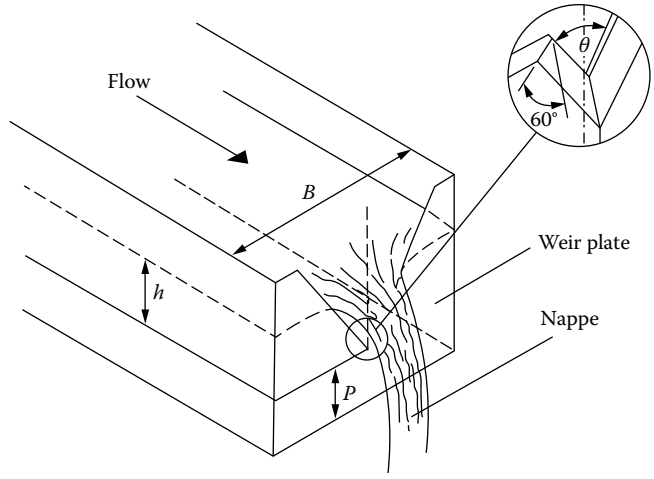


FIGURE 52.7 Thin-plate V-notch weir: discharge coefficient C_d is almost independent of the flow.

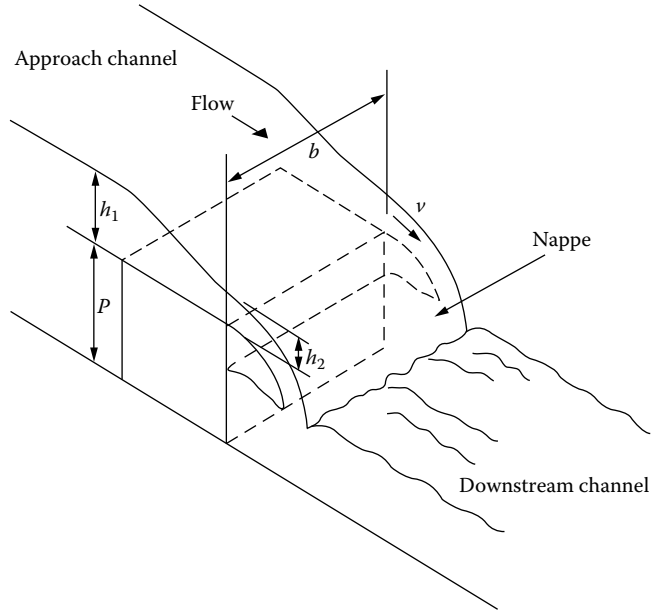


FIGURE 52.8 Broad-crested weir: maximum discharge when $h_2 = 2h_1/3$.

assumed that the channel width remains constant so that a contraction of the nappe occurs only in the vertical direction as the water accelerates over the obstacle. When the weir is narrower than the upstream channel, there is an additional horizontal contraction. The effect of a significant upstream velocity is accounted for empirically.

As an alternative to the weir, the flume (Figure 52.9) provides a method for flowmetering with relatively low pressure loss. By restricting the channel width, analogous to the venturi tube used in pipe flow, the flow velocity in the narrow portion of the channel is increased and the water level sinks accordingly. Most of the head of water is recovered in the diffusing section of the weir. The water levels upstream and in the throat of the weir can be determined by simple floats and recorded on a chart by pens driven mechanically from the floats. For remote monitoring, the use of echo sounders is advantageous.

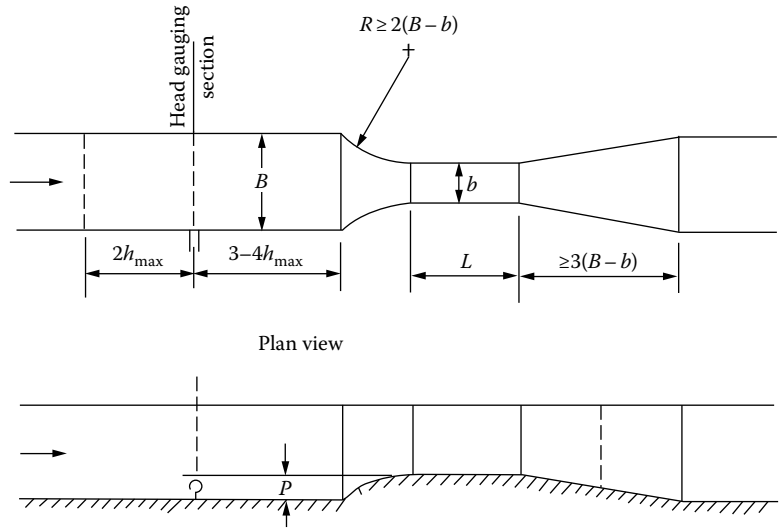


FIGURE 52.9 Venturi flume: maximum discharge when throat depth = 2/3 × total head.

The weir is preferentially used in natural riverbeds and the flume in canalized water courses. The flume must be used in streams with sediment transport to avoid the accumulation of deposits that would occur at the approach to a weir. Weirs and flumes are characterized by

- Simple measurement of the water level
- Simple maintenance
- Reliable measurement of large flow rates at low stream velocity
- Limited measurement accuracy (at best about 2%)
- High installation costs, particularly for flumes

Recommendations for installation of weirs and flumes, for the location of the head measuring station upstream, and for determining the discharge from the measured head are given, for example, in [2, 3] as well as in appropriate standards (e.g., [4–6]).

52.2 Measuring Principles of Variable Area Flowmeters

52.2.1 Rotameter

52.2.1.1 Flow Rate Analysis

The forces acting on the bob lead to equilibrium between the weight of the bob $\rho_b g V_b$ acting downward and the buoyancy force $\rho g V_b$ and the drag force F_d acting upward, where V_b is the volume and ρ_b is the density of the bob, ρ is the density of the fluid, and g is the gravitational acceleration:

$$\rho_b g V_b = \rho g V_b + F_d \tag{52.1}$$

The drag force results from the flow field surrounding the bob and particularly from the wake of the bob. In flow analyses based on similarity principles, these influences are accounted for by empirical coefficient C_L or C_T in the drag law for

$$\text{Laminar flow } F_d = C_L \mu D_b U \tag{52.2}$$

$$\text{Turbulent flow } F_d = C_T \rho D_b^2 U^2 \tag{52.3}$$

where

μ is the fluid viscosity
D_b is the maximum bob diameter
U is the velocity in the annular gap around the bob at the minimum cross section

The volume flow rate through the rotameter is

$$\dot{Q} = \frac{\pi}{4}(D^2 - D_b^2)U \tag{52.4}$$

or

$$\dot{Q} = m\frac{\pi}{4}D_b^2 U \tag{52.5}$$

where m is the open area ratio, defined as

$$m = \frac{D^2 - D_b^2}{D_b^2} \tag{52.6}$$

and D is the tube diameter at the height of the bob.

Combining Equations 52.1, 52.2, and 52.4 for laminar flow gives

$$\dot{Q}_L = \alpha D_b^4 \frac{(\rho_b - \rho)g}{\mu} \tag{52.7}$$

where the parameter α is defined in terms of a constant $K = V_b/D_b^3$ characteristic of the shape of the bob

$$\alpha = \frac{\pi m K}{4C_L} \tag{52.8}$$

Using Equation 52.3 instead of Equation 52.2 yields for turbulent flow

$$\dot{Q}_T = \beta D_b^{5/2}\sqrt{\frac{(\rho_b - \rho)g}{\rho}} \tag{52.9}$$

where

$$\beta = \frac{\pi m}{4}\sqrt{\frac{K}{C_T}} \tag{52.10}$$

With either laminar or turbulent flow through the rotameter, it is clear from Equations 52.7 and 52.9 that the flow rate is proportional to m. If the cross-sectional area of the tube is made to increase linearly with length, that is,

$$D = D_b(1 + h\tan\phi) \tag{52.11}$$

then since the cone angle ϕ of the tube is small, Equation 52.6 can be written as

$$m = 2h\tan\phi \tag{52.12}$$

and the flow rate is directly proportional to the height h of the bob.

52.2.1.2 Similarity Analysis

In early studies of floating element flowmeters, Ruppel and Umpfenbach [7] proposed the introduction of characteristic dimensionless quantities, to permit the use of experimentally determined flow coefficients in flowmeter analysis. Lutz [8] extended these ideas by showing that the transfer of flow coefficients from one flowmeter to another is possible if geometrical similarity exists. More recent works [9–12] have used these principles to produce graphical or computer-based design schemes and have proposed general guidelines for laying out practical flowmetering systems.

The basic scaling parameter for flow is the Reynolds number, defined as

$$\mathrm{Re} = \frac{\rho U_{\mathrm{IN}} D_{\mathrm{b}}}{\mu} \tag{52.13}$$

where U_{IN} is the velocity at the rotameter inlet and the tube diameter D is represented by its value at the inlet, equal to the bob diameter D_{b}. Through the Reynolds number, regimes of laminar or turbulent flow, and particularly important for the rotameter flow regimes with strong or weak viscosity dependence, can be distinguished. Originating in the work [7], it has been found to be practical for rotameters to use an alternative characteristic number, the Ruppel number, defined as

$$\mathrm{Ru} = \frac{\mu}{\sqrt{m_{\mathrm{b}} g \rho (1 - \rho/\rho_{\mathrm{b}})}} \tag{52.14}$$

where $m_{\mathrm{b}} = \rho_{\mathrm{b}} D_{\mathrm{b}}^3$ is the mass of the bob. By combining Equations 52.9, 52.10, and 52.14, the mass flow $\dot{m}$ through the rotameter can be written as

$$\dot{m} = \frac{\pi}{4} \frac{m D_{\mathrm{b}} \mu}{\sqrt{C_{\mathrm{T}}\,\mathrm{Ru}}} \tag{52.15}$$

Alternatively, from the definition

$$\dot{m} = \rho \dot{Q} = \rho U_{\mathrm{IN}} \frac{\pi}{4} D_{\mathrm{b}}^2 \tag{52.16}$$

and Equation 52.13, the flow rate is

$$\dot{m} = \frac{\pi}{4} D_{\mathrm{b}} \mu \mathrm{Re} \tag{52.17}$$

Equations 52.15 and 52.17 give the following relationship between the Ruppel number and the Reynolds number:

$$\mathrm{Ru} = \frac{m}{\sqrt{C_{\mathrm{T}} \mathrm{Re}}} \tag{52.18}$$

An analysis for laminar flow leads to the relationship

$$\mathrm{Ru} = \sqrt{\frac{m}{C_{\mathrm{L}} \mathrm{Re}}} \tag{52.19}$$

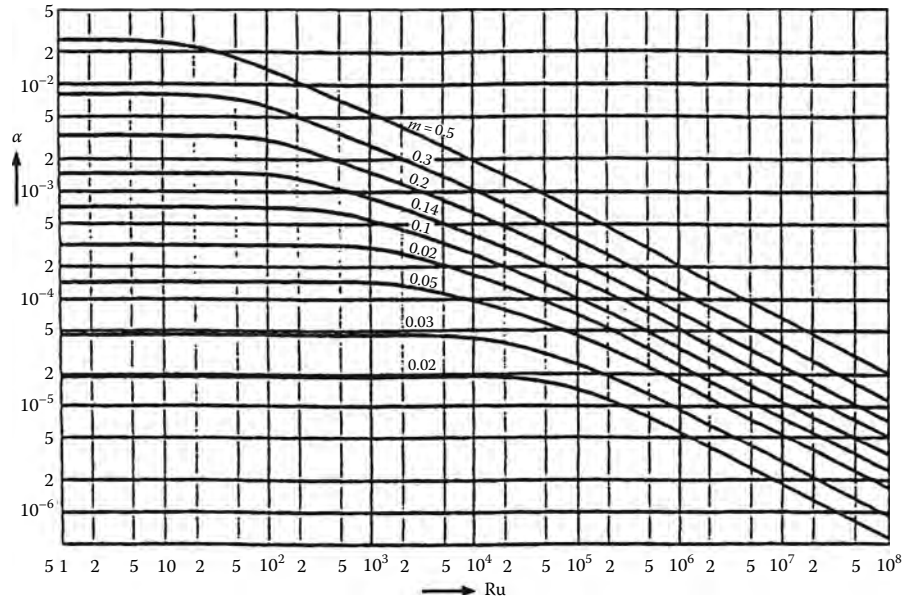


FIGURE 52.10 Rotameter flow coefficient α as a function of Ruppel number Ru and open area ratio m. For a given geometry, α is constant for low Ru (laminar flow).

The advantage of the Ruppel number is its independence of the flow rate. Since the Ruppel number contains only fluid properties and the mass and the density of the bob, it is a constant for a particular instrument.

At low Ruppel numbers, the linear resistance law assumed in Equation 52.2 applies and α is a constant for given m, as shown in Figure 52.10. At higher Ruppel numbers, the flow is transitional or turbulent, and log α decreases linearly with log Ru. In Figure 52.11, curves for β against Ruppel number show a linear increase of log β with log Ru in the laminar region, followed by a gradual transition to horizontal curves in fully turbulent flow.

Similarity analysis of rotameters allows the easy calculation of the flow reading with changing density and viscosity of the fluid. For lower viscosities, only the density effect has to be taken into account; hence, for most gas measurements, simple conversion factors can be used. With higher viscosity, manufacturers offer either Ruppel number–related conversion factors or 2D tables of conversion factors to be applied to different heights of the bob. Some manufacturers offer recalibration services, allowing the user to order new scales for instruments to be used in changed applications and environments. For large-order users, simple computer programs are available from some manufacturers for scale calculations.

Theories based on similarity considerations can predict the variation of the flow coefficient with the Ruppel number in the laminar and turbulent flow regimes but not in laminar–turbulent transitional flow. Detailed experimental and computational studies can assist the flow analysis of floating element flowmeters. An example is given in the section "Rotameter Internal Flow Analysis."

52.2.2 Weir and Flume

52.2.2.1 Flow Rate Analysis for Weirs

Although the determination of the discharge across a weir or a flume generally requires the knowledge of one or more empirical coefficients, the basic equations describing the head-discharge characteristics of the various geometrical forms are readily derived from simple fluid mechanics considerations.

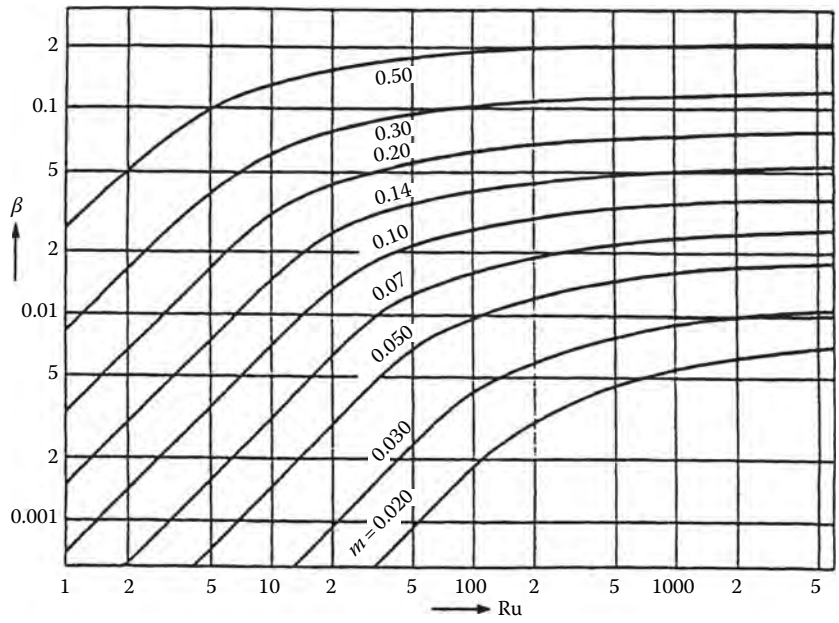


FIGURE 52.11 Rotameter flow coefficient β as a function of Ruppel number Ru and open area ratio m. For a given geometry, β is constant for high Ru (turbulent flow).

The discharge over a sharp-crested weir of width b (Figure 52.5) when the water level over the sill is h is analyzed by considering a horizontal strip of water of thickness δy at a depth y below the water surface. Since the velocity of the water through this strip is $\sqrt{2gy}$, the discharge is

$$\delta\dot{Q} = b\delta y\sqrt{2gy} \tag{52.20}$$

The total discharge is obtained by integration, introducing a coefficient of discharge C_d:

$$\dot{Q} = C_d\sqrt{2gb}\int_0^h \sqrt{y}dy \tag{52.21}$$

$$\dot{Q} = \frac{2}{3}C_d\sqrt{2g}bh^{3/2} \tag{52.22}$$

Empirical corrections to Equation 52.22 to account for the weir height p are given, for example, in [3], but provided h/p does not exceed 0.5, it is adequate to use Equation 52.22 with a discharge coefficient of 0.63 to achieve a flow rate tolerance of 3%.

For the rectangular notch, Equation 52.22 also applies with a discharge coefficient of 0.59 and a flow rate tolerance of 3%. Since in practice, the discharge coefficient varies slightly with the head of water for both weirs and notches, it is sometimes preferred to replace Equation 52.22 by $\dot{Q} = Kbh^n$, where K and n are determined by calibration.

The triangular notch (Figure 52.7) offers the advantage that the length of the wetted edge varies with the head of water, in contrast with the rectangular notch where it is constant. Consequently, the discharge coefficient of the triangular notch is constant for all heads. Furthermore, the achievable range (ratio of maximum to minimum flow rates) is higher for the triangular geometry. If h is the height of the

water surface and θ is the angle of the notch, then a horizontal strip of the notch of thickness δy at depth y has a width $2(h - y)\tan(\theta/2)$. The discharge through this strip is

$$\delta\dot{Q} = 2(h-y)\tan\frac{\theta}{2}\delta y\sqrt{2gy}C_{\mathrm{d}} \tag{52.23}$$

giving a total discharge through the notch

$$\dot{Q} = 2C_{\mathrm{d}}\sqrt{2g}\tan\frac{\theta}{2}\int_0^h (h-y)\sqrt{y}\,dy \tag{52.24}$$

$$\dot{Q} = \frac{8}{15}C_{\mathrm{d}}\sqrt{2g}\tan\frac{\theta}{2}h^{5/2} \tag{52.25}$$

The discharge coefficient is about 0.58.
Commonly used notch angles are 90° (tan $\theta/2 = 1$, $C_{\mathrm{d}} = 0.578$), 53.13° (tan $\theta/2 = 0.5$, $C_{\mathrm{d}} = 0.577$), and 28.07° (tan $\theta/2 = 0.25$, $C_{\mathrm{d}} = 0.587$). Notches with $\theta = 53.13°$ and $\theta = 28.07°$ deliver, respectively, one-half and one-quarter of the discharge of the 90° notch at the same head.

The aforementioned derivations assume that the water is discharging from a reservoir with cross-sectional area far exceeding the flow area at the weir, so that the velocity upstream is negligible. When the weir is built into a channel of cross-sectional area A, the water will have a finite velocity of approach $v_1 = \dot{Q}/A$. As a first approximation, $\dot{Q}$ is obtained from Equation 52.22 assuming zero velocity of approach. Assuming further that v_1 is uniform over the weir, there will be additional head $v_1^2/2g$ acting over the entire weir. Referring to Figure 52.5, the total discharge is then

$$\dot{Q} = C_{\mathrm{d}}\sqrt{2gb}\int_{v_1^2/2g}^{h+v_1^2/2g}\sqrt{y}\,dy \tag{52.26}$$

$$\dot{Q} = \frac{2}{3}C_{\mathrm{d}}\sqrt{2gb}\left[(h+v_1^2/2g)^{3/2} - (v_1^2/2g)^{3/2}\right] \tag{52.27}$$

From Equation 52.27, a corrected value of v_1 can be determined. Further iterations converge rapidly to the final value of the discharge $\dot{Q}$. Equation 52.27 can be written as

$$\dot{Q} = \frac{2}{3}C_{\mathrm{v}}C_{\mathrm{d}}\sqrt{2g}bh^{3/2} \tag{52.28}$$

where

$$C_{\mathrm{v}} = \left(1+\frac{v_1^2}{2gh}\right)^{3/2} - \left(\frac{v_1^2}{2gh}\right)^{3/2} \tag{52.29}$$

is the coefficient of velocity, which is frequently determined empirically.

Over a broad-crested weir (Figure 52.8), the discharge depends on the head h_1, the width b, and the length l of the sill. There is also a dependence on the roughness of the sill surface and the viscosity. Consequently, there is a loss of head as the water flows over the sill. It is assumed that the sill is of suf-

ficient length to allow the velocity to be uniform at a value v throughout the depth h_2 of the water at the downstream edge of the sill. Neglecting losses,

$$v = \sqrt{2g(h_1 - h_2)}$$

and the discharge is

$$\dot{Q} = C_d b h_2 v = C_d b \sqrt{2g\left(h_1 h_2^2 - h_2^3\right)} \tag{52.30}$$

From Equation 52.30, the discharge is a maximum when $(h_1 h_2^2 - h_2^3)$ reaches a maximum, that is, when $h_2 = 2h_1/3$. This discharge is then

$$\dot{Q} = \left(\frac{2}{3}\right)^{3/2} C_d b \sqrt{g} h_1^{3/2} \tag{52.31}$$

The stable condition of the weir lies at the maximum discharge.

52.2.2.3 Flow Rate Analysis for Flumes

The discharge through a flume depends on the water level h, the channel width b, and the velocity v at each of the stations 1 (upstream) and 2 (at the minimum cross section) in Figure 52.9. Applying Bernoulli's equation to the inlet and the throat and neglecting all losses, one obtains

$$\rho g H = \rho g h_1 + \frac{1}{2}\rho v_1^2 = \rho g h_2 + \frac{1}{2}\rho v_2^2 \tag{52.32}$$

where H is the total head. The analysis is then formally the same as that for the broad-crested weir, giving the same expression for the discharge as in Equation 52.31. The flow through the flume is a maximum when the depth at the throat is two-thirds of the total head. Normally, a coefficient of velocity is introduced so that the discharge equation is

$$\dot{Q} = \left(\frac{2}{3}\right)^{3/2} C_v C_d b \sqrt{g} h_1^{3/2} \tag{52.33}$$

The combined coefficient is determined empirically. In [3], a value $C_v C_d = 1.061 \pm 0.085$ is quoted.

The validity of the empirical coefficients quoted earlier, or of other coefficients given in the various standards for metering of open-channel flows, can only be guaranteed with a given confidence level for meter installations satisfying certain geometrical constraints. Limits on quantities such as *b, h,* and *h*/*b* are given in [3] and the standards.

52.2.3 Rotameter: Internal Flow Analysis

52.2.3.1 Computation of Internal Flow

The improvement of rotameter design could be assisted by detailed knowledge of the internal flow field, which is characterized by steep velocity gradients and regions of separated flow. The measurements of the internal flow field are complicated by the small dimensions of the gap around the bob and the strongly curved glass tube. Bückle et al. [13] successfully used laser Doppler anemometry [14] for velocity measurements in a rotameter. The working fluid was a glycerin solution with an index of refraction (1.455) close to that of glass (1.476). Problems with refraction of the laser beams at the curved tube wall were thus avoided, but the high viscosity of glycerin restricted the experiments to laminar flow at Reynolds numbers $\mathrm{Re} = \rho U_{IN} D_b/\mu < 400$.

The application of computational fluid dynamics to the flow in a rotameter [13] involves the finite volume solution of the conservation equations for mass and momentum. For a 2D laminar flow, these equations can be written in a cylindrical polar coordinate system as

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho u)}{\partial z} + \frac{1}{r}\frac{\partial(\rho r v)}{\partial r} = 0 \tag{52.34}$$

$$\frac{\partial(\rho u)}{\partial t} + \frac{\partial}{\partial z}\left(\rho uu - 2\mu\frac{\partial u}{\partial z}\right) + \frac{1}{r}\frac{\partial}{\partial r}\left(\rho ruv - \mu r\left(\frac{\partial u}{\partial r} + \frac{\partial v}{\partial z}\right)\right) = -\frac{\partial p}{\partial z} \tag{52.35}$$

$$\frac{\partial(\rho v)}{\partial t} + \frac{\partial}{\partial z}\left(\rho uv - \mu r\left(\frac{\partial u}{\partial r} + \frac{\partial v}{\partial z}\right)\right) + \frac{1}{r}\frac{\partial}{\partial r}\left(\rho rvv - 2\mu r\frac{\partial v}{\partial r}\right) = -\frac{\partial p}{\partial r} \tag{52.36}$$

where

μ is the density

u, v and r, and z are the velocity components and coordinate directions in the axial and radial directions, respectively

μ is the dynamic viscosity

p is the pressure

For the numerical solution, the governing equations for a generalized transport variable ϕ (i.e., u or v) are formally integrated over each control volume (CV) of the computational grid. The resulting flux-balance equation can in turn be discretized as an algebraic equation for ϕ at the center P of each CV in terms of the values ϕ_{nb} at the four nearest neighbors of point P and known functions A:

$$A_{p\phi p} + \sum_{nb} A_{nb}\phi_{nb} = S_{\phi} \tag{52.37}$$

For the whole solution domain, a system of equations results that can be solved by a suitable algorithm (e.g., [15]).

For a rotameter, the symmetry of the problem allows the computational solution domain to be chosen as one-half diametral plane of the rotameter. The boundary conditions are

$$u = v = 0 \tag{52.38}$$

along the walls and

$$\frac{\partial u}{\partial r} = v = 0 \tag{52.39}$$

along the axis of symmetry. At the input boundary, the initial profile for the u velocity is taken from the experiment. At the outlet boundary, zero gradients are assumed for all dependent variables.

52.2.3.2 Computed Flow Field

Representative calculations for Reynolds number 220 are shown as velocity vectors (Figure 52.12) and streamlines (Figure 52.13). Experimental and computed velocity profiles over a radius of the flowmeter tube are shown in Figure 52.14.

At $z = 45$ mm, the computations indicate a stagnation point at the bob tip, but the measurements show a low but finite axial velocity. The deviation from zero is attributable to slight unsteadiness in the position of the bob and to smearing of the steep radial velocity gradient along the measuring volume. This



FIGURE 52.12 Computed velocity vectors for laminar flow through a rotameter. Computations in a half diametral plane for axisymmetric flow at Reynolds number 220.



FIGURE 52.13 Computed streamlines for laminar flow through a rotameter (Reynolds number 220).

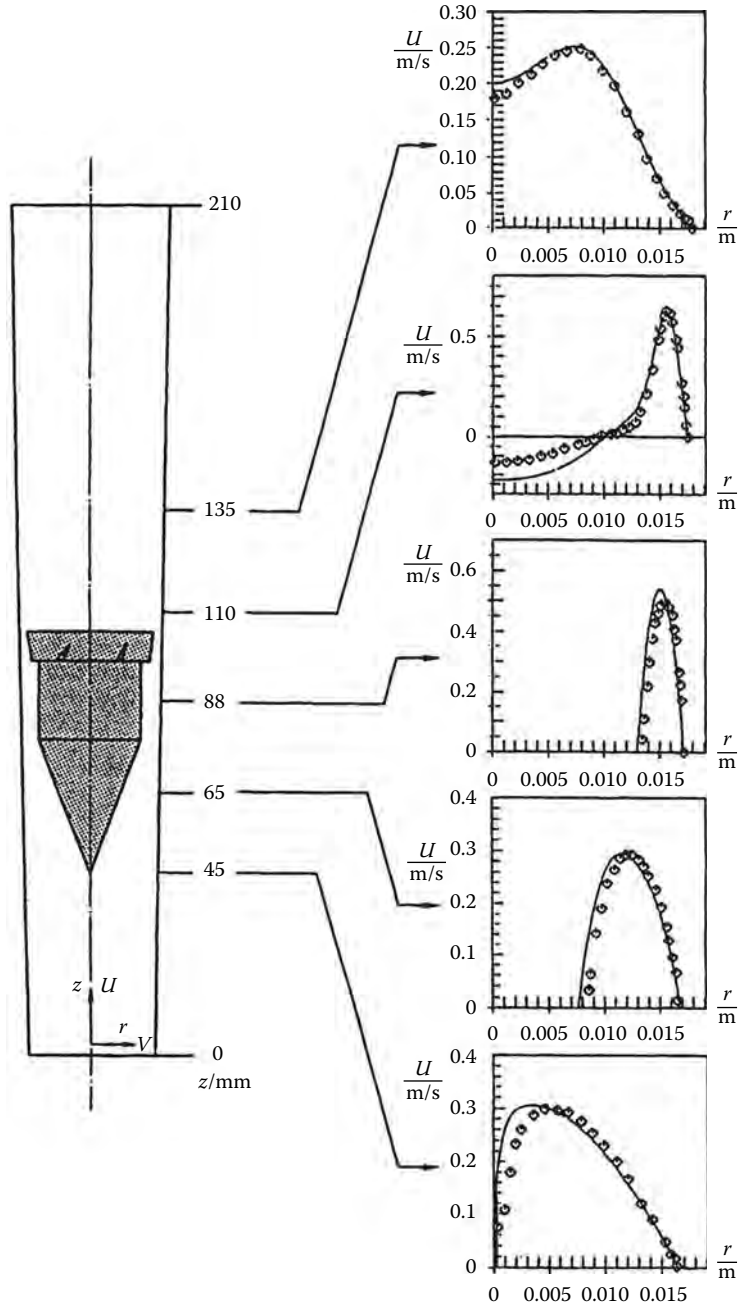


FIGURE 52.14 Comparison of measured and computed velocity profiles for laminar flow through a rotameter (Reynolds number 220). U = axial component (m/s), z = axial position (mm), and r = radial position (m).

effect is also apparent at radii up to about 5 mm, where the computations show a notably steeper velocity gradient than the measurements. The measured and computed peak velocities agree well, although the measured peak is located at a larger radius.

In the strongly converging annulus between the bob and the tube wall (z = 65 mm) and in the plane z = 88 mm, the computations reproduce well the trend of the measured results, particularly near the flowmeter wall. Discrepancies in the region adjacent to the bob are attributable to asymmetry in measured profiles arising from the piping upstream of the rotameter tube.

Above the bob (z = 110 mm), there is a strong upward flow in an annular region near the tube wall and a recirculation region around the axis occupying almost half the tube cross section. The forward velocities show a good match between computations and measurements, but there is a very marked discrepancy in the recirculation zone. At z = 135 mm, the recirculation zone is finished, but the wake of the bob is still very evident. Except for a discrepancy of about 10% on the axis, the results from the two methods agree remarkably well.

The effect of rotation of the bob on the flow field was considered by computations, including an equation for the azimuthal velocity component w:

$$\frac{\partial(\rho w)}{\partial t}+\frac{\partial}{\partial z}\left(\rho uw-\mu\frac{\partial w}{\partial z}\right)+\frac{1}{r}\frac{\partial}{\partial r}\left(\rho rvw-\mu r\frac{\partial w}{\partial r}\right)=-\rho\frac{vw}{r}-\mu\frac{w}{r^2} \tag{52.40}$$

Calculations for 1 Hz rotation frequency showed the highest swirl velocities close to the axis above the bob. There was no observable influence of the rotation on the u- and v-velocity components [16].

52.3 Summary

For pipe flows, variable area flowmeters are most suitable for low flow rates of gases or liquids at moderate temperatures and pressures. Favorable features include rugged construction, high reliability, low pressure drop, easy installation, and low cost. Disadvantages include measurement uncertainty of 1% or more, limited range (10:1), slow response, and restrictions on the meter orientation. A generally good price/performance ratio has led to widespread use of these meters in numerous scientific and medical instruments and in many industrial applications for flow monitoring.

Variable area flowmeters in open-channel flows have applications for flow measurements in wastewater plants, waterworks, rivers and streams, irrigation, and drainage canals. Hydrological applications of weirs with adjustable sill or crest height include flow regulation, flow measurement, upstream water-level control, and discharge of excess flow in streams, rivers, and canals. Ruggedness, simplicity, and low maintenance costs are favorable characteristics for field applications of these meters.

References

1. R. A. Furness, BS7045: the principles of flowmeter selection, *Flow Meas. Instrum.*, 2, 233–242, 1991.
2. W. Boiten, Flow-measuring structures, *Flow Meas. Instrum.*, 4, 17–24, 1993.
3. R. Hershey, General purpose flow measurement equations for flumes and thin plate weirs, *Flow Meas. Instrum.*, 6, 283–293, 1995.
4. ISO 1438, *Thin Plate Weirs*, International Standards Organisation, Geneva, Switzerland, 1980.
5. ISO 4359, *Rectangular, Trapezoidal and U-Shaped Flumes*, International Standards Organisation, Geneva, Switzerland, 1983.
6. ISO 8368, *Guidelines for the Selection of Flow Gauging Structures*, International Standards Organisation, Geneva, Switzerland, 1985.

7. G. Ruppel and K. J. Umpfenbach, Stromungstechnische Untersuchungen an Schwimmermessern, *Technische Mechanik und Thermodynamik*, 1, 225–233, 257–267, 290–296, 1930.
8. K. Lutz, Die Berechnung des Schwebekörper-Durchfluβmessers, *Regelungstechnik*, 10, 355–360, 1959.
9. D. Bender, Ähnlichkeitsparameter und Durchfluβgleichungen für Schwebekörperdurchfluβmesser, *ATM-Archiv für Technisches Messen*, 391, 97–102, 1968.
10. VDE/VDI-Fachgruppe Meβtechnik 3513, *Schwebekörperdurchfiu*βmesser, Berechnungsverfahren, VDE/VDI 3513, 1971.
11. H. Nikolaus, Berechnungsverfahren für Schwebekörperdurchfluβmesser, theoretische Grundlagen und Aufbereitung für die Datenverarbeitungsanlage, *ATM-Archiv für Technisches Messen*, 435, 49–55, 1972.
12. H. Nikolaus and M. Feck, Graphische Verfahren zur Bestimmung von Durchfluβkennlinien für Schwebekörperdurchfluβmeβgeräte, *ATM-Archiv für Technisches Messen*, 1247–5, 171–176, 1974.
13. U. Bückle, F. Durst, B. Howe, and A. Melling, Investigation of a floating element flow meter, *Flow Meas. Instrum*., 3, 215–225, 1992.
14. F. Durst, A. Melling, and J. H. Whitelaw, *Principles and Practice of Laser-Doppler Anemometry*, 2nd ed., Academic Press, London, U.K., 1981.
15. M. Perić, M. Schäfer, and E. Schreck, Computation of fluid flow with a parallel multigrid solver, *Proc. Conf. Parallel Computational Fluid Dynamics*, Stuttgart, 1991.
16. U. Bückle, F. Durst, H. Köchner, and A. Melling, Further investigation of a floating element flowmeter, *Flow Meas. Instrum*., 6, 75–78, 1995.

Further Readings

BES Flowmeters, MPB Variable Area Flowmeters for Air & Water. http://www.besflowmeters.com.au/mpb/mpb.html (accessed on July 27, 2012).

Flow Meter Directory, Variable Area Flowmeter Principles, On Line 22/05/2007. http://www.flowmeterdirectory.com/flowmeter_artc_07052201.html (accessed on July 26, 2012).

Lake Monitors, Variable Area Flow Meters. http://www.lakemonitors.com/variable_area.htm (accessed on July 25, 2012).

Partial List of Vendors and Suppliers

Bailey Fischer & Porter Co., County Line Rd., Warminster, PA 18974, Tel.: (215) 674-6000, Fax: (215) 674-7183,

Bopp & Reuther Heinrichs Messtechnik GmbH, Stolberger Str. 393, D-50933 Köln, Germany, Tel.: +49-221-49708-0, Fax: +49-221-497088

Brooks Instrument Division, 407 W. Vine St., Hatfield, PA 19440, Tel.: (215) 362-3500, Fax: (215) 362-3745, http://www.brooksinstrument.com/

CT Platon, Jays Close, Viables, Basingstoke, Hampshire RG22 4BS, Great Britain, Tel.: +44-1256-470456, Fax: +44-1256-363345

ERDCO Engineering Corp., 721 Custer Ave.8, P.O. Box 6318, Evanston, IL 60204–6318, Tel.: 847-328-0550, 800-553-0550 (toll free), Fax: 847-328-3535, www.erdco.com

Erdman Corp., 1900 W. Muhammad Ali Blvd., Louisville, KY 40203, Tel.: 502-584-1271, 800-523-0420 (toll free), Fax: 502-584-3611, www.erdmanncorp.com

KDG-Mobrey Ltd., Crompton Way, Crawley, West Sussex RH10 2YZ, Great Britain, Tel.: +44-1293-518632, Fax: +44-1293-533095

Kobold Messring GmbH, Nordring 22-24, D-65719 Hofheim/Taunus, Germany, Tel.: +49 6192 2990, Fax: +49 6192 23398, http://kobold.com/

Krohne America Inc., 7 Dearborn Rd., Peabody, MA 01960, Tel.: (508) 535-6060, (800) 356-9464, Fax: (508) 535-1720,

Orange Research, Inc., 140 Cascade Blvd., Milford, CT 06460, Tel.: 203-877-5657, 800-989-5657 (toll free), Fax: 203-783-9546, www.orangeresearch.com

Porter Instrument Div., parker Hannifin Corp., 245 Township Line Rd., P.O. Box 907-TR, Hatfield, PA 19440, Tel.: 215-723-4000, 800-457-2001 (toll free), Fax: 215-723-2199, www.porterinstrument.com

Rota Yokogawa GmbH & Co. KG, Rheinstrasse 8, D-79664 Wehr, Germany, Tel.: +49-7761-567-0, Fax: +49-7761-567-126

Ryan Herco Flow Solutions, 3010 N. San Fernando Blvd., Burbank, CA 91504, Tel.: 818-841-1141, 800-848-1141 (toll free), Fax: 818-973-2650, www.rhfs.com

Tokyo Keiso Co. Ltd., Shiba Toho Building, 1–7–24, Shibakoen, Minato-ku, Tokyo 105, Japan, Tel.: +81-3-3431-1625, Fax: +81-3-3433-4922, http://www.tokyokeiso.co.jp/english/index.html

Wallace & Tiernan, Inc., 25 Main St., St. Belleville, NJ 07109, Tel.: (201) 759-8000, Fax: (201) 759-9333, http://www.wallaceandtiernan-usa.com/

Webster Instruments USA, 1290 E. Waterford Ave., Milwaukee, WI 53235, Tel.: 414-769-6400, 800-932-8378 (toll free), Fax: 414-769-6591, 866-356-9637 (toll free), www.webster-inst.com/

53

Positive Displacement Flowmeters

53.1 Design and Construction **53**-1
53.2 Some Commercially Available PD Meter Designs **53**-2
Sliding-Vane-Type PD Meter • Trirotor-Type PD Meter • Birotor PD Meter • Piston-Type PD Meter • Oval-Gear PD Meter • Nutating-Disk-Type PD Meters • Roots PD Meter • CVM Meter • Diaphragm Meter
53.3 Advantages and Disadvantages of PD Meters **53**-7
53.4 Applications **53**-7
Accessories • Price and Performance
Further Information **53**-9
Partial List of Vendors and Suppliers **53**-9

Zaki D. Husain
Daniel Flow Products, Inc.

Donald J. Wass
Daniel Flow Products, Inc.

A *positive displacement flowmeter*, commonly called a PD meter, measures the volume flow rate of a continuous flow stream by momentarily entrapping a segment of the fluid into a chamber of known volume and releasing that fluid back into the flow stream on the discharge side of the meter. By monitoring the number of entrapments for a known period of time or number of entrapments per unit time, the total volume of flow or the flow rate of the stream can be ascertained. The total volume and the flow rate can then be displayed locally or transmitted to a remote monitoring station.

The PD flowmeter has been in use for many decades to measure both liquid and gas flows. A PD meter can be viewed as a hydraulic motor with high volumetric efficiency that generally absorbs a small amount of energy from the flowing fluid. The energy absorption is primarily to overcome the internal resistance of the moving parts of the meter and its accessories. This loss of energy is observed as the pressure drop across the meter. The differential pressure across the meter is the driving force for the internals of the PD meter.

53.1 Design and Construction

A PD meter has three basic components: an outer housing, the internal mechanism that creates the dividing chamber through its repetitive motion, and the display or counter accessories that determine the number of entrapments of fluid in the dividing chamber and infer the flow rate and the total volume of flow through the meter.

The external housing acts as the holding chamber of the flowing fluid before and after the entrapments by the internals. For low line pressures, the housing is usually single walled, while for higher operating pressures, the housing is double walled where the inner wall is the containment wall for the entrapment chamber and the outer wall is the pressure vessel. For the double-walled housing, the entrapment chamber walls are subjected to the differential pressure across the meter, while the external

housing is subjected to the total line pressure. This allows fabrication of a thin-walled inner chamber that can retain its precise shape and dimensions independent of line pressure.

The measuring mechanism consists of precise metering elements that require tight tolerances on the mating parts. The metering elements consist of the containment wall of the metering chamber and the moving components of the meter that form the entrapment volume of the flowing fluid by cyclic or repetitive motion of those elements. The most common types of PD meters are the oscillating-piston, nutating-disk, oval-gear, sliding-vane, birotor, trirotor, and diaphragm designs.

The counter or output mechanism converts the motion of the internal measuring chamber of a PD meter and displays the flow rate or total flow by correlating the number of entrapments and each entrapped volume. Many PD meters have a mechanical gear train that requires seals and packing glands to transmit the motion of the inner mechanism to the outside counters. This type of display requires more driving power to overcome resistance of the moving parts and the seals, which results in an additional pressure drop for the meter. Many PD meters transmit the motion of the inner mechanism to the counters through switch output utilizing electromechanical, magnetic, optical, or purely electronic techniques in counting the entrapments and displaying the flow rate and total flow volume. The latter processing techniques normally have less pressure drop than the all-mechanical or partly mechanical transmission methods. All-mechanical drive counters do not require external power and, through proper selection of gear train, can display the actual flow volume or the flow rate. Thus, meters can be installed at remote locations devoid of any external power source. Meters with mechanical display cannot easily correct for the changes in volume due to thermal expansion or contraction of the measuring chamber due to flow temperature variation and, in the case of single-walled housing, the changes in the entrapment volume due to variations of the line pressure. An electronically processed output device could monitor both the pressure and the temperature of the meter and provide necessary corrections to the meter output. With constantly changing and improving electronics technology, many PD meters are now installed with solar- or battery-powered electronic output for installations at remote locations with no external power source. Many electronic displays allow access to meter data from a central monitoring station via radio or satellite communication.

53.2 Some Commercially Available PD Meter Designs

Commercially available PD meters have many noticeably different working mechanisms, and a few designs are unique and proprietary. Although each design and working mechanism can be noticeably different from another, all PD meters have a stationary fluid retaining wall and a mechanism that momentarily entraps inlet fluid into a partitioned chamber before releasing it to the downstream side of the meter. This entrapment and release of the flowing fluid occur with such repetitive and sweeping motion that, for most practical purposes, the flow rate appears to be uniform and steady, even though, in reality, the exit flow does have some pulsation. The flow pulsation out of the meter may be more pronounced for some designs of PD meters than others. These flow pulsations are more pronounced at the lower flow rates for all designs. Some designs are more suitable for either liquid or gas flows, while some designs can measure both gas and liquid. For liquid applications, PD meters work best for liquids with heavy viscosities. Almost all PD meters require precisely machined, high-tolerance mating parts; thus, measured fluid must be clean for longevity of the meter and to maintain the measurement precision.

53.2.1 Sliding-Vane-Type PD Meter

Figure 53.1 shows a working cycle of a sliding-vane-type PD meter where vanes are designed to move in and out of the rotating inner mechanism. The position of each sliding vane relative to specific angular rotation of the rotor is usually maintained by a mechanical cam. In the design shown in Figure 53.1,



FIGURE 53.1 Sliding-vane-type PD meter. (Courtesy of Daniel Industries, Inc., Houston, TX.)



FIGURE 53.2 Trirotor-type PD meter. (Courtesy of Liquid Controls LLC, Lake Bluff, IL.)

each blade is independently retained in the slots. There are designs with even numbers of blades where diametrically opposite blades are one integral unit. High-pressure application would utilize dual-wall construction.

53.2.2 Trirotor-Type PD Meter

This design has three rotating parts that entrap fluid between the rotors and an outer wall. The working cycle of this type of meter is shown in Figure 53.2. In this design, for one rotation of the top mechanism, two blades rotate twice. The driving mechanism and rotation of each blade with respect to the others is maintained by a three-gear assembly where each rotating blade shaft is connected to one gear of the three-gear assembly to maintain relative rotational speed. This design is used to measure liquid flows.

53.2.3 Birotor PD Meter

The measuring unit of the birotor PD meter has two spiral-type rotors kept in perfect timing by a set of precision gears and is used to measure liquid flows. Flow can enter the meter either perpendicular or parallel to the axis of rotation of the mating rotors, as shown in Figure 53.3. The axial design birotor is identical to the standard model in component parts and in principle; however, it utilizes a measuring unit mounted parallel, rather than perpendicular, to flow. Meter output is registered mechanically through a gear train located outside the measuring chamber or, electronically, using a pickup assembly mounted with a timing gear. Axial orientation results in compact installation, improved accuracy, and low pressure loss. The axial design is ideal for high-flow-rate applications.

53.2.4 Piston-Type PD Meter

A typical design of a piston-type PD meter is shown in Figure 53.4. A centrally located rotating part generates the reciprocating motion for each of the four pistons of the meter. The reciprocating motion of each of the piston is timed such that the discharge from an individual piston–cylinder occurs in a cycle to generate a semicontinuous discharge from the meter. These PD meters are used for very low flow rates of liquid flows. A piston–cylinder design can withstand large differential pressures across the meter, so highly viscous liquids can be measured by this type of meters, and measurements are very precise. Mechanical components of this type of meter require very precise mechanical tolerances, which increase the product cost.

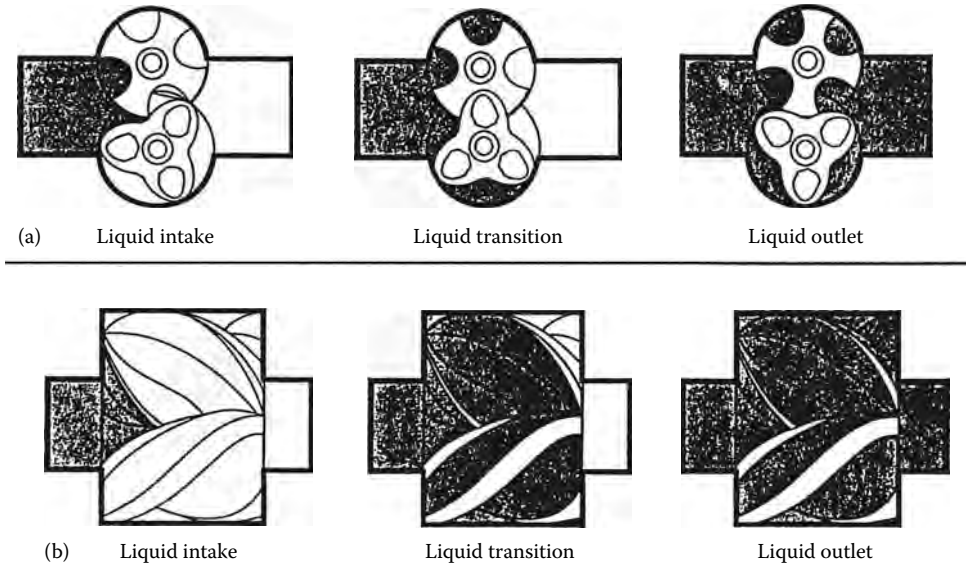


FIGURE 53.3 Birotor-type PD meter. (a) The standard birotor principle and (b) the axial flow principle. (Courtesy of Brooks Instruments Division, Emerson Electric Company, Hatfield, PA.)

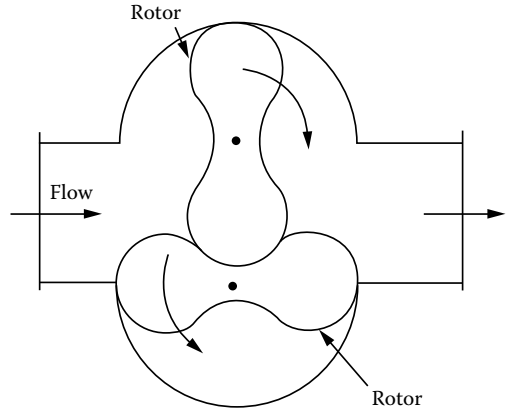


FIGURE 53.4 Piston-type PD meter. (Courtesy of Pierburg Instruments, Inc., Auburn Hills, MI.)

53.2.5 Oval-Gear PD Meter

The measurement of volumetric flow of an oval-gear meter is obtained by separating partial volumes formed between the oval gears and the measuring chamber wall, as shown in Figure 53.5. The rotation of the oval gears results from the differential pressure across the flowmeter. During one revolution of the oval gears, four partial volumes are transferred. The rotation of the gears is transmitted from the measuring chamber to the output shaft directly with mechanical seals or via a magnetic coupling. This type of meter is used to measure liquids having a wide range of viscosity.



FIGURE 53.5 Oval-gear meter. (Courtesy of Daniel Industries, Inc., Houston, TX.)

53.2.6 Nutating-Disk-Type PD Meters

A disk placed within the confines of a boundary wall at a specific orientation can induce a flow instability that can generate a wobbling or nutating motion to the disk. The operating cycle of one such design is shown in Figure 53.6. The entrapment and discharge of the fluid to and from the two chambers occur during different phases of the repetitive cycle of the nutating disk. This type of meter is used to measure liquids. The design provides economic flow measurement where accuracy is not of great importance and is often used in water meters.

Meter designs described hereafter are for gas flows. PD meters for gas flows require very tight tolerances of the moving parts to reduce leak paths, and therefore, the gas must be clean because the meters are less tolerant to solid particles.

53.2.7 Roots PD Meter

The roots meter design is shown in Figure 53.7. This is the most commonly used PD meter to measure gas flows. This is one of the oldest designs of a PD meter. The mechanical clearance between the rotors and the housing requires precisely machined parts. Roots meters are adversely affected if inlet flow to the meter has a relatively high level of pulsation. The main disadvantage of this meter is that it introduces cyclic pulsations into the flow. This meter cannot tolerate dirt and requires filtering upstream of

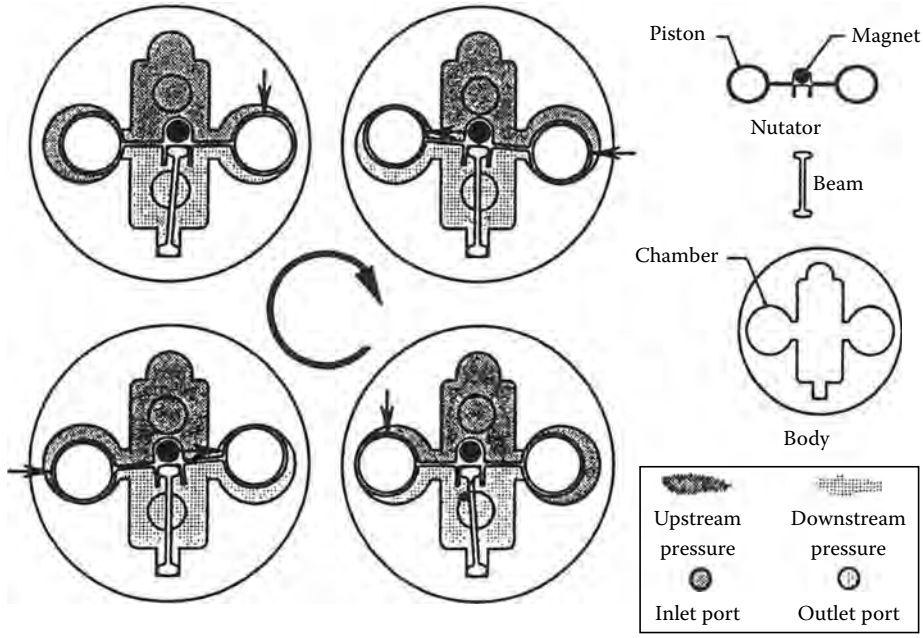


FIGURE 53.6 Nutating-disk-type PD meter. (Courtesy of DEA Engineering Company, Mission Viejo, CA.)

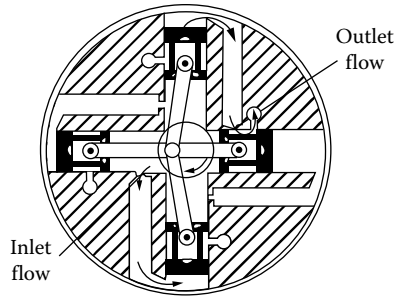


FIGURE 53.7 Schematic diagram of roots meter.

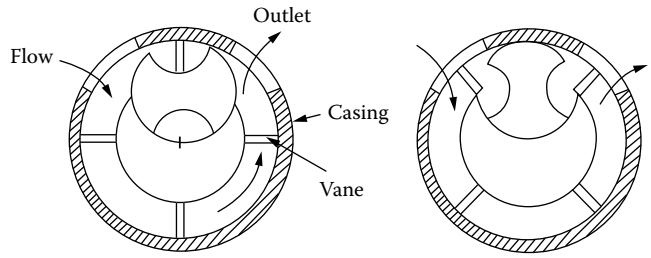


FIGURE 53.8 Operation of CVM meter.

the meter. A roots meter in single-case design is used at near-ambient line pressures. This meter design is very widely used in the low-pressure transmission and distribution market of natural gas. If this meter ever seizes up at any position, the flow is completely blocked.

53.2.8 CVM Meter

The CVM meter (Figure 53.8) is a proprietary design of a PD meter used to measure gas flows. This meter has a set of four vanes rotating in an annular space about the center of the circular housing. A gate similar in shape to the centerpiece of the trirotor design (Figure 53.2) allows the vanes to pass—but not the gas. The measurement accuracy of the CVM meter is similar to the roots meter, but the amplitude of exit pulsation is lower than the roots meter for similar meter size and flow rate. This design is also not as sensitive to inlet flow fluctuations and is widely used in the low-pressure, natural gas distribution market in Europe.

53.2.9 Diaphragm Meter

The diaphragm meter, also known as a bellows-type meter, is simple and relatively inexpensive, has reliable metering accuracy, and is widely used as a domestic gas meter. This design, shown in Figure 53.9,

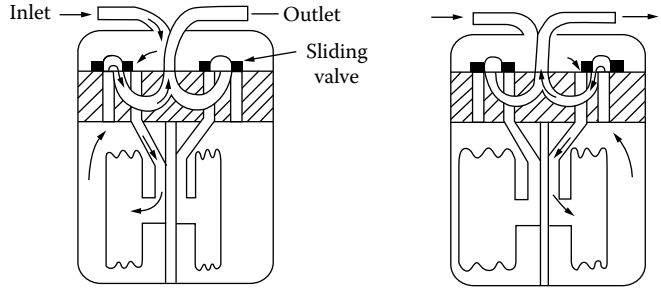


FIGURE 53.9 Operation of a diaphragm meter.

acts like a reciprocating piston meter where the bellows behave as pistons. The operation is controlled by a double-slide valve. In one position, the slide valve connects the inlet to the inside of the bellow, while the outside of the bellow is open to the outlet. In the other position, the outside is connected to the inlet, while the inside of the bellow is open to the outlet. This design can measure extremely low flow rates (e.g., pilot light of a gas burner).

53.3 Advantages and Disadvantages of PD Meters

High-quality PD meters will measure with high accuracy over a wide range of flow rates and are very reliable over long periods. Unlike most flowmeters, a PD meter is insensitive to inlet flow profile distortions. Thus, PD meters can be installed in close proximity to any upstream or downstream piping installations without any loss of accuracy. In general, PD meters have minimal pressure drop across the meter; hence, they can be installed in a pipeline with very low line pressures. Until the introduction of electronic correctors and flow controls on other types of meters, PD meters were most widely used in batch loading and dispensing applications. All-mechanical units can be installed in remote locations.

PD meters are generally bulky, especially in the larger sizes. Due to the tight clearance necessary between mating parts, the fluid must be clean for measurement accuracy and longevity of the meter. More accurate PD meters are quite expensive. If a PD meter ever seizes up, it would completely block the flow. Many PD meters have high inertia of the moving parts; therefore, a sudden change in the flow rate can damage the meter. PD meters are normally suitable over only limited ranges of pressure and temperature. Some designs will introduce noticeably high pulsations into the flow. Most PD meters require a good maintenance schedule and are high-repair and high-maintenance meters. Recurring costs in maintaining a PD flowmeter can be a significant factor in overall flowmeter cost.

53.4 Applications

Liquid PD meters are capable of measuring fluids with a wide range of viscosity and density. Minor changes in viscosity of the fluid have minimal influence on the accuracy of a PD meter. However, no one specific design of PD meter is capable of such a broad application range. The physical properties of the fluid, especially the fluid viscosity, must be reviewed for proper selection of a meter for the application. Many PD meter designs accurately measure flow velocities over a wide range.

Careful dimensional control is required to produce well-regulated clearances around the moving parts of the PD meter. The clearances provide a leakage path around the flow measurement mechanism, and the amount of leakage depends on the relative velocity of the moving parts, the viscosity of the fluid, and the clearances between parts. In general, the leakage flow rate follows the Poiseuille equation; that is, leakage is directly proportional to the differential pressure and inversely proportional to the absolute viscosity and the length of the leakage path.

The leakage flow, in percentage of the total flow rate, is useful information for error measurement. At low flow rates, percent leakage error can be significant, while with modest flow rates, leakage can be insignificant. However, the differential pressure across the PD meter increases exponentially with increasing flow rates. Therefore, at very high flow rates, leakage flow can again be a significant portion of the total flow. As a result, PD meters tend to dispense more fluid than is indicated by the register at the very low and very high flow rates. The amount of error due to leakage depends on the meter design and on the viscosity of the fluid.

PD meters are driven by the differential pressure across the meter. The primary losses within the meter can be attributed to the friction losses and the viscous drag of the fluid on the moving parts. At very low flow rates, the bearing losses are predominant, while at high flow rates, the viscous drags predominate.

The viscous drag is directly proportional to the fluid viscosity and the relative velocity of the moving parts and inversely proportional to the clearance. Within each design, the differential pressure is limited to a predetermined value. Excessive differential pressures can damage the meter. High differential pressure can be avoided by limiting the maximum flow rate, lowering viscosity by heating the fluid, increasing clearances, or various combinations of these techniques. In general, if the differential pressure limit can be addressed, PD meters measuring high-viscosity fluids have less measurement error over a wide flow-rate range.

The viscosity of most gases is too low to cause application problems relating to viscous drag of the form discussed for liquid PD meters. Gas PD meters utilize smaller clearances than liquid PD meters; therefore, gas must be very clean. Even small particles of foreign material can damage gas PD meters. PD meters used in gas flow measurement are primarily in the low line pressure application.

53.4.1 Accessories

Many accessories are available for PD meters. Among the most common and useful accessories for liquid PD meters are the automatic air eliminator and the direct-coupled shutoff valve for batch loading operations. The automatic air eliminator consists of a pressure-containing case surrounding a float valve. When gas in the liquid flow enters the chamber, the float descends by gravity and allows the float valve to open and purge the gas. An air eliminator improves the precision of flow measurement in many applications and is an absolute necessity in applications where air or gas is trapped in the flow stream. The direct-coupled shutoff valve is directly linked to the mechanical register. A predetermined fluid quantity can be entered into the register. When the dispensed amount equals the amount entered in the register, the mechanical linkage closes the valve. Two-stage shutoff valves are often used to avoid line shock in high-flow-rate dispensing applications. Similarly, all-electronic units can control batch operations. The all-electronic units provide great convenience in the handling of data. However, the register-driven mechanical valves have advantages in simplicity of design, field maintenance or repair, and initial cost.

Gas PD meters can be equipped with mechanical registers that correct the totals for pressure and temperature. The pressure and temperature compensation range is limited in the mechanical units but with appropriate application; the range limitation is not a problem. Electronic totalizers for gas PD meters can correct for broad ranges of temperature and pressure. Electronic accessories can provide various data storage, logging, and remote data access capabilities.

53.4.2 Price and Performance

In general, the accuracy of the PD meter is reflected in the sales price. However, several additional factors influence the cost. The intended application, the materials of construction, and the pressure rating have a strong influence on the cost. Although they provide excellent accuracy, small PD meters for residential water or residential natural gas service are very inexpensive. Meters produced for the general industrial market provide good accuracy at a reasonable cost, and meters manufactured for the pipeline or petrochemical markets provide excellent accuracy and are expensive.

Very inexpensive PD meters are available with plastic cases or plastic internal mechanisms. Aluminum, brass, ductile-iron, or cast-iron PD meters are moderately priced. Steel PD meters are relatively expensive, especially for high-pressure applications. If stainless steel or unusual materials are necessary for corrosive fluids or special service requirements, the costs are very high. PD meters are manufactured in a wide range of sizes, pressure ratings, and materials, with a multitude of flow-rate ranges and with an accuracy to match most requirements. The cost is directly related to the performance requirements.

Further Information

API, Measurement of liquid hydrocarbons by displacement meters, *Manual of Petroleum Measurement Standard*, Chapter 5.2, Washington, DC: American Petroleum Institute, 1992.

Bean, H. S., *Fluid Meters: Their Theory and Application*, 6th ed., Report of ASME Research Committee on Fluid Meters, New York: American Society of Mechanical Engineers, 1971.

Mechanical Flowmeters, Flow and Level Measurement. http://www.omega.com/literature/transactions/volume4/T9904–08-MECH.html (accessed on June 23, 2012).

Miller, R. W., *Flow Measurement Engineering Handbook*, New York: McGraw-Hill, 1989.

SmartMeasurement, Positive displacement. http://www.smartmeasurement.com/en/documents/flowmeter/PositiveDisplacement.asp (accessed on June 21, 2012).

Spitzer, D. W. (Ed.), *Flow Measurement: Practical Guides for Measurement and Control*, Research Triangle Park, NC: ISA, 1991.

Streeter, V. L., *Fluid Mechanics*, 4th ed., New York: McGraw-Hill, 1966.

Streeter, V. L., *Handbook of Fluid Dynamics*, New York: McGraw-Hill, 1961.

Universal Flow Monitors, Positive displacement flowmeters. http://www.flowmeters.com/ufm/index.cfm?task=positive_displacement (accessed on June 11, 2012).

Partial List of Vendors and Suppliers

Controls Warehouse, 356 Cypress Rd., Ocala, FL 34472, Phone: 888-212-5531 (toll free), Fax: 352-687-8925, www.controlswarehouse.com

GPI Meters, P.O. Box 1008, Sparta, NJ 07871, Phone: 888-722-3569 (toll free), Fax: 973-940-1651, www.gpimeters.com/

Great Plains Industries, 5252 E. 36th St. N., Wichita, KS 67220-3205, Phone: 316-686-7361, 800-835-0113 (toll free), Fax: 316-686-6746, www.gpi.net

Imac Systems, Inc., 90 Main St., Tullytown, PA 19007, Phone: 215-946-2200, 800-955-4427 (toll free), 800-955-4GAS (toll free), Fax: 215-943-2984, www.imacsystems.com

Ives Equipment Corp., 601 Croton Rd., King of Prussia, PA 19406-3111, Phone: 610-265-3071, 877-768-1600 (toll free), Fax: 610-768-1605, www.ivesequipment.com

Keco Engineering Controls, 1200 River Ave., Bldg. 3A, Lakewood, NJ 08701-5657, Phone: 732-901-5900, Fax: 732-901-5904, www.kecocontrols.com



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

54

Turbine and Vane Flowmeters

54.1 Axial Turbine Flowmeters **54**-1
General Performance Characteristics • Theory • Calibration, Installation, and Maintenance • Design and Construction • Propeller Meters • Helical Turbine Meters • Swirl-Type Axial Turbine Meters • Dual-Rotor Axial Turbines
54.2 Insertion Axial Turbine Flowmeters **54**-10
54.3 Angular Momentum Flowmeters **54**-11
54.4 Multijet Turbine Flowmeters **54**-12
54.5 Cylindrical Rotor Flowmeter **54**-12
54.6 Manufacturers **54**-12
References **54**-15
Standards and Recommendations Applicable in the United States **54**-16

David Wadlow
Consultant

This section describes a range of closed-conduit flowmeters that utilize rotating vaned transduction elements, with particular emphasis on axial turbine flowmeters. The various vaned flowmeters used for open channel and free-field flow measurement along with single jet and insertion tangential turbines, also known as paddle wheel flowmeters, are not included in this section.

54.1 Axial Turbine Flowmeters

The modern axial turbine flowmeter, when properly installed and calibrated, is a reliable device capable of providing the highest accuracies attainable for both liquid and gas volumetric flow measurement. It is the product of decades of intensive innovation and refinements to the original axial vaned flowmeter principle first credited to Woltman in 1790 and at that time applied to measuring water flow. The initial impetus for the modern development activity was largely the increasing needs of the U.S. natural gas industry in the late 1940s and 1950s for a means to accurately measure the flow in large-diameter, high-pressure, interstate natural gas lines. Today, due to the tremendous success of this principle, axial turbine flowmeters of different and often proprietary designs are used for a variety of applications where accuracy, reliability, and rangeability are required in numerous major industries besides water and natural gas, including oil, petrochemical, chemical process, cryogenics, milk and beverage, aerospace, and biomedical.

Figure 54.1 is a schematic longitudinal section through the axis of symmetry depicting the key components of a typical meter. As one can see, the meter is an in-line sensor comprising a single turbine rotor, concentrically mounted on a shaft within a cylindrical housing through which the flow passes. The shaft or shaft bearings are located by end supports inside suspended upstream and downstream aerodynamic structures called diffusers, stators, or simply cones. The flow thus passes through an annular

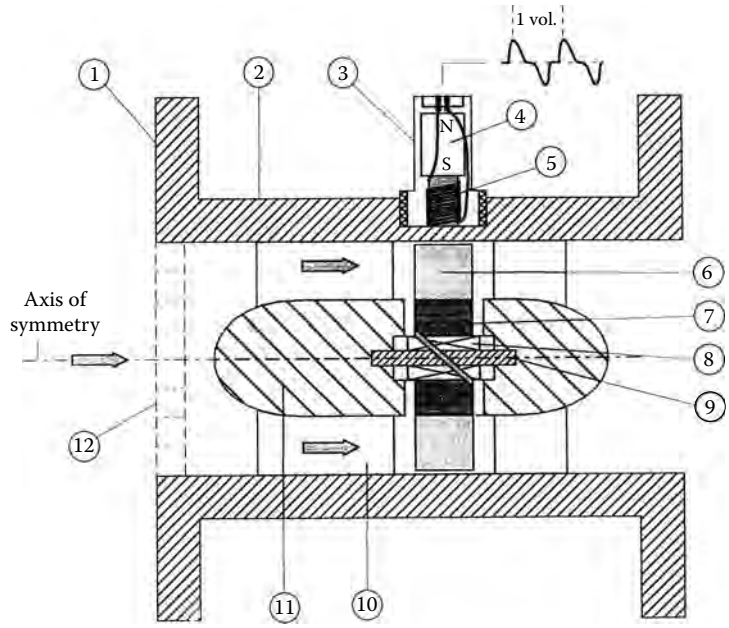


FIGURE 54.1 Longitudinal section of an axial turbine flowmeter depicting the key components. The flowmeter body is usually a magnetically transparent stainless steel such as 304. Common end fittings include face flanges (depicted), various threaded fittings, and tri-clover fittings. The upstream and downstream diffusers are the same in bidirectional meters and generally supported by three or more flat plates or sometimes tubular structures, aligned with the body, which also act as flow straighteners. The relative size of the annular flow passage at the rotor varies among different designs. Journal rotor bearings are frequently used for liquids, while ball bearings are often used for gases. Magnetic reluctance pickups (depicted) are frequently used. Others types include mechanical and modulated carrier pickups. (1) End fitting—flange shown; (2) flowmeter body; (3) rotation pickup—magnetic, reluctance type shown; (4) permanent magnet; (5) pickup cold wound on pole piece; (6) rotor blade; (7) rotor hub; (8) rotor shaft bearing—journal type shown; (9) rotor shaft; (10) diffuser support and flow straightener; (11) diffuser; (12) flow conditioning plate (dotted)—optional with some meters.

region occupied by the rotor blades. The blades, which are usually flat but can be slightly twisted, are inclined at an angle to the incident flow velocity and hence experience a torque that drives the rotor. The rate of rotation, which can be up to several $\times 10^4$ rpm for smaller meters, is detected by a pickup, which is usually a magnetic type, and registration of each rotor blade passing infers the passage of a fixed volume of fluid.

54.1.1 General Performance Characteristics

Axial turbines perform best when measuring clean, conditioned, steady flows of gases and liquids with low kinematic viscosities (below about 10^{-5} m^2 s^{-1}, 10 cSt, although they are used up to 10^{-4} m^2 s^{-1}, 100 cSt) and are linear for subsonic, turbulent flows. Under these conditions, the inherent mechanical stability of the meter design gives rise to excellent repeatability performance. Not including the special case of water meters, which are described later, the main performance characteristics are

- Sizes (internal diameter) range from 6 to 760 mm (1/4 to 30 in.).
- Maximum measurement capacities range from 0.025 to 25,500 Am^3 h^{-1} (0.015–15,000 ACFM), for gases, and 0.036–13,000 m^3 h^{-1} (0.16–57,000 gpm or 82,000 barrels per hour), for liquids, where A denotes actual.

- Typical measurement repeatability is ±0.1% of reading for liquids and ±0.25% for gases with up to ±0.02% for high-accuracy meters. Typical linearities (before electronic linearization) are between ±0.25% and ±0.5% of reading for liquids and ±0.5% and ±1.0% for gases. High-accuracy meters have linearities of ±0.15% for liquids and ±0.25% for gases, usually specified over a 10:1 dynamic range below maximum rated flow. Traceability to National Institute of Standards and Technology (NIST) is frequently available, allowing one to estimate the overall absolute accuracy performance of a flowmeter under specified conditions. Under ideal conditions, absolute accuracies for optimum designs and installations can come close to the accuracy capabilities at the NIST, which are stated as ±0.13% for liquid flows and ±0.25% for air.
- Rangeability, when defined as the ratio of flow rates over which the linearity specification applies, is typically between 10:1 and 100:1.
- Operating temperature ranges span –270 °C to 650 °C (–450 °F to 1200 °F).
- Operating pressure ranges span coarse vacuum to 414 MPa (60,000 psi).
- Pressure drop at the maximum rated flow rate ranges from around 0.3 kPa (0.05 psi) for gases to in the region of 70 kPa (10 psi) for liquids.

54.1.2 Theory

There are two approaches described in the current literature for analyzing axial turbine performance. The first approach describes the fluid driving torque in terms of momentum exchange, while the second describes it in terms of aerodynamic lift via airfoil theory. The former approach has the advantage that it readily produces analytical results describing basic operation, some of which have not appeared via airfoil analysis. The latter approach has the advantage that it allows more complete descriptions using fewer approximations. However, it is mathematically intensive and leads rapidly into computer-generated solutions. One prominent pioneer of the momentum approach is Lee (Lee and Evans 1965), who, using this approach, later went on to invent one of the few, currently successful, dual-rotor turbine flowmeters, while Thompson and Grey (1970) later provided a comprehensive model using the airfoil approach, which, for example, took into account blade interference effects. Stoltenkamp (2007) reviewed the theory in detail and also extended the airfoil model of Thompson and Grey. In the following, the momentum exchange approach is used to highlight the basic concepts of the axial turbine flowmeter.

In a hypothetical situation, where there are no forces acting to slow down the rotor, it will rotate at a speed that exactly maintains the fluid flow velocity vector at the blade surfaces. Figure 54.2 is a vector diagram for a flat-bladed rotor with a blade pitch angle equal to β. Assuming that the rotor blades are flat and that the velocity is everywhere uniform and parallel to the rotor axis, then referring to Figure 54.2, one obtains

$$r\omega_i = V \tan\beta \tag{54.1}$$

When one introduces the total flow rate, this becomes

$$\frac{\omega_i}{Q} = \frac{\tan\beta}{\bar{r}A} \tag{54.2}$$

where

ω_i is the "ideal" rotational speed
Q is the volumetric flow rate
A is the area of the annular flow cross section
$\bar{r}$ is the root mean square of the inner and outer blade radii (R,a)

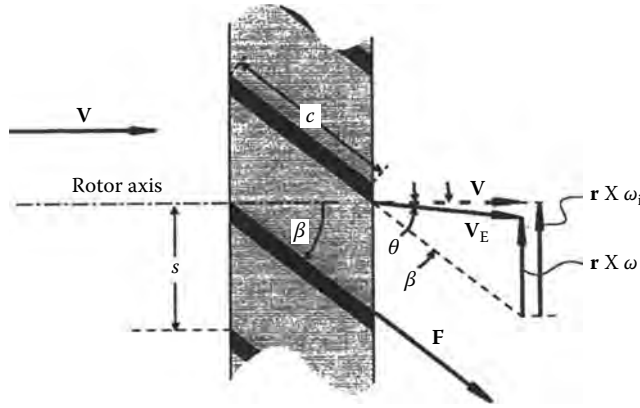


FIGURE 54.2 Vector diagram for a flat-bladed axial turbine rotor. The difference between the ideal (subscript i) and actual tangential velocity vectors is the rotor slip velocity and is caused by the net effect of the rotor retarding torques. This gives rise to linearity errors and creates swirl in the exit flow. **V**, incident fluid velocity vector; $\mathbf{V}_E$, exit fluid velocity vector; θ, exit flow swirl angle due to rotor retarding torques; β, blade pitch angle, same as angle of attack for parallel flow; ω, rotor angular velocity vector; **r**, rotor radius vector; **F**, flow-induced drag force acting on each blade surface; c, blade chord; s, blade spacing along the hub; c/s, rotor solidity factor.

Eliminating the time dimension from the left-hand-side quantity reduces it to the number of rotor rotations per unit fluid volume, which is essentially the flowmeter K factor specified by most manufacturers. Hence, according to Equation 54.2, in the ideal situation, the meter response is perfectly linear and determined only by geometry. (In some flowmeter designs, the rotor blades are helically twisted to improve efficiency. This is especially true of blades with large radius ratios (R/a). If the flow velocity profile is assumed to be flat, then the blade angle in this case can be described by $\tan\beta$ = constant × r. This is sometimes called the "ideal" helical blade.) In practice, there is instead a number of rotor retarding torques of varying relative magnitudes. Under steady flow, the rotor assumes a speed that satisfies the following equilibrium:

$$\begin{aligned}\text{Fluid driving torque} = {}& \text{rotor blade surfaces fluid drag torque} + \text{rotor hub} \\ & \text{and blade tip clearance fluid drag torque} + \text{rotation} \\ & \text{sensor drag torque} + \text{bearing friction retarding torque}\end{aligned} \tag{54.3}$$

Referring again to Figure 54.2, the difference between the actual rotor speed, $r\omega$, and the ideal rotor speed, $r\omega_i$, is the rotor slip velocity due to the combined effect of all the rotor retarding torques as described in Equation 54.3, and as a result of which, the fluid velocity vector is deflected through an exit or swirl angle, θ. Denoting the radius variable by r and equating the total rate of change of angular momentum of the fluid passing through the rotor to the retarding torque, one obtains

$$\int_a^R \frac{\rho Q 2\pi r^2 (r\omega_i - r\omega)}{\omega(R^2 - a^2)} dr = N_T, \tag{54.4}$$

which yields

$$\overline{r}^2 \rho Q(\omega_i - \omega) = N_T \tag{54.5}$$

where ρ is the fluid density and N_T is the total retarding torque. Combining Equations 54.1 and 54.5 and rearranging yields

$$\frac{\omega}{Q} = \frac{\tan\beta}{\bar{r}A} - \frac{N_T}{\bar{r}^2 \rho Q^2} \tag{54.6}$$

The trends evident in Equation 54.6 reflect the characteristic decline in meter response at very low flows and why lower friction bearings and lower drag pickups tend to be used in gas versus liquid applications and small-diameter meters. In most flowmeter designs, especially for liquids, the latter three of the four retarding torques described in Equation 54.3 are small under normal operating conditions compared with the torque due to induced drag across the blade surfaces. As shown in Figure 54.2, the force, *F*, due to this effect acts in a direction along the blade surface and has a magnitude given by

$$F = \frac{\rho V^2}{2} C_D S \tag{54.7}$$

where C_D is the drag coefficient and *S* is the blade surface area per side. Using the expression for drag coefficient corresponding to turbulent flow, as described by Pate et al. (1984) and others, this force can be estimated by

$$F = \rho V^2 0.074 \mathrm{Re}^{-0.2} S \tag{54.8}$$

where Re is the flow Reynolds number based on the blade chord shown as dimension *c* in Figure 54.2. Assuming θ is small compared with β, then after integration, the magnitude of the retarding torque due to the induced drag along the blade surfaces of a rotor with *n* blades is found to be

$$N_D = n(R+a)\rho V^2 0.037 \mathrm{Re}^{-0.2} S \sin\beta \tag{54.9}$$

Combining Equations 54.9 and 54.6 and rearranging yields

$$\frac{\omega}{Q} = \frac{\tan\beta}{\bar{r}A} - \frac{0.036n(R+a)SA^2 \mathrm{Re}^{-0.2} \sin\beta}{\bar{r}^2} \tag{54.10}$$

Equation 54.10 is an approximate expression for the *K* factor because it neglects the effects of several of the rotor retarding torques and a number of important detailed meter design and aerodynamic factors, such as rotor solidity and flow velocity profile. Nevertheless, it reveals that linearity variations under normal, specified operating conditions are a function of certain basic geometric factors and Reynolds number. These results reflect general trends that influence design and calibration. Additionally, the marked departure from an approximate ρV^2 (actually $\rho^{0.8} V^{1.8} \mu^{-0.2}$ via Equation 54.8) dependence of the fluid drag retarding torque on flow properties under turbulent flow, to other relationships under transitional and laminar flow, gives rise to major variations in the *K* factor versus flow rate and media properties for low-flow Reynolds numbers. This is the key reason why axial turbine flowmeters are generally recommended for turbulent flow measurement.

54.1.3 Calibration, Installation, and Maintenance

Axial turbine flowmeters have a working dynamic range of at least 10:1 over which the linearity is specified. The maximum flow rate is determined by design factors related to size versus maximum pressure drop and maximum rotor speed. The minimum of the range is determined by the linearity specification itself. Due to small, unavoidable, manufacturing variances, linearity error curves are unique to

individual meters and are normally provided by the manufacturer. However, although recommended where possible, the conditions of the application cannot usually and need not necessarily duplicate those of the initial or even subsequent calibrations. This has pivotal importance in applications where actual operating conditions are extreme or the medium is expensive or difficult to handle. Figure 54.3 depicts a typically shaped calibration curve of linearity versus flow rate expressed in terms of multiple alternative measures, various combinations of which can be found in current use. The vertical axis thus represents either the linearity error as a percentage of flow rate, a *K* factor expressed in terms of the number of pulses from the rotation sensor output per volume of fluid, or the deviation from 100% registration; the latter only applies to flowmeters with mechanical pickups. The horizontal axis can be expressed in terms of flow rate in volume units/time, Reynolds number (Re), or pulse frequency (from the rotation sensor for nonmechanical) divided by kinematic viscosity, (f/ν), in units of Hz per $m^2\ s^{-1}$ (Hz/cSt or Hz/SSU; $10^{-6}\ m^2\ s^{-1}$ = 1 centistoke = 31.0 seconds, Saybolt Universal), where kinematic viscosity is the ratio of absolute viscosity (μ) to density. Calibrations are preferably expressed versus Re or f/ν, which is proportional to Re. The hump shown in the curve is a characteristic frequently observed at lower Re and is due to velocity profile effects. *K* factor versus f/ν calibration curves are specifically called universal viscosity curves (UVC) and, for most meters, are available from the manufacturer for an extra charge. A key utility of UVC is that where media type and properties differ significantly from those of the original calibration, accuracies much greater than the overall linearity error can still readily be obtained via the flowmeter UVC if the kinematic viscosity of the application is known. An alternative calibration technique (Olivier and Ruffner 1992) is to provide response in terms of Strouhal number versus Re

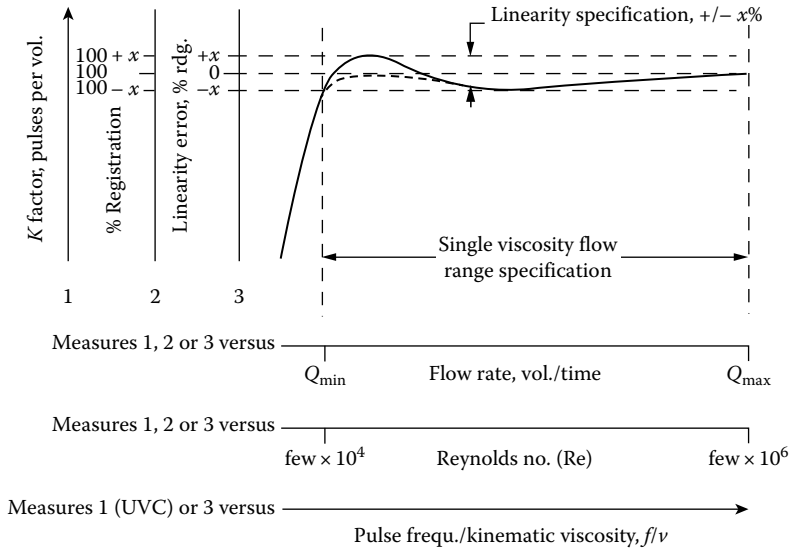


FIGURE 54.3 A typical single-rotor axial turbine linearity error, or calibration, curve for a low-viscosity fluid showing the main alternative presentations in current use. Higher-accuracy specifications usually correspond to a 10:1 flow range down from Q_{max}, while extended operating ranges usually correspond to reduced accuracies. The hump in the depicted curve is a characteristic feature caused by flow velocity profile changes as Re approaches the laminar region. This feature varies in magnitude between meters. Sensitivity and repeatability performance degrade at low Re. Percent registration is only used with meters that have mechanical pickups. All other meters have a *K* factor. UVC and Re calibrations remain in effect at different known media viscosities provided Re or f/ν stays within the specified range. Re is referenced to the connecting conduit diameter and is less within the flowmeter. The Re range shown is therefore approximate and can vary by an order of magnitude, depending on the meter. Linearity error can also be expressed in terms of Strouhal number (fD/V) versus Re (VD/υ) or Roshko number (fD^2/υ), when instead *D* is a flowmeter reference diameter [4]. UVC — the effect of a rotor shroud in a viscosity compensated flowmeter.

or Roshko number. This approach is particularly relevant to high-accuracy and extreme temperature applications because it further allows correct compensation for flowmeter thermal expansion errors.

The accuracy of axial turbine flowmeters is reduced by unconditioned flow, especially swirl. An installation incorporating flow conditioners along with specific upstream and downstream straight pipe lengths is generally recommended. See for instance ISA-RP 31.1, 1977. Some axial turbine flowmeters can be purchased with additional large flow straighteners that mount directly ahead of the flowmeter body or conditioning plates that are integral to the body. The manufacturer is the first source of information regarding installation. Errors due to flow velocity pulsations are another concern, particularly in natural gas installations. Flow pulsation errors tend to be positive because turbine rotors tend to speed up faster than they slow down. However, although the problem of flow pulsation-induced errors has been carefully studied, no standard technique for effectively counteracting this source of error has yet been adopted (Jungowski and Weiss 1996, Laurantzon et al. 2012). Periodic maintenance, testing, and recalibration are required because the calibration will shift over time due to wear, damage, or contamination. For certain applications, especially those involving custody transfers of oil and natural gas, national standards, international standards, and other recommendations exist that specify the minimum requirements for turbine meters with respect to these aspects. A number of these are identified separately in the reference section.

54.1.4 Design and Construction

There are numerous, often proprietary, designs incorporating variations in rotors, bearings, pickups, and other components in format and materials that are tailored to different applications. Meter bodies are available with a wide range of standard end fittings. Within application constraints, the primary objective is usually to optimize the overall mechanical stability and fit in order to achieve good repeatability performance. Design for performance, application, and manufacture considerations impacts every internal component but most of all the rotor with respect to blade shape and pitch; blade count; balance and rigidity versus drag, stress, and inertia; bearings with respect to precision versus friction, speed rating, and durability; and rotation pickup versus performance and drag.

Most low-radius ratio blades are machined flat, while high-ratio blades tend to be twisted. The blade count varies from about 6 to 20 or more, depending on the pitch angle and blade radius ratio so that the required rotor solidity is achieved. Rotor solidity is a measure of the "openness" to the flow such that higher solidity rotors are more highly coupled to the flow and achieve a better dynamic range. The pitch angle, which primarily determines the rotor speed, is typically 30° to 45° but can be lower in flowmeters designed for low density, gas applications. Rotor assemblies are usually a close fit to the inside of the housing. In large diameter meters, the rotor often incorporates a shroud around the outer perimeter for enhanced stability. Also, since large meters are often used for heavy petroleum products, via selection of a suitable wall clearance, the fluid drag resulting from this clearance gap is often designed to offset the tendency at high media viscosities for the meter to speed up at lower Reynolds numbers. The materials of construction range from nonmagnetic to magnetic steels to plastics.

Stainless steel ball bearings are traditionally used for gas meters and low-lubricity liquids, while hybrid and full ceramic ball bearings may be preferred for corrosive and cryogenics applications. Combination tungsten carbide, graphite, ceramic, and sometimes Teflon journal and thrust bearings are also available for many liquid meters with material selection depending on the medium lubricity and corrosiveness. Fluid bearings (sometimes called "bearingless" designs) are often used in conjunction with the latter, but also sometimes with gases, for reducing the drag. They operate by various designs that use flow-induced forces to balance the rotor away from the shaft ends. Either bearing lubrication is derived from the metered medium or an internal or external system is provided. The more fragile, jeweled pivot bearings are also used in certain gas applications and small meters. Sanitary meters can incorporate flush holes in the bearing assembly to meet 3A crack and crevice standards.

The most common types of rotation sensor are magnetic, modulated carrier, and mechanical, while optical, capacitive, and electric resistance are also used. In research, a modulated nuclear radiation flux rotation sensor for use in certain nuclear reactors has also been reported (Van Der Hagen 1993, Termaat et al. 1995). Mechanical pickups, which sometimes incorporate a magnetic coupling, are traditional in some applications and can have high resolution; one advantage is that they require no electric power. However, the pickup drag tends to be high. The magnetic and modulated carrier types utilize at least a coil in a pickup assembly that screws into the meter housing near the rotor. In magnetic inductance types, which are now less common, the blades or shroud carries magnetized inserts, and signals are induced in the coil by the traversing magnetic fields. In the more prevalent magnetic reluctance type, an example of which is schematically depicted in Figure 54.1, the coil is wrapped around a permanent magnet or magnet pole piece in the pickup assembly, which is mounted next to a high magnetic permeability bladed rotor (or machined shroud). The latter is then typically made of a magnetic grade of stainless steel such as 416, 430, or 17–4Ph. As the rotor turns, the reluctance of the magnetic circuit varies, producing signals at the coil. In the more expensive modulated carrier types, the rotor need only be electrically conductive. The coil is part of a radio-frequency (RF) oscillator circuit, and the proximity of the rotor blades changes the circuit impedance, giving rise to modulation at a lower frequency that is recovered. The RF types have much lower drag, have higher signal levels at low flow, and can operate at temperatures above the Curie point of typical ferromagnetic materials. They are preferred for wide dynamic range and high-temperature applications. Bidirectional flowmeters usually have two magnetic pickups to determine flow direction. This is useful, for example, in the monitoring of container filling and emptying operations often encountered in sanitary applications. Multiple magnetic pickups are also used in some designs to provide increased measurement resolution. Regarding output, various pulse amplifiers, totalizers, and flow computers, for instance gas pressure and temperature correction, along with 4–20 mA and other standard interface protocols, are available to suit particular applications. As an example of advanced transmitters, at least one manufacturer (FTI Flow Technology, Inc.) provides a real-time, miniature, reprogrammable transmitter that is integrated into the pickup housing along with a meter body temperature sensor, for full viscosity compensation and UVC linearization. These are sometimes used in dedicated applications, such as airborne fuel management, where the medium viscosity–temperature relationship is known.

Certain applications have uniquely different design requirements and solutions. Three important variations are discussed in the following.

54.1.5 Propeller Meters

Propeller meters are used in municipal, irrigation, and wastewater measurement, although there is increasing competition in those applications from electromagnetic flowmeters. In some designs, propeller and turbine meters look almost identical and operate on the same axial rotor principle. However, this type of flowmeter is currently commercially and officially [AWWA C704] distinguished as a separate category distinct from the axial turbine. Diameters up to 2440 mm (96 in.) are available. The flow rate capacity of a 1800 mm (72 in.) diameter propeller meter is up to about 25,000 $m^3 h^{-1}$ (110,000 gpm). Typical accuracies are ±2% of reading. A primary requirement is ruggedness, and it is in the designs most suited to harsh environments that the formats are most distinctive. Rotor and pickup assemblies are generally flanged to the housing and removable. The rotors have large clearances, are often cantilevered into the flow, and are supported via a sealed bearing without stators. The rotors are typically made of plastic or rubber and carry as few as three highly twisted, high-radius ratio blades. Pickups are always mechanical and frequently have magnetic couplings.

54.1.6 Helical Turbine Meters

Instead of a flat or slightly twisted multibladed rotor, these meters incorporate a helical, high-radius ratio, two-bladed rotor. This design provides greater accuracy when used to meter higher-viscosity liquids compared with that of the more conventional turbine. Dynamic range is also improved. The reason

for this general improvement is discussed by Mitchell (2010). These meters are sold as helical turbine meters for use in liquid media, and accuracies can be as high as ±0.15%. Applications include the custody transfer of crude oil. Where viscosity can vary, accuracy can be maintained by means of a separate real-time viscosity measurement, flow computer, and UVC correction. When used to meter crude oil, this design has the added advantage of being less susceptible to fouling by wax and other filamentary debris often found in crude oil. The meters are also used for metering chemicals such as ammonia, ethylene, propylene, and acids.

54.1.7 Swirl-Type Axial Turbine Meters

This description is solely due to the author and is in order to distinguish this design. A number of companies manufacture swirl-type axial turbine flowmeters, and the design departs significantly from that of the traditional axial turbine. In these devices, a fixed upstream stator intentionally imparts a high degree of azimuthal velocity or swirl to the flow. Immediately downstream of the stator is an axially mounted rotor where the blades are flat and parallel to the axis. The rotor is driven into rotation by the swirling flow and at a rate that is proportional to the volumetric flow rate. Upstream flow conditioning is typically absent, and two-bladed rotor variants of this design find application as the sensor in diagnostic spirometers where such a conditioning requirement would otherwise be restrictive. Diagnostic spirometers are used to monitor the degree and nature of respiration. With these, a clinician can determine patient respiratory condition by various measures and clinical maneuvers. Low cost, light weight, speed of response, and patient safety are major considerations. Measurement capabilities include the gas volume of a single exhalation and also the peak expiratory flow, measured in liters and liters per second, respectively. A number of manufacturers also provide small, portable diagnostic spirometers for personal use by patients at home. The spirometry market is large and competitive, and other flow sensing technologies are also used in spirometry. However, (swirl-type) axial turbine-based diagnostic spirometers command a significant market share. The devices typically utilize an infrared, optical pickup and have a microprocessor-controlled display. Rotors and housings tend to be plastic, and the entire turbine element may also be disposable. The meters are typically accurate to ± a few percent of reading. In the United States, spirometers are designated as class 2 medical devices, and as such certain FDA approvals are required concerning manufacture and marketing. In the EU, they are class IIb medical devices under a different system, and other approvals are required.

One manufacturer (Equflow BV) makes a range of sub-inch diameter (swirl-type) axial turbines for single or multiple use liquid applications ranging from fuel consumption measurement and biotechnology applications to corrosive liquids with accuracies in the region of 1% of reading. In these particular devices, the rotor has six blades. Housings and wetted parts may be perfluoroalkoxy perfluoroalkoxy alkanes (PFA)/Teflon®, polyvinylidene difluoride (PVDF), or in combination with stainless steel, while all models feature ruby bearings. The pickup is infrared where instead the pulse signal is obtained via reflection from a region at the tip of each blade as it passes the sensor. By this means, the flow of opaque liquids can also be monitored.

54.1.8 Dual-Rotor Axial Turbines

Dual-rotor axial turbines have performance features not found in single-rotor designs. In 1981, Lee et al. were issued a U.S. patent for a self-correcting, self-checking dual-rotor turbine flowmeter that is currently manufactured exclusively by Sensus Metering Systems and sold as the Auto-Adjust. This is a high-accuracy flowmeter primarily intended for use on large natural gas lines where even small undetected flow measurement errors can be costly. It incorporates two closely coupled turbine rotors that rotate in the same direction. The upstream rotor is the main rotor, and the second rotor, which has a much shallower blade angle, is the sensor rotor. Continuous and automatic correction of measurement errors due to varying bearing friction is achieved by calculating the flow rate based on the

difference between the rotor speeds. As shown in Figure 54.2 and discussed in the theory section, the flow exit angle is due to the net rotor retarding torque. If this torque increases in the main rotor, thereby reducing its speed, the exit angle increases and the speed of the sensor rotor is then also reduced. The meter is also insensitive to inlet swirl angle because the swirl affects both rotor speeds in the same sense and the effect is then subtracted in the flow calculation. The meter also checks itself for wear and faults by monitoring the ratio of the two rotor speeds and comparing this number with the installation value (Lee et al. 1982).

A dual-rotor liquid flowmeter, invented by Ruffner and Olivier (1997), is manufactured and sold by Exact Flow. It is a high-accuracy flowmeter (up to ±0.25% accuracy and ±0.02% repeatability), which has an extraordinarily wide dynamic range of 500:1 with a single-viscosity liquid. The meter comprises two, closely and hydraulically coupled helical rotors that rotate in opposite directions. Due to the exit angle generated by the first rotor, the second rotor continues to rotate to much lower flow rates compared with the first. Flow conditioning is not essential since swirl impacts both rotors. However, if the flow is conditioned, then changes in the ratio of the two rotor speeds provide a useful bearing diagnostic. This flowmeter is intended for liquid measurement only, in particular fuel and other measurements in aerospace applications.

54.2 Insertion Axial Turbine Flowmeters

These flowmeters comprise a small axial rotor mounted on a stem that is inserted radially through the conduit wall, often through a shutoff valve. They measure the flow velocity at the rotor position from which the volumetric flow rate is inferred. They are an economical solution to flow measurement problems where pipe diameters are high and accuracy requirements are moderate and also may be technically preferred where negligible pressure drop is an advantage, as in high-speed flows. They are typically more linear than insertion tangential turbine flowmeters and compete also with magnetic and vortex shedding insertion flowmeters. They are available for the measurement of a range of liquids and gases, including steam, similar to the media range of full-bore axial turbines, and have a similarly linear response. The rotors, which are usually metal but can be plastic, typically have diameters of 25–51 mm (1–2 in.). They can be inserted into pipes with diameters ranging from 51 to 2032 mm (2–80 in.). Velocity measurement ranges cover 0.04691 ms^{-1} (9–18,000 fpm) for gases and 0.03–30 ms^{-1} (6–6,000 fpm) for liquids. Dynamic ranges vary between 10:1 and 100:1. The maximum flow rate measurement capacity in a 1,836 mm (72 in.) diameter pipe can be as high as nearly 56,500 $m^3 h^{-1}$ (about 250,000 gpm). Since these devices are local velocity sensors, calculating the volumetric flow rate requires knowledge of the area velocity profile and the actual flow area. Flow conditioning is therefore particularly important for accurate volumetric measurements, while radial positioning, which is a further responsibility of the user, must be according to the manufacturer's recommendation, which can either be centerline, one-third of the diameter, 12% of the diameter, or determined by "profiling." Quick (1996) discussed operations and installations for natural gas measurements. Although linearities or "accuracies" can be quoted up to ±1% of velocity, achieving the same accuracy for the volumetric flow rate, although possible, can be difficult or impractical. In this respect, a unique dual-rotor design, exclusive to Onicon, Inc., and primarily used for chilled-water flow measurement in HVAC systems, requires less flow conditioning than single-rotor designs. An Onicon dual-rotor turbine assembly is depicted in Figure 54.4. It comprises two rotors that rotate in opposite directions. The output is based on the average rotor speed. Any flow swirl present due to poor flow conditioning changes the speed of rotation of each rotor by the same but opposite amounts. Swirl-induced error is thus virtually absent in the averaged output. Also, flow profile sampling is improved over that of a single rotor. The devices are calibrated using a volumetric prover, and the specified accuracy of ±2% of reading is for volumetric flow rate rather than velocity. This is the total error and includes an allowance for dimensional variations in industry standard pipes.



FIGURE 54.4 The rotor assembly of a dual-rotor, insertion axial turbine flowmeter for water flow measurement. This patented design renders the flowmeter insensitive to errors due to flow swirl; an important source of potential error in single-rotor axial turbine flowmeters. The rotations are sensed by two separate, electric impedance sensors. (Courtesy of Onicon Incorporated.)

54.3 Angular Momentum Flowmeters

These are accurate, linear, liquid mass flowmeters that utilize vaned components and are used in aerospace applications. They are currently the instrument of choice for airborne, jet engine fuel flow measurement for large commercial aircraft and some military aircraft for afterburner applications. They are more expensive than the equivalent turbine flowmeters for this application, but they provide mass flow rate measurements directly and are unaffected by fuel density variations. Typical accuracies lie between ±0.5% and ±1.0% over a 40:1 or greater dynamic range. Measurement ranges are available for fuel flows from about 0.01 to 6 kg s^{-1} (70–46,000 PPH).

The principle of operation is long established and based on imparting angular momentum to the fluid flow using a driven, flat-bladed impeller. The force required to drive the impeller at constant speed is monitored as a proportional indication of mass flow rate as this quantity varies. Some designs use an electric motor to drive the impeller. However, the current trend in design is motorless. In such a device, a constant driving speed in the region of 100–200 rpm is provided by one or more turbine rotors driven by the flow. A variable shunt metering valve assembly adjacent to the turbine rotor mechanically opens and closes in response to flow dynamic pressure and thereby automatically maintains a constant speed of rotation provided that the flow rate is above the minimum of the range. The driven impeller carries vanes that are parallel to the flow and resides on a common axis with the turbine rotor. A flow straightener ensures parallel flow past the impeller. There is a carefully engineered constant rate spring connection between the turbine shaft and the impeller so that the angular deflection between the two is proportional to the applied torque, and this quantity is directly proportional to the mass flow rate. Two pickup coils sense the rotations of the turbine and the vaned impeller, and only the time difference between the pulses in these two signals is measured and used to calculate the flow rate. Chiles et al. (1977) provide a detailed illustration and explanation of the intricate mechanism.

54.4 Multijet Turbine Flowmeters

These are linear, volumetric flowmeters designed for liquid measurements and comprise a single, radial vaned impeller, vertically mounted on a shaft bearing within a vertically divided flow chamber, sometimes called a distributor. The impeller is often plastic and can even be neutrally buoyant in water. There are various designs, but typically, both chambers access a series of radially distributed and angled jets. The lower chamber jets connect to the flowmeter input port and distribute the flow tangentially onto the lower region of the impeller blades, while the upper series, which are angled oppositely, allow the flow to exit. The flow pattern within the flow chamber is thus a vertical spiral, and the dynamic pressure drives the impeller to track the flow. This design gives the meters good sensitivity at low flow rates. Due to the distribution principle, the meters are also insensitive to upstream flow condition. Impeller rotation pickups are always mechanical, often magnetically coupled, and frequently also connect with electric contact transmitters. They are primarily used in water measurement, including potable water measurement for domestic and business billing purposes and in conjunction with energy management systems such as hot water building or district heating, in water flood projects for secondary oil recovery, and to a lesser extent in some chemical and pharmaceutical industries for dosing and filling systems involving solvents, refrigerants, acids, and alkalis with absolute viscosities less than 4.5 MPa s (0.045 Poise). Available sizes range from 15 to 50 mm. Dynamic ranges lie between 25:1 and 130:1, and flow measurement ranges cover 0.03–30 $m^3 h^{-1}$ (0.13–130 gpm). Measurement linearities range between ±1% and ±2%, with typical repeatabilities of ±0.3%. Operating temperatures range from normal to 90 °C (200°F), and maximum operating pressures are available up to 6.9 MPa (1000 psi). A number of potable water measurement systems come with sophisticated telemetry options that allow remote interrogation by radio or telephone. For potable water applications in the United States, these meters normally comply with the applicable American Water Works Association (AWWA) standard, while in Europe other national standards apply.

54.5 Cylindrical Rotor Flowmeter

Instead of coupling to a turbulent flow using the fluid dynamic pressure or momentum flux, as in most turbine flowmeters, a demonstrated research device due to Wadlow and Layden (1994) provides a linear, volumetric, low gas flow rate measurement using a single, low-inertia, smooth cylindrical tangential rotor that couples to a laminar flow via surface friction. The geometry is conceptually that of a single-surfaced rotating vane. A plane Poiseuille flow is created that passes azimuthally over most of the curved rotor surface in a narrow annular passage between the rotor and a concentric housing. The rotor is motor driven via a feedback control loop connected to an error signal producing differential pressure sensor connected across the gas ports so as to maintain the meter pressure drop equal to zero. Under this condition, the response, as indicated by motor shaft speed, is exactly linear, determined only by geometry, and is independent of gas density and viscosity. The demonstration device reported has a 40 mm diameter rotor and 15 mm diameter gas ports. It measures up to a maximum of 25 Lpm bidirectionally, has a linear dynamic range greater than 100:1, is insensitive to upstream flow condition, and has a $1/e$ step response time of 42 ms, limited only by the external motor torque and combined motor and rotor inertia.

54.6 Manufacturers

There is significant and dynamic competition among the numerous manufacturers of the different types of flowmeters described in this section. In all flowmeter types, most manufacturers have exclusive patent rights concerning one or more detailed design aspects that make their products perform differently

from those of the competition. Every few years, one or more major turbine flowmeter company can be identified that has changed ownership and name or formed a new partnership. Identifying the competition and selecting the manufacturer are important and sometimes time-consuming parts of the flowmeter selection and specification process. To assist with this, Table 54.1 gives a few selected examples from different manufacturers, of all the different flowmeter types described in this section. Table 54.2 gives the corresponding contact information for those selected manufacturers, along with the general types offered by each.

TABLE 54.1 Examples of turbine and vaned flowmeters

Type	Size(s), inches	Description	Example application(s)	Manufacturer	Approx. price/ range, USD
Axial	0.5	FT 4-8 with linear link transmitter	Aerospace; marine – fuel management	FTI Flow Technology	2,500
Axial	18	Daniel Series 1500	Pipeline, oil	Emerson Process Measurement	58,000 starting price
Axial	1	3A Sanitary Series	Raw milk	Sponsler	2,800
Propeller	8	MW508	Municipal water	McCrometer	1,750
Helical	3 – 16	Smith Meter® MV Series Turbine Meter	Custody transfer and volume balance, oil	*FMC Technologies	Not available
Swirl-type axial	1 approx.	FlowMIR® Disposable turbine, code 910004	Spirometry	MIR	75 for box of 50
Swirl-type axial	¼	Disposable PFA turbine model 0085	Pharmaceutical industry	Equflow	72 each
Axial dual rotor	12	Auto-Adjust II Turbo-Meter; AAT230	Custody transfer, gas	Sensus	44,000 starting price
Axial dual rotor	1	DX/DL 16 Series	Flow transfer standard	Exact Flow	2,400
Axial insertion	1.5 – 12	TX101	Irrigation	Seametrics	450
Axial insertion	1.5 - 46	ITM100	Natural gas, steam, water, compressed air.	InFlow, Inc.	3,000 – 3,500
Axial insertion dual rotor	2.5 - 72	F-1200	Water - HVAC	Onicon	1,062 – 1,239
Angular momentum	-	Model 9-217 True Mass Fuel Flowmeter	Large air transports	Eldec	15,000 - 30,000
Multi-jet	¾ - 2	MJ Series	Chemical dosing – cooling tower water treatment.	Seametrics	140 – 500
Multi-jet	1	Model 200	Water flood projects for secondary oil recovery.	Blancett	768
Multi-jet	5/8 – 2	PMN Nitro 1 Series	Potable cold water	Zenner	57 - 721

*Marked line entry ©FMC Technologies, All rights reserved.

TABLE 54.2 Manufacturer Contact Information

Manufacturer	Relevant Types
FTI Flow Technology Inc. 8930 South Beck Ave Suite #107 Tempe, AZ 85284 (480) 240 3400	Axial turbines—liquid and gas, including sanitary
Daniel Measurement and Control Emerson Process Measurements 11100 Brittmoore Park Drive Houston, TX 77041 (713) 467 6000	Axial turbines—liquid and gas
Liquid Controls Sponsler, Inc. A Unit of IDEX Corporation. 105 Albrecht Drive Lake Bluff, IL 60044 (847) 295 1050	Axial turbines—liquid and gas, including sanitary Insertion axial turbines—liquid and gas
McCrometer 3255 West Stetson Ave. Hemet, CA 92545 (951) 652 6811	Propeller meters
FMC Technologies, Inc. 500 N. Sam Houston Pkwy West Suite 100 Houston, TX 77067 (281) 260 2190	Axial turbines—liquid and gas, including sanitary Helical turbine meters
Sensus Metering Systems 8601 Six Forks Road Suites 300 and 700 Raleigh, NC 27615 (800) 375 8875	Axial turbines—gas Dual-rotor axial turbines—gas
Exact Flow 15555 North 79th Place Scottsdale, AZ 85260 (480) 948–3789	Dual-rotor axial turbines—liquid
Seametrics Inc. P.O. Box 1589 Kent, WA 98035 (800).975.8153	Axial turbines—water Insertion axial turbines—water Multijet turbines
InFlow, Inc. 1525 Skyline Lane. Longmont, CO 80501 (303) 652 0444	Insertion axial turbines—liquid and gas
Onicon, Inc. 1500 N. Belcher Road Clearwater, FL 33765 (727) 447–6140	Axial turbines—liquid Insertion axial turbines, single and dual rotor—liquid

TABLE 54.2 (continued) Manufacturer Contact Information

Manufacturer		Relevant Types
Eldec, A Crane Company 16700 13th Avenue W. Lynnwood, WA 98037 (800) 464 0261		Angular momentum flowmeters
Blancett (Div. Racine Federated, Inc.) 8635 Washington Avenue Racine, WI 53406 (800) 235 1638		Axial turbines—liquid and gas, including sanitary Multijet turbines
Zenner Performance P.O. Box 895 Banning, CA 92220 (951) 849 8822		Multijet turbines
Equflow B.V. Postbus 35 5370 AA Ravenstein Langakker 5F 5371 EV Ravenstein The Netherlands (31) 24 3792666	In the United States Clark Solutions 10 Brent Drive Hudson, MA 01749 (978) 568 3400 HP-NA High Purity Unit #4, 25 Thurber Blvd. Smithfield, RI 02917 (401) 349 4477	Swirl-type axial turbines—liquid
MIR Medical International Research S.r.l. via del Maggiolino, 125 00155 Rome, Italy (+39) 0622754777	MIR Medical International Research Inc. 1900 Pewaukee Road Suite O Waukesha, WI 53188 (262)565-6797	Swirl-type axial turbines—diagnostic spirometers (Use a disposable turbine)

References

Chiles, W. E., Vetsch, L. E., and Peterson, J. V., Shrouded flowmeter turbine and improved fluid flowmeter using the same, U.S. Patent No. 4,012,957, 1977.

ISA-RP 31.1, *Specification, Installation and Calibration of Turbine Flowmeters*, Research Triangle Park, NC: ISA, 1977.

Jungowski, W. M. and Weiss, M. H., Effects of flow pulsation on a single-rotor turbine meter, *Trans. ASME J. Fluids Eng.*, 118(1), 198–201, 1996.

Laurantzon, F., Orlu, R., Tillmark, N. et al., Response of common flowmeters to unsteady flow, *Technical Report in Flow Measurements Related to Gas Exchange*, pp. 135–166, Royal Inst. Techn., KTH Mechanics, SE-100 44 Stockholm, Sweden, 2012.

Lee, W. F. Z. and Evans, H. J., Density effect and Reynolds number effect on gas turbine flowmeters, *Trans. ASME J. Basic Eng.*, 87(4), 1043–1057, 1965.

Lee, W. F. Z., White, R. V., Sciulli, F. M. et al., Self-correcting self-checking turbine meter, U.S. Patent No. 4,305,281, 1981.

Lee, W. F. Z., Blakeslee, D. C., and White R. V., A self-correcting and self-checking gas turbine meter, *Trans. ASME J. Fluids Eng.*, 104, 143–149, 1982.

Miller, M.R. et al, Standardisation of Spirometry, *Eur. Respir. J.* 2005; Volume 26 number 2: pp.319–338.

Mitchell, P. R., Helical turbine meters for liquid measurement, *Proc. 85th International School of Hydrocarbon Measurement*, 2010.

Olivier P. D. and Ruffner, D., Improved turbine meter accuracy by utilization of dimensionless data, *Proc. 1992 National Conf. Standards Labs*. (NCSL) Workshop and Symp., Boulder, CO: NCSL, 1992, pp. 595–607.

Pate, M. B., Myklebust, A., and Cole, J. H., A computer simulation of the turbine flow meter rotor as a drag body, *Proc. Int. Comput. in Eng. Conf. and Exhibit 1984*, pp. 184–191, Las Vegas, New York: ASME, 1984.

Quick, L. A., Gas measurement by insertion turbine meter, *Proc. 71st Int. School Hydrocarbon Meas.*, 1996.

Stoltenkamp, P. W., Dynamics of turbine flow meters, PhD Thesis, Eindhoven University of Technology, Eindhoven, the Netherlands, 2007.

Termaat, K., Oosterkamp, W. J., and Nissen, W., Nuclear turbine coolant flow meter, U.S. Patent No. 5,425,064, 1995.

Thompson, R. E. and Grey, J., Turbine flowmeter performance model, *Trans. ASME J. Basic Eng.*, 92(4), 712–723, 1970.

Van Der Hagen, T. H. J. J., Proof of principle of a nuclear turbine flowmeter, *Nucl. Technol.*, 102(2), 167–176, 1993.

Wadlow, D. and Layden, L. M., Controlled flow volumetric flowmeter, U.S. Patent No. 5,284,053, 1994.

Standards and Recommendations Applicable in the United States

AGA Rep. No. 7, *Measurement of Natural Gas by Turbine Meter*, Arlington, VA: AGA (Amer. Gas Assoc.), 2006.

API MPM, *Measurement of Liquid Hydrocarbons by Turbine Meters*, 5th ed., Washington, D.C.: API (Amer. Petroleum Inst.), 2005, Chapter 5.3.

API MPMS, Manual of Petroleum Measurement Standards, *Electronic Liquid Volume Measurement Using Positive Displacement and Turbine Meters*, 1st ed., Washington, D.C.: API (Amer. Petroleum Inst.), 1998, Chapter 21.2. (Reaffirmed 2004).

ASME MFC-4M, *Measurement of Gas Flow by Turbine Meters*, 1986. (Reaffirmed 2008).

ATS, Standardization of spirometry, 1994 update, *Am. J. Respir. Care Med.,* 152, 1107–1136, 1995 is the latest version of the official U.S. guideline for spirometry generated by the American Thoracic Society (ATS). Miller, M.R. et al, (2005), as stated therein, "brings the views of the ATS and the ERS (European Respiratory Society) together in an attempt to publish standards that can be applied more widely".

AWWA C701-12, *Cold Water Meters—Turbine Type, for Customer Service*, Denver, CO: Amer. Water Works Assoc., 2012.

AWWA C704-08, *Propeller-Type Meters for Waterworks Applications*, Denver, CO: Amer. Water Works Assoc., 2008.

AWWA C708-11, *Cold-Water Meters, Multi-Jet-Type*, Denver, CO: Amer. Water Works Assoc., 2011.

Int. Recommendation OIML R 137-1, *Gas Meters*, Paris, France: Int. Organization of Legal Metrology, Edition 2006 (E).

Int. Recommendation OIML R 117-1, *Dynamic Measuring Systems for Liquids Other Than Water*, Edition 2007 (E).

ISO 4064:2005, *Measurement of Water Flow in Fully Charged Closed Conduits—Meters for Cold Potable Water and Hot Water*, Geneva, Switzerland: Int. Organization for Standardization.

ISO 9951:1993, *Measurement of Gas Flow in Closed Conduits—Turbine Meters*, Geneva, Switzerland: Int. Organization for Standardization, (also available ANSI), 1993.

NIST Handbook 44, 2012 Edition. Specifications, Tolerances, and Other Technical Requirements for Weighing and Measuring Devices, Section 3, National Institute of Standards and Technology, Office of Weights and Measures, Gaithersburg, MD.

55

Impeller Flowmeters

55.1 Introduction **55**-1
55.2 Sensing Principles **55**-3
55.3 Flow Ranges **55**-4
55.4 Installation and Applications **55**-4
Pipe Sizes • Piping System Restrictions • Operating Pressure • Calibration • Accuracy • Design Features • Areas of Application
55.5 Manufacturers **55**-6
Data Industrial • G. Fischer (Signet) • Seametrics • Roper Flow Technology • Hoffer • McMillan • Flowmetrics • Proteus
Bibliography **55**-7

Harold M. Miller
Data Industrial Corporation

55.1 Introduction

Impeller flowmeters, sometimes called *paddle wheel* meters, are one of the most commonly used flowmeter variety. The impeller flow sensor is a direct offshoot of the old-fashioned undershot waterwheel. They are a cost-effective alternative to turbine meters and can be used in applications that are difficult to handle with other types of flow metering instruments. In their mechanical construction, they are usually very simple, with any sophistication residing in the electronics used to detect the rotation rate of the impeller and in the choice of materials of construction for chemical corrosion attributes of the metered fluids. They are related to turbine meters in that both types use a rotating mechanical element to produce the output signal. They differ in the fact that the impeller meter has its rotary axis transverse to the flow stream, as opposed to the turbine meter axis, which is parallel thereto.

These devices are available in two basic types. The insertion style is the most common (see Figures 55.1b and 55.2). The sensor is directly installed into a hole in a pipe, with saddle or welded fitting installed at the entry to seal the sensor to the pipe. The sensor can also be preinstalled in appropriate tees in the pipeline. Most suppliers have designs that can be installed in operating pipe systems, with little, if any, loss of the fluid during the installation process. The impeller design is also supplied in in-line (through-flow) sensors for those applications where such use is desirable (Figures 55.1a). The in-line meters are commonly of somewhat higher accuracy than the insertion style.

In-line meters are *linear*, *inferential*, *volumetric* flowmeters for measuring *liquid and gas flows* and are more *sensitive at lower flow rates* compared with, for example, axial turbine flowmeters. (This is of course because the flow blade incidence angle is much greater.) Insertion impeller flowmeters instead measure the flow velocity in a small region within a flow conduit.

Impeller flowmeters are generally suited to much lower flow rate ranges than the same size axial turbines and hence often find applications where axial turbines cannot be used. This and a tendency for lower cost and high reliability are key strengths. Competition with in-line axial turbines can only occur in overlapping design flow ranges.

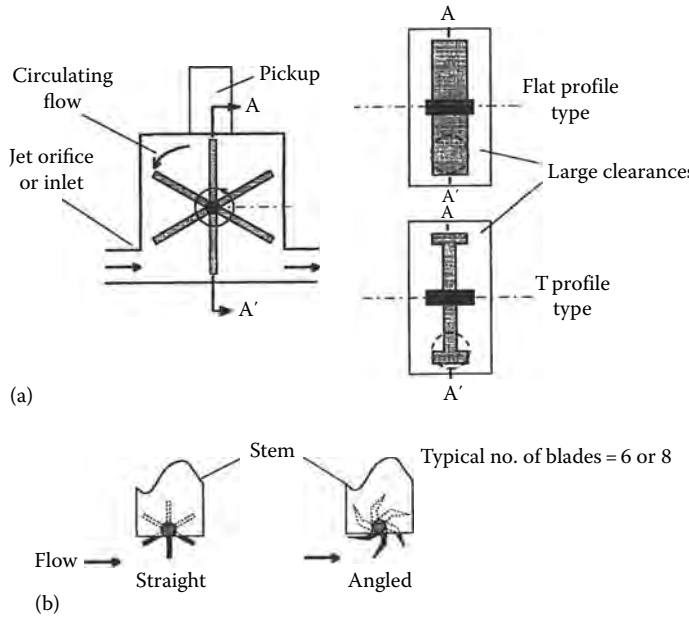


FIGURE 55.1 Impeller flowmeter rotor design variations: (a) in-line meters and (b) insertion meters. (Courtesy of David Wadlow.)



FIGURE 55.2 A typical insertion-type impeller flowmeter.

Some in-line meters have interchangeable orifice sizes, allowing the same body to be used over different flow ranges. The orificed in-line meters are typically insensitive to flow condition. Insertion meters are, of course, very sensitive to flow condition if volumetric measurements are inferred. This is an important distinction and can extend to differences between requirements for specific in-line versions. Installation is important. It affects the user when he or she is designing an installation, selecting a flowmeter, and deciding on how much confidence to place in the measurement reading.

Paddle wheel flowmeters are *never* used in liquid hydrocarbon custody transfer applications. The flow measurement capacity is insufficient to compete in monitoring large volume transfers, and accuracy is not sufficient for valuable fluids.

These meters go by a variety of names, depending on which name the manufacturer selects and also the application. One often-quoted, historical root is also the impulse turbine invented by Pelton in the nineteenth century and has given rise to the "Pelton wheel turbine" description currently used by some suppliers. A device like this was originally used to drive milling wheels directly about the vertical axis, rather than through a right angle gear as in the "undershot" wheel described. Another common name, besides paddle wheel and impeller type, based at least on suppliers' descriptions for the in-line variety, is single-jet tangential, or simply tangential, turbine. The former description distinguishes this design from that of multijet tangential turbine flowmeters and therefore deserves mention. There are thus a confusing variety of names in current use, all relating to essentially the same impeller-like, vaned flowmetering principle:

1. Impeller type: insertion and in-line
2. Paddle wheel type: insertion and in-line
3. Pelton turbine wheel type: in-line
4. Tangential turbine flowmeter: insertion and in-line
5. Single-jet tangential turbine flowmeter: in-line
6. Impulse turbine flowmeter: in-line

Historically, the impeller sensor is based on early electronic speedometers used in pleasure boating. Signet, followed quickly by Data Industrial, moved to modify the basic design of such marine instruments to meet the significantly more demanding service life requirements of industrial flow measurement. The resulting flowmeters, and their descendants, have been widely used since 1975. Significant engineering efforts have resulted in highly reliable instruments used in many extremely demanding applications.

55.2 Sensing Principles

All impeller flow sensors must detect the rotation of the impeller and in their usual form transmit a pulse train, at a frequency related to the rotational velocity of the impeller. Being essentially a digital output, impeller sensors can typically transmit signals over quite long distances, up to 1 km when so required. Detection principles used include

1. One or more magnets retained in the impeller or mechanically connected thereto, using the zero crossing of an induced AC field to generate the pulse train.
2. One or more coupling devices contained within the impeller, modulating a transmitted frequency that is processed to produce the pulse train.
3. One or more metallic targets installed within the sensor, sensed by any of the proximity pickup techniques commercially available, to produce output pulses.
4. One or more magnets retained in the impeller, used to switch a Hall effect device, producing the output pulse train.
5. Optical devices, both transmissive and reflective, have been used to sense the passage of the impeller blades to produce the output pulse train.
6. Measurement of the change in electric reactance due to the passage of impeller vanes through the measurement field area, conditioned to produce the output pulse train.

Any given supplier can produce several types of impeller flowmeter, each type using a different detection method depending on market requirements. Since the impeller can operate to rotational velocities of 4000 rpm or higher, output frequencies can be as high as 500 Hz. At low flow rates, the frequencies can be as low as 0.2 Hz. This factor should be addressed early in the selection procedure because the chosen output device must be capable of the frequency output range of the sensor.

Most sensor manufacturers can either supply or specify an appropriate meter to relate sensor output to flow rate in engineering units, either U.S. customary or SI. These meters are usually capable of displaying both flow rate and accumulated flow for the sensor to which they are connected. Additional outputs are available, either stand-alone or in combination with the meter, providing periodic pulse outputs at definable flow increments or analog outputs scaled to flow rate. In addition, certain control functions, alarms, and other special features required by the various markets served are often incorporated in these meters.

55.3 Flow Ranges

The insertion style of impeller meter, even when Tee installed, is a local sensing device, measuring the flow velocity in only a part of the flow stream. The manufacturer calibrates the meter for average flow across the entire cross section of the pipe. The paddle wheel location is usually close to the inner diameter of the pipe in a region of flow with a velocity significantly below the average flow velocity in the pipe. Proper calibration practices by the manufacturer make the meters effective at flow rates as low as 10 cm s^{-1} average velocity in spite of the low local velocities at the impeller. Generally speaking, flows at Reynolds numbers as low as 5000 can be run with no requirement for special calibration, and frequently lower numbers can be handled with no difficulty. The usual range specification is a flow rate equivalent to 0.3–10 m s^{-1} average flow velocity. The diversity of products available allows operation to velocities considerably lower than the 0.3 m s^{-1} range, indeed to as low as 0.07 m s^{-1} in certain specialty impeller flowmeters.

55.4 Installation and Applications

55.4.1 Pipe Sizes

Pipe sizes in which these sensors have been installed run the gamut from small-bore tubing to 2.3 m outside diameter. The larger pipe sizes are those that show the greatest installed cost savings over alternative metering systems. The impeller meter, in fact, can be cost-effective in any flowmeter application that is consistent with the accuracy of the instrument, particularly if the application involves pump or valve control.

55.4.2 Piping System Restrictions

Most suppliers require at least 10 pipe diameter lengths of straight pipe upstream of the installed meter and 5 downstream. These conditions are required to minimize the asymmetry of the flow stream in the neighborhood of the impeller, in the installed piping, which can be caused by elbows, tees, and valves. More is better; no supplier is likely to complain that there is too great a straight pipe length upstream. Less can adversely affect the calibration of the sensor due to the local variations of velocity resulting from flow disturbances.

55.4.3 Operating Pressure

Manufacturers' standard offerings are usually consistent with the pressure limitations of the materials of construction of the piping system with plastic piping systems and are commonly as high as 2.7 MPa with steel or brass piping. Higher pressures are available but are usually somewhat more expensive. Pressure drop generated by the installed flowmeter is usually low. Manufacturers can usually supply information on the anticipated pressure drop.

55.4.4 Calibration

Calibration of the sensors is usually specified by the manufacturer. Some manufacturers provide a calibration factor in terms of gallons per pulse or in pulses per gallon. Others, such as Data Industrial, provide data relating frequency to flow rate in GPM or other volumetric rate units. For insertion-style sensors, any such instrument must be field calibratable to accommodate the variation in the relation of impeller rotational velocity with average flow rate with pipe-size variation.

55.4.5 Accuracy

Accuracies for in-line impeller flowmeters vary considerably but can be high, ranging from ±0.2% reading for liquids for one manufacturer to several percent of full scale for several others. The difference between accuracies specified as % full scale and % reading should be particularly noted. Manufacturers use both, particularly with this class of meter. A ±1% full-scale device is often far less accurate than, for example, a ±2% reading device. For example, it is in error by 10% at the minimum of its flow range if the turndown is 10:1, whereas the ±2% reading device is still only in error by ±2%.

Output accuracy is usually specified as ±1% of full rated flow. This specification is broad enough to handle the dimensional and frictional variations in the sensors as manufactured. Some producers will custom calibrate the sensors to meet special needs, but this task should be performed only when the mating pipe entry and exit can be shipped to, and accommodated by, the manufacturer. Alternatively, the manufacturer can usually provide new calibration values if the meter used is identified, the meter reading is known, and the actual flow is known for at least two points on the flow curve. When calibration in place is required, the anticipated accuracy is in the range of ±0.5% of full scale or 1% of indicated flow, whichever is greater.

Repeatability of readings is usually on the order of 0.5%. Linearity, except at the extremes of the flow range, is also expected to be no worse than 0.5% and, over the full design range of the meter, is normally accurate within the ±1% of full rated flow.

To achieve the accuracies noted earlier, careful attention must be paid to proper installation, particularly with regard to insertion depth.

55.4.6 Design Features

These flow sensors are offered in materials of construction compatible with a broad range of aqueous solutions, of both high and low pH and with deionized water. For this latter service, the history of the sensor, particularly for particle generation, should be reviewed. Flow streams with high concentrations of solids, and particularly of fibrous solids, should be carefully reviewed when the impeller sensor is under consideration. Construction is reasonably forgiving with particulates, but caution in such applications is warranted. Heat transfer fluids are also compatible with sensors from at least two of the suppliers. Basic application limits should be discussed with the supplier, who should have a broad history of successful applications to back up recommendations, as well as sensor materials with test results specific to the measured fluid. The so-called "hot tap" or "wet tap" versions are available for installation in operating pipelines and in-line or on-line service. A submersible version capable of metering in flooded well pits for extended periods of time is available from some suppliers. At least two manufacturers provide "intrinsically safe" sensors for application in hazardous environments.

55.4.7 Areas of Application

Impeller flow sensors are widely used in the following fields with considerable success:

- Agricultural and horticultural irrigation
- Deionized water systems, including silicon wafer fabrication
- Heating, ventilating, and air conditioning and energy management
- Industrial waste treatment
- Industrial filtration systems
- Chemical reagent metering and batching
- Municipal water systems (potable water)

The sensors can be supplied with analog output for control applications, if desired.

Some suppliers have developed specialty impeller meters adapted to provide long life, ruggedness, and maintainability required in more demanding applications. The history of acceptable operation in applications similar to that proposed for the metering system should be reviewed with the supplier. Custodial transfer applications are normally unsuitable for the impeller sensor.

55.5 Manufacturers

The following is a selection of U.S. manufacturers. Their locations and phone numbers are listed in Table 55.1.

55.5.1 Data Industrial

In-line and insertion meters: 0.07–18 m s^{-1} (0.25–60 fps) fluid velocity. Available materials to accommodate most industrial fluids. Temperatures from –29 °C to 152 °C (–20 °F to 305 °F). Energy monitoring. Digital and/or analog meter outputs. NEMA 4, 4X, and 6P constructions. FM, CSA Approvals.

55.5.2 G. Fischer (Signet)

In-line and insertion meters: 0.1–7 m s^{-1} (0.3–20 fps) fluid velocity. Available materials to accommodate most industrial fluids. Temperatures to 149 °C (300 °F). Energy monitoring. Digital and/or analog meter outputs. NEMA 4, 4X constructions. FM, CE Approvals (http://www.gfsignet.com/go/45C24DDB1999 3E1D34D4906307B08581).

55.5.3 Seametrics

In-line meter. Sizes 3/8–2 in. Plastic, brass. Clean water applications. Flows from 0.05 to 40.0 GPM. Accuracy ±1% of full scale (FS). Insertion. Range of IP probes. Plastic, brass, stainless. Sapphire bearings and carbide shafts. Up to 250°F. 0.3–30 fps. Accuracy ±1% to ±2% FS (http://products.seametrics.com/).

55.5.4 Roper Flow Technology

In-line meters: Omni flow®, liquids and gases, highly precise, range 7.6–5677 mL min^{-1} (liquid), 0.0025–0.68 Am^3 h^{-1} (gases). Cryogenics to 593 °C, pressure to 60,000 psi, viscosity compensated.

Optiflo®, plastic construction for liquids only with some reduction in performance capabilities.

55.5.5 Hoffer

MF series (Miniflow). High-performance, in-line low flowmeters using a "Pelton rotor" for liquids and gases. Liquid flow ranges from 57 mL min^{-1} to 11.5 rpm. Linearity ±1% reading over 10:1 range.

TABLE 55.1 U.S. Manufacturers of Impeller Flowmeters

Manufacturer	City, State	Phone Number
Data Industrial Corp.	Mattapoisett, MA	(508) 758-6390
G. Fischer (Signet)	Tustin, CA	(714) 731-8800
Seametrics	Kent, WA	(206) 872-0284
Roper Flow Technology, Inc.	Phoenix, AZ	(602) 437-1315
Blancett	Altus, OK	(405) 482-0036
Hoffer Flow Controls, Inc.	Elizabeth City, NC	(800) 628-4584, (919) 331
McMillan Co.	Georgetown, TX	(512) 863-0231
Flowmetrics, Inc.	Canoga Park, CA	(818) 346-4492
Proteus Industries, Inc.	Mountain View, CA	(415) 964-4163

Repeatability ±0.1% typical. Gas flow ranges depending on density from 0.02 to 1.0 ACFM. Linearity ±1.5% reading. Repeatability ±0.2%. Various bearings available. Ball bearings used for high-accuracy applications. UVC curves available. "Smart" transmitter available for temperature–viscosity correction + linearization (http://hofferflow.com/).

55.5.6 McMillan

Low-viscosity liquids: In-line meters. Ranges from 13 mL min^{-1} to 10 L min^{-1}. Accuracies ±1%–±3% FS. IR rotation sensors. Have a Teflon version for corrosives.

Gases: Teflon version for chlorine, fluorine, etc. Accuracy ±3% FS. Ranges: 25–5 AmL min^{-1}.

55.5.7 Flowmetrics

Series 600: Stainless steel, in-line tangential turbine for gases and liquids. Ranges: 0.001–2.0 GPM liquids; 0.001–2.0 ACFM gases. Linearities vary according to range from ±1% to ±5% FS. Repeatabilities ± 0.05%, traceable NIST (http://www.flometrics.com/).

55.5.8 Proteus

In-line paddle wheel flowmeters and flow switches—liquids only. Typically used for water cooling lines on electrical or vacuum equipment. Quoted for use with liquids up to 120 cSt. Media include water, treated water, deionized water, ethylene glycol, and light oils. Range 0.08–60 GPM. Accuracies vary between models in the ±3% to ±4% FS range. Meter bodies can incorporate optional integral temperature and pressure transducers. These meters are also available with a FluidTek™ computer interface protocol and PC software for constructing embedded control systems for flow, temperature, and pressure (http://proteuseng.com.au/).

Bibliography

Bean, H. S. (ed.), *Fluid Meters*, 6th ed., New York: American Society of Mechanical Engineers, 1971, p. 99.

Cheremisinoff, N. P., *Applied Fluid Flow Measurement*, New York: Marcel Dekker, 1979, 106.

Impeller Flowmeter, Geophysics done right. http://www.enviroscan.com/html/impeller_flowmeter.html (accessed on July 25, 2012).

Products, Impeller meters, Veris Industries, http://www.veris.com/Category/Flow/Impeller-spcMeters.aspx (accessed on March 15, 2013).

Products, Positive displacement meters, FMC Technologies, http://www.fmctechnologies.com/en/MeasurementSolutions/Technologies/LiquidProducts/PDMeters.aspx (accessed on March 15, 2013).

Kjelstrom, L. C., Methods of measuring pumpage through closed conduit irrigation systems, *J. Irrig. Drain. Eng. ASCE*, 117, 748–757, 1991.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

56

Electromagnetic Flowmeters

56.1 Faraday's Law of Induction **56**-1
56.2 Construction and Operation of Electromagnetic Flowmeters **56**-4
56.3 Types of Electromagnetic Flowmeters **56**-5
AC-Type Electromagnetic Flowmeters • DC-Type Electromagnetic Flowmeters
56.4 Installations and Practical Applications **56**-7
Effects of Electric Conductivity of Fluid
56.5 Signal Pickup and Demodulation Techniques **56**-10
Bibliography **56**-11
Partial List of Manufacturers and Suppliers **56**-12

Halit Eren
Curtin University

Electromagnetic flowmeters have been widely used in industry for many decades. Unlike many other types of flowmeters, they offer true noninvasive measurements. They can easily be installed in the processes. These flowmeters can be used to monitor the reverse flows and are insensitive to viscosity, density, turbulences, and other flow disturbances. *Electromagnetic flowmeters* can rapidly respond to flow changes and exhibit linear characteristics for a wide measurement range. In recent years, technological refinements have resulted in the production and usage of much more economical, accurate, and smaller instruments than the previous versions.

As in the case of many electric devices, the underlying principle of the electromagnetic flowmeter is the Faraday's law of induction. The induced voltages in an electromagnetic flowmeter are linearly proportional to the mean velocity of liquids or to the volumetric flow rates. As is the case in many applications, if the pipe walls are made from nonconducting elements, then the induced voltage is largely independent of the properties of the fluid.

The accuracy of these meters can be as low as 0.25%, and in most cases, an accuracy of 1% is used. At worst, 5% accuracy is obtained in some difficult applications where impurities of liquids and the contact resistances of the electrodes are inferior as in the case of low-purity sodium liquid solutions.

56.1 Faraday's Law of Induction

Faraday's law of induction states that if a conductor of length l (m) is moving with a velocity v (m s^{-1}) perpendicular to a magnetic field of flux density B (Tesla), then the induced voltage e across the ends of conductor can be expressed by

$$e = Blv \tag{56.1}$$

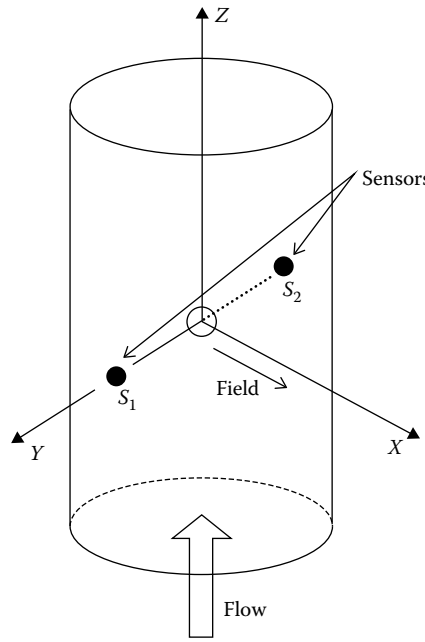


FIGURE 56.1 Operational principles of electromagnetic flowmeters. Faraday's law states that a voltage is induced in a conductor moving in a magnetic field. In electromagnetic flowmeters, the direction of movement of the conductor, the magnetic field, and the induced emf are perpendicular to each other on *x*-, *y*-, and *z*-axes. Sensors S_1 and S_2 experience a virtual conductor due to the flow of conductive fluid in the pipe.

The principle of application of Faraday's law to an electromagnetic flowmeter is given in Figure 56.1. The magnetic field, the direction of the movement of the conductor, and the induced electromotive force (emf) are all perpendicular to each other.

Figure 56.2 illustrates a simplified electromagnetic flowmeter in greater detail. By using externally located electromagnets, a homogeneous magnetic field that passes through the pipe and the liquid inside it is created. When a conducting flowing liquid inside the pipe cuts through the magnetic field, a voltage is induced along the liquid path between the two electrodes positioned on the opposite sides of the pipe.

In the case of electromagnetic flowmeters, the conductor is the liquid flowing through the pipe, and the length of the conductor is the distance between the two electrodes, which is equal to the tube diameter. The velocity of the conductor is proportional to the mean flow velocity of the liquid. Hence, the induced voltage becomes

$$e = BDv \tag{56.2}$$

where D (m) is the diameter of pipe.

If the magnetic field is constant and the diameter of the pipe is fixed, the magnitude of the induced voltage will only be dependent on the velocity of the liquid. If the ends of the conductor, in this case the sensors, are connected to an external circuit, the induced voltage causes a current, i, to flow, which can be processed suitably as a measure of the flow rate.

Electromagnetic flowmeters are often calibrated to determine the volumetric flow of the liquid. The volume of liquid flow, Q (L/s), can be related to the average fluid velocity as

$$Q = Av \tag{56.3}$$

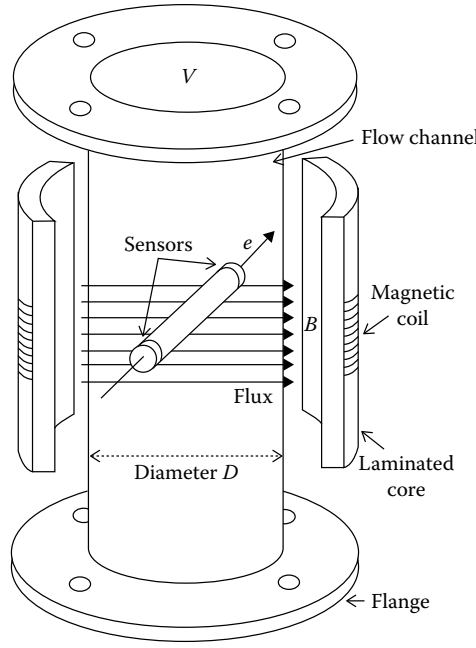


FIGURE 56.2 Construction of practical flowmeters. External electromagnets create a homogeneous magnetic field that passes through the pipe and the liquid inside. Sensors are located 90° to the magnetic field and to the direction of the fluid flow. Sensors are insulated from the pipe walls. Flanges are provided for fixing the flowmeter to external pipes. Usually, manufacturers supply information about the minimum lengths of the straight portions of external pipes.

Writing the area, A (m^2), of the pipe as

$$A = \frac{\pi D^2}{4} \tag{56.4}$$

gives the induced voltage as a function of the flow rate:

$$e = \frac{4BQ}{\pi D} \tag{56.5}$$

Equation 56.5 indicates that in a carefully designed flowmeter, if all other parameters are kept constant, then the induced voltage can be expressed to be linearly proportional to the volume of the liquid flowing through the pipe.

There are many different types of electromagnetic flowmeters available, such as AC, DC, and permanent magnets. The two most commonly used ones are the AC and DC types. This section concentrates mainly on AC- and DC-type flowmeters.

Although the induced voltage is directly proportional to the mean value of the liquid flow, the main difficulty in the use of electromagnetic flowmeters is that the amplitude of the induced voltage is small relative to extraneous voltages and noise. Noise sources include

- Stray voltage in the process liquid
- Capacitive coupling between signal and power circuits
- Capacitive coupling in connection leads
- Electromechanical emf induced in the electrodes and the process fluid
- Inductive coupling of the magnets within the flowmeter

56.2 Construction and Operation of Electromagnetic Flowmeters

Common to both AC and DC electromagnetic flowmeters, the magnetic coils create a magnetic field that passes through the flow tube and process fluid. As the conductive fluid flows through the flowmeter, a voltage is induced between the electrodes in contact with the process liquid. The electrodes are placed at positions where maximum potential differences occur. The electrodes are electrically isolated from the pipe walls by nonconductive liners to prevent short-circuiting of electrode signals. The liner also serves as protection to the flow tube to eliminate galvanic action and possible corrosion due to metal contacts. Electrodes are held in place by holders that also act as sealing.

The resistance of the moving conductor R can be expressed as $R = \rho D/a^2$ where ρ is the resistivity of the fluid and a is the effective area of the sensor. If surfaces of electrodes are not insulated, there will be an associated leakage current due to conductivity of the fluid.

Dimensionally, magnetic flowmeters are manufactured from 2 mm in diameter to 1.2 m. In a particular application, the determination of the size of the flowmeter is a matter of selecting the one that can handle the anticipated liquid velocities. The anticipated velocity of the liquid must be within the linear range of the device. As an example of a typical guide for selection, the capacity of various size flowmeters is given in Figure 56.3.

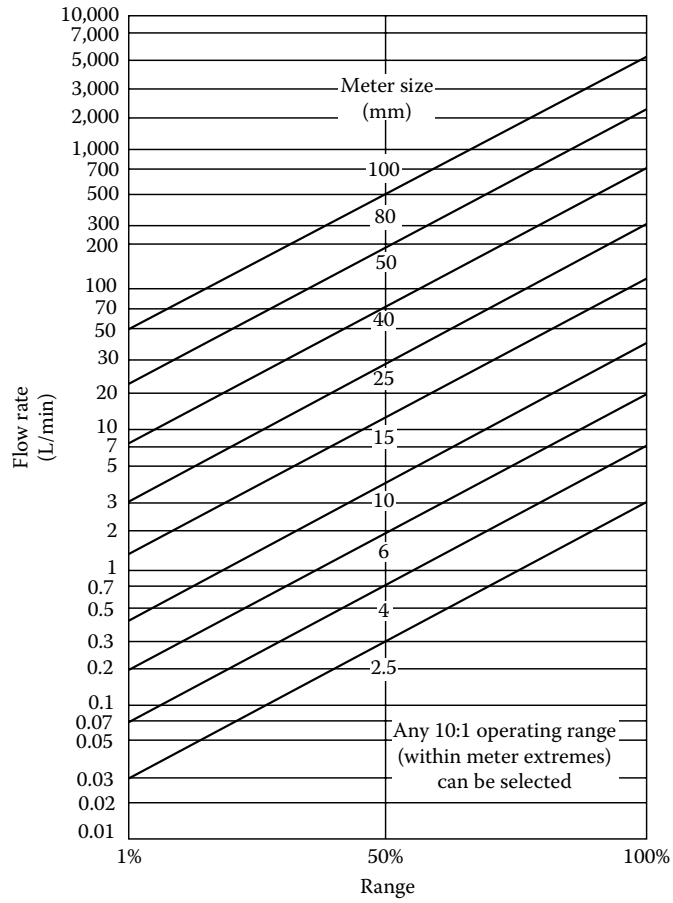


FIGURE 56.3 Selection of flowmeters. In the selection of a suitable flowmeter for a particular application, care must be exercised in handling the anticipated liquid velocities. The velocity of liquid must be within the linear range of the device. For example, a flowmeter with 100 mm internal diameter can handle flows between 50 and 4000 L min^{-1}. An optimum operation will be achieved typically at a flow rate of 500 L min^{-1}.

Some electromagnetic flowmeters are made from replaceable flow tubes, whereby the field coils are located external to the tubes. In these flowmeters, the flanges are located far apart from each other in order to reduce their adverse effects on the accuracy of the measurements; hence, they are relatively larger in dimensions. In other cases, the field coils are integrated with the flow tube itself. In these cases, the flanges are located closer to the magnets and the electrodes, thus yielding to relatively smaller dimensions. On the other hand, miniature and electrodeless magnetic flowmeters are also available, compact enough to allow them to be installed between two closely spaced flanges.

The wetted parts of a magnetic flowmeter include the liners, electrodes, and electrode holders. Many different materials such as rubber, Teflon, polyurethane, and polyethylene are used in the construction of these flowmeters to suit the process corrosion properties, the abrasiveness, and the temperature constraints. The main body and the electrodes are manufactured from stainless steel, tantalum, titanium, and various other alloys. Liners are selected mainly to withstand the abrasive and corrosive properties of the liquid. The electrodes must be selected such that they cannot be coated with the insulating deposits of the process liquid during long time periods of operations.

The pipe between the electromagnets of a flowmeter must be made from nonmagnetic materials to allow the field to penetrate the fluid without any distortion. Therefore, the flow tubes are usually constructed using stainless steel or plastics. The use of steel is a better option because it adds strength to the construction. Flanges are protected with appropriate liners and not allowed to make contact with the process fluid.

In some electromagnetic flowmeters, electrodes are cleaned continuously or periodically by ultrasonic or electric means. Ultrasonic cleaning methods are suitable for both AC- and DC-type magnetic flowmeters when frequent severe insulating coating is expected to be deposited on the electrodes, which might cause the flowmeter to cease to operate in an anticipated manner.

The operation of a magnetic flowmeter is generally limited by factors such as linear characteristics, pressure ratings of flanges, and temperatures of the process fluids. The maximum temperature limit is largely dependent on the liner material selection and usually is set to around 200°C. For example, ceramic liners can withstand high temperatures but are subject to cracking in the cases of sudden changes in temperatures of the process fluid.

During the selection of electromagnetic flowmeters, the velocity constraints should be evaluated carefully to secure accurate performance over the expected range. The full-scale (FS) velocity of the flowmeter is typically 0.3–10 m s^{-1}. Some flowmeters can measure lower velocities with somewhat poorer accuracy. Generally, employment of electromagnetic flowmeters over a velocity of 5 m s^{-1} should be considered carefully because erosion of the pipe and damages to liners can be significant.

The accuracy of a conventional magnetic flowmeter is usually expressed as a function of FS, typically 0.5%–1% FS. However, DC flowmeters have a well-defined zero due to automatic zeroing capabilities; therefore, they have a percentage rate of accuracy better than AC types, typically 0.5%–2%.

56.3 Types of Electromagnetic Flowmeters

56.3.1 AC-Type Electromagnetic Flowmeters

In many commercial electromagnetic flowmeters, an alternating current of 50–60 Hz in coils creates the magnetic field to excite the liquid flowing within the pipe. A voltage is induced in the liquid due to Faraday's law of induction, as explained earlier. A typical value of the induced emf in an AC flowmeter fixed on a 50 mm internal diameter pipe carrying 500 L min^{-1} is about 2.5 mV.

Historically, AC magnetic flowmeters were the most commonly used types because they reduced polarization effects at the electrodes. They are less affected by the flow profiles of the liquid inside the pipes. They allow the use of high Z_{in} amplifiers with low-drift and high-pass filters to eliminate slow and spurious voltage drifts emanating mainly from thermocouple and galvanic actions. These flowmeters

find many applications as diverse as the blood flow measurements in living specimens. Miniaturized sensors allow measurements on pipes and vessels as small as 2 mm in diameter. In these applications, the excitation frequencies are higher than industrial types, 200–1000 Hz.

A major disadvantage of the AC flowmeter is that the powerful AC field induces spurious AC signals in the measurement circuits. This necessitates periodic adjustment of zero output at zero velocity conditions—more frequently than for DC counterparts. Also, in some harsh industrial applications, currents in the magnetic field can vary due to voltage fluctuations and frequency variations in the power lines. The effect of fluctuations in the magnetic fields can be minimized by the use of a reference voltage proportional to the strength of the magnetic field to compensate for these variations. To avoid the effects of noise and fluctuations, special cabling and proper calibration practices recommended by the manufacturers must be followed to ensure accurate operation. Usually, the use of two conduits is required—one for the signal transmission and one for the power supply. The cable lengths should also be set to certain levels to minimize noise and sensitivity problems.

AC flowmeters operating at 50, 60, or 400 Hz are readily available. AC flowmeters can operate from 10 to 5000 Hz. High frequencies are preferred in determining the instantaneous behavior of transients and pulsating flows. Nevertheless, in applications where extremely good conducting fluids and liquid metals are used, the frequency must be kept low to avoid skin effects. On the other hand, if the fluid is a poor conductor, the frequency must not be so high such that dielectric relaxation is not instantaneous.

56.3.2 DC-Type Electromagnetic Flowmeters

Unlike AC magnetic flowmeters, direct current or pulsed magnetic flowmeters excite the flowing liquid with a field operating at 3–8 Hz. As the current to the magnet is turned on, a DC voltage is induced at the electrodes. The signals observed at the electrodes represent the sum of the induced voltage and the noise, as illustrated in Figure 56.4. When the current in the magnetic coils is turned off, the signal represents only the noise. The effect of noise can be eliminated by subtracting the signals of the flowmeter when no current flows through from the signals when the current flows through.

If the magnetic field coils are energized by normal direct current, then several problems can occur: polarization, which is the formation of a layer of gas around the measured electrodes, as well as electrochemical and electromechanical effects. Some of these problems can be overcome by energizing the field coils at higher frequencies or AC. However, higher frequencies and AC generate transformer action in the connectors and the fluid path. Therefore, the coils are excited by DC pulses at low repetition rates to eliminate the transformer action. In some flowmeters, by appropriate sampling and digital signal processing techniques, the zero errors and the noise can be rejected substantially.

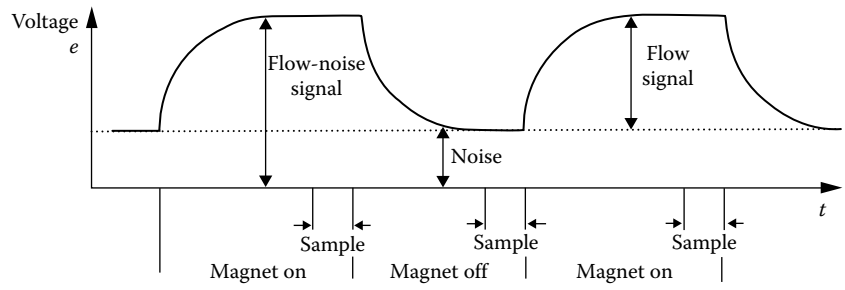


FIGURE 56.4 The signals observed at the electrodes represent the sum of the induced voltage and the noise. When the current in the magnetic coils is turned off, the signal across the electrodes represents only the noise. Subtracting the signal of the flowmeter when there is no current flow through the magnet from the measurement when current flows through the magnet, effectively cancels out the effect of noise.

The zero compensation inherent in DC magnetic flowmeters eliminates the necessity of zero adjustment. This allows the extraction of flow signals regardless of zero shifts due to spurious noise or electrode coating. Unlike AC flowmeters, a larger insulating electrode coating can be tolerated that could shift the effective conductivity significantly without affecting performance. If the effective conductivity remains high enough, a DC flowmeter will operate satisfactorily. Therefore, DC flowmeters are less susceptible to drifts, electrode coatings, and changes in process conditions in comparison to conventional AC flowmeters.

DC magnetic flowmeters do not have good response times due to the slow pulsed nature of operations. However, as long as there are not rapid variations in the flow patterns, zero to FS response times of a few seconds do not create problems in the majority of applications. Power requirements are also much less as the magnet is energized part of the time. This gives an advantage in power saving of up to 75%.

If the DC current to the magnet is constant, the proportional magnetic field can be kept steady. Therefore, the amplitudes of the DC voltages generated at the electrodes will be linearly proportional to the flow. However, in practice, the current to the magnet varies slightly due to line voltage and frequency variations. As in the case of AC flowmeters, voltage and frequency variations could necessitate the use of a reference voltage. Because the effect of noise can be eliminated better, the cabling requirements are not as stringent.

To avoid electrolytic polarization of the electrodes, bipolar pulsed DC flowmeters have been designed. Also, modification of DC flowmeters led to the development of miniature DC magnetic flowmeters with wafer design for a limited range of applications. The wafer design reduces the weights as well as the power requirements.

56.4 Installations and Practical Applications

Conventional AC and DC magnetic flowmeters have flanges at the inlet and the outlet that need to be bolted to the flanges of the pipe. Face-to-face dimensions of magnetic flowmeters differ between manufacturers; therefore, once the flowmeter is installed, new piping arrangements could be necessary if a flowmeter is replaced by one from a different manufacturer.

Upstream and downstream straight piping requirements can vary from one flowmeter to another, depending on the manufacturer's specifications. As a rule of thumb, the straight portion of the pipe should be at least 5D/2D from the electrodes and 5D/5D from the face of the flowmeter in upstream and downstream directions, respectively. For good accuracy, one should adhere carefully to the recommendations of manufacturers for piping requirements. In some magnetic flowmeters, coils are used in such a way that the magnetic field is distributed in the coil to minimize the piping effect.

For accurate measurements, magnetic flowmeters must be kept full of liquid at all times. If the liquid does not contact the electrodes, measurements cannot be taken. Figure 56.5 illustrates this point. If the measurements are made in other than vertical flows, the electrodes should be located in horizontal directions to eliminate the possible adverse effect of the air bubbles, because the air bubbles tend to concentrate on the top vertical part of the liquid.

In the selection of magnetic flowmeters, a number of considerations must be taken into account, as

- Cost, simplicity, precision, and reproducibility
- Metallurgical aspects
- Velocity profiles and upstream disturbances

Most processes employ circular piping that adds simplicity to the construction of the system. The flowmeters connected to circular pipes give relatively better results compared to rectangular- or square-shaped pipes, and velocity profiles of the liquid are not affected by the asymmetry. However, in circular pipes, the fringing of the magnetic field can be significant, making it necessary to employ empirical calibrations.

Selection of materials for constructing the channel of the magnetic flowmeter demands care. If the fluid is nonmetallic, a nonconducting or an insulated channel should be sufficient. In this case, wetted

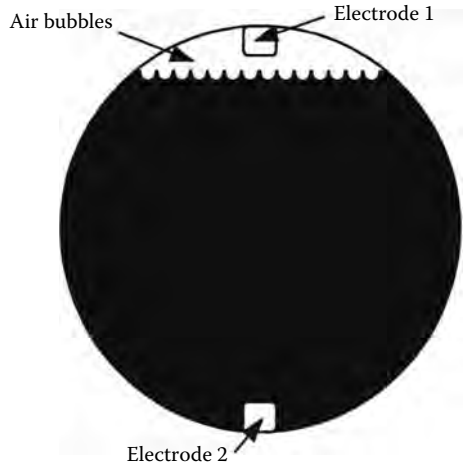


FIGURE 56.5 The pipes of electromagnetic flowmeters must be full of liquid at all times for accurate measurement. If the liquid does not make full contact with electrodes, the high impedance prevents the current flow; hence, measurements cannot be taken. Also, if the pipe is not full, even if the contact is maintained between the liquid and electrodes, the empty portions of the pipe will lead to miscalculated flow rates.

electrodes must be used. Electrodes must be designed to have sufficiently large dimensions to keep the output impedance at acceptable levels. Also, careful handling of electrode signals must be observed because, in many cases, malfunctioning of reference signal electronics is the main cause of flowmeter failure.

Magnetic flowmeters do not require continuous maintenance, except for periodic calibrations. Nevertheless, electrode coating, damage to the liners, and electronic failures can occur. Any modification or repair must be treated carefully because, when installed again, some accuracy can be lost. After each modification or repair, recalibration will be necessary.

Often, magnetic flowmeter liners are damaged by the presence of debris and solids in the process liquid. Also, the use of incompatible liquid with the liners, wear due to abrasion, excess temperature, and installation and removals can contribute to the damage of liners. The corrosion in the electrodes can also be a contributing factor for the damage. In some cases, magnetic flowmeters can be repaired onsite even if severe damage occurs; in other cases, they must be shipped to the manufacturer for repairs. Usually, manufacturers supply spare parts for electrodes, liners, flow tubes, and electronic components.

Calibration of electromagnetic flowmeters is achieved with a magnetic flowmeter calibrator or by electronic means. The magnetic flowmeter calibrators are precision instruments that inject simulated output signals of the primary flowmeter into the transmitter. Effectively, this signal is used to check correct operation of electronic components and make adjustments to the electronic circuits. Alternatively, calibrations can also be made by injecting suitable test signals to discrete electronic components. In some cases, empirical calibrations must be performed at zero flow while the flowmeter is filled with the stationary process liquid.

Application of magnetic flowmeters can only be realized with conductive liquids such as acids, bases, slurries, foods, dyes, polymers, emulsions, and suitable mixtures that have conductivities greater than the minimum conductivity requirements. Generally, magnetic flowmeters are not suitable for liquids containing organic materials and hydrocarbons. As a rule of thumb, magnetic flowmeters can be applied if the process liquids constitute a minimum of about 10% conductive liquid in the mixture.

The lack of any direct Reynolds number constraints and the obstructionless design of magnetic flowmeters make it practical for applications that involve conductive liquids that have high viscosity, which could plug other flowmeters. They also measure bidirectional flows.

Despite the contrary belief, magnetic flowmeters demonstrate a certain degree of sensitivity to flow profiles. Another important aspect is the negative effect of the turbulence. Unfortunately, there is very

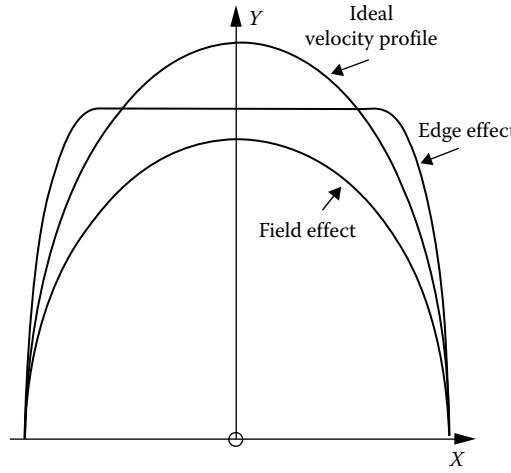


FIGURE 56.6 Flow profiles in the pipes. Magnetic flowmeters demonstrate a certain degree of sensitivity to flow profiles. The ideal velocity profile can be distorted due to edge effects and also field effects known as magnetohydrodynamics. In some applications, the velocity perturbation due to the magnetohydrodynamic effect can be serious enough to severely influence the accuracy of operations.

little information available on the behavior of turbulent flows when they are in transverse magnetic fields. Figure 56.6 shows an example of flow profile in which the velocity profile is perturbed. The fluid is being retarded near the center of the channel and accelerated at the top and bottom near the electrodes.

An important point in electromagnetic flowmeters is the effect of magnetohydrodynamics, especially prominent in fluids with magnetic properties. Hydrodynamics is the ability of a magnetic field to modify the flow pattern. In some applications, the velocity perturbation due to the magnetohydrodynamic effect can be serious enough to influence the accuracy of operations, as in the case of liquid sodium and its solutions.

56.4.1 Effects of Electric Conductivity of Fluid

For electromagnetic flowmeters to operate accurately, the process liquid must have minimum conductivity of 1–5 μS cm^{-1}. Most common applications involve liquids with conductivities greater than 5 μS cm^{-1}. Nevertheless, for accurate operation, the requirement for the minimum conductivity of liquid can be affected by length of leads from sensors to the electronic signal processors.

AC flowmeters are sensitive to the nonconductive coating of the electrodes, which may result in calibration shift or complete loss of signals. The DC flowmeters, on the other hand, should not be affected by nonconductive coating to such a great extent, unless the conductivity between electrodes is less than the minimum required value. In some flowmeters, electrodes can be replaced easily, while in others, electrodes can be cleaned by suitable methods. If coating is a continual problem, ultrasonic and other cleaning methods should be considered.

Zero adjustment of AC magnetic flowmeters requires compensation for noise. If the zero adjustment is performed with any fluid other than the process fluid, serious errors can result because of possible differences in conductivities. Similarly, if the electrodes are coated with an insulating substance, the effective conductivity of the electrodes can be altered, thereby causing a calibration shift. If the coating changes in time, the flowmeter can continually require calibration for repeatable readings.

The resistance between electrodes can be expressed by $R = D/\delta\pi r^2$, where D is the pipe diameter, δ is the fluid conductivity, and r is the electrode radius. For tap water, δ = 200 μS cm^{-1}; for gasoline, δ = 0.01 μS cm^{-1}; and for alcohol, δ = 0.2 μS cm^{-1}. A typical electrode with a 0.74 cm diameter in contact with tap water results in a resistance of 6756 Ω.

56.5 Signal Pickup and Demodulation Techniques

Electromagnetic flowmeters are four-wire devices that require an external power source for operations. Particularly in AC magnetic flowmeters, the high-voltage power cables and low-voltage signal cables must run separately, preferably in different conduits, whereas in DC magnetic flowmeters, the power and signal cables can be in one conduit. This is because in DC-type magnetic flowmeters, the voltage and the frequency of excitation of the electromagnets are relatively much lower. Some manufacturers supply special cables along with their flowmeters.

In AC flowmeters, the electrode signals can be amplified much more readily compared to their DC counterparts. That is the reason why AC flowmeters have been used successfully to measure very low flow rates, as well as the flow of very weakly conducting fluids. Nevertheless, AC flowmeters tend to be more complicated, bulky, and expensive, and they require electromagnets with laminated yokes together with stabilized power supplies. In some magnetic flowmeters, it is feasible to obtain sufficiently large flow signal outputs without the use of a yoke by means of producing a magnetic field by naked coils. In this case, the transformer action to the connecting leads can be reduced considerably.

One of the main drawbacks of AC-type flowmeters is that it is difficult to eliminate the signals due to transformer action from the useful signals. The separation of the two signals is achieved by exploiting the fact that the flow-dependent signal and the transformer signal are in quadrature. That is, the useful signal is proportional to the field strength, and the transformer action is proportional to the time derivative of the field strength. The total voltage v_T can be expressed as

$$v_T = v_F + v_t = V_F \sin(\omega t) + V_t \cos(\omega t) \tag{56.6}$$

where

$v_F = e$ is the induced voltage due to liquid flow

v_t is the voltage due to transformer action on wires

Phase-sensitive demodulation techniques can be employed to eliminate the transformer action voltage. The coil magnetizing current, $i_m = I_m \sin(\omega t)$, is sensed and multiplied by the total voltage v_T, giving

$$v_T i_m = [V_F \sin(\omega t) + V_t \cos(\omega t)]\, I_m \sin(\omega t) \tag{56.7}$$

Integration of Equation 56.7 over one period between 0 and 2π eliminates the transformer voltage, yielding only the voltage that is proportional to the flow:

$$V_f = V_F I_m \pi \tag{56.8}$$

where V_f is the voltage after integration like its DC value once the sinusoidal part is eliminated. This voltage is proportional to the induced voltage modified by constants I_m and π.

In reality, this situation can be much more complicated because of phase shift due to eddy currents in nearby solids and conductors. Other reasons for complexity include the harmonics because of nonlinearity such as hysteresis and capacitive pickup.

A good electrical grounding of magnetic flowmeters, as illustrated in Figure 56.7, is required to isolate relatively high common-mode potential. The sources of ground potential can be in the liquid or in the pipes. In practice, if the pipe is conductive and makes contact with the liquid, the flowmeter should be grounded to the pipe. If the pipe is made from nonconductive materials, the ground rings should be installed to maintain contact with the process liquid.

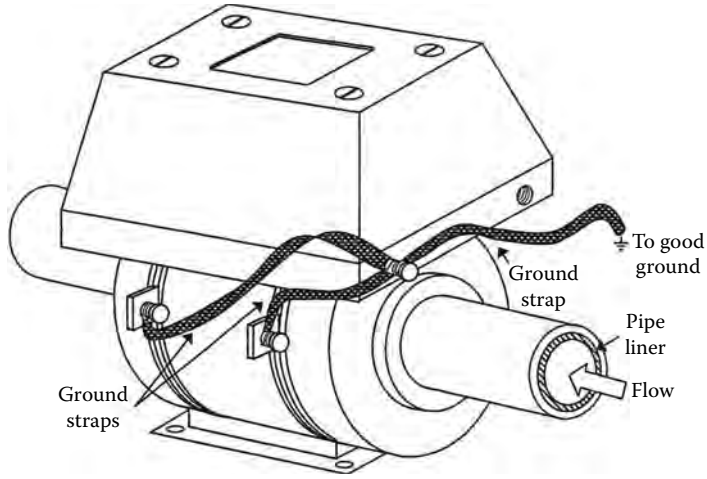


FIGURE 56.7 Grounding of electromagnetic flowmeters. A good grounding is absolutely essential to isolate noise and high common-mode potential. If the pipe is conductive and makes contact with the liquid, the flowmeter should be grounded to the pipe. If the pipe is made from nonconductive materials, ground rings should be installed to maintain contact with the process liquid. Improper grounding results in excessive common-mode voltages that can severely limit the accuracy and also damage the processing electronics.

If the flowmeter is not grounded carefully relative to the potential of the fluid in the pipe, then the flowmeter electrodes could be exposed to excessive common-mode voltages that can severely limit the accuracy. Excessive ground potential can damage the electronic circuits because the least-resistance path to the ground for any stray voltage in the liquid would be via the electrodes.

Some commercial magnetic flowmeters that can operate on sawtooth or square waveforms have been developed. Universally standardized magnetic flowmeters and generalized calibration procedures still do not exist, and manufacturers use their own particular design of flow channels, electromagnets, coils, and signal processors. Most manufacturers provide their own calibration data.

Bibliography

Baker, R. C., *Flow Measurement Handbook*, Cambridge, U.K.: Cambridge University Press, 2000.

Bates, C. J. and Frankilin, B., The performance characteristics of a novel multi-electrode electromagnetic flowmeter, *Measurement*, London, U.K.: Elsevier, Vol. 35, No. 4, pp. 399–408, 2004.

Bates, C. J. and Turner, R. B., Fluid flow studies associated with a new electromagnetic flowmeter, *Measurement*, London, U.K.: Elsevier, Vol. 33, No. 1, pp. 85–94, 2003.

Committee report: Magnetic inductive flowmeters, *Journal of American Water Works Association*, 99(6), 154–160, 2007.

Consumer Guide to Magnetic Flowmeters, Flowmeter Directory. http://www.flowmeterdirectory.com/flowmeter_artc_03112701.html (accessed on September 7, 2011).

Digiacomo, R. W., Understanding electromagnetic flow meters, *Chemical Engineering Progress*, Los Angeles, CA: American Institute of Chemical Engineers, Vol. 106, No. 5, pp. 42–47, 2010.

Doebelin, E. O., *Measurement Systems: Application and Design*, 4th edn., New York: McGraw-Hill, 1990.

Magnetic Flowmeters. Measuring Principle, Rockwell Automation. http://www.instrumentacionycontrol.net/ (accessed on September 2, 2011).

Magnetic Flowmeters, Instrumart. http://www.instrumart.com/Content.aspx?ContentID=223 (accessed on September 5, 2011).

Miller, R. W., *Flow Measurement Engineering Handbook*, Boston, MA: McGraw-Hill, 1996.

Omega Engineering, Introduction to Magnetic Flowmeters. http://www.omega.com/prodinfo/magmeter.html (accessed on September 5, 2011).

Sanderson, M. L., Factors affecting the dynamic performance of electromagnetic flow meters, *Measurement and Control*, London, U.K.: Institute of Measurement Control, Vol. 36, No. 9, pp. 270–274, 2003.

Partial List of Manufacturers and Suppliers

ABB, Inc., 125 E. County Line Rd. Warminster, PA 18974, Phone: 215-674-6580, 800-829-6001 (toll free), Fax: 215-674-7180, www.abb.us/instrumentation

Accu-Chek, Inc., 8385 State Rd. 64 Georgetown, IN 47122, Phone: 877-891-8830 (toll free), Fax: 812-951-2262, aciquality.rtrk.com/?scid=2091472&rl_alt…

American Controls, Inc., 20764 Whitlock Farmington Hills, MI 48336-5168, Phone: 248-476-7782, 800-470-3441 (toll free), Fax: 248-476-3390, www.americancontrolsinc.com

AMETEK TIP (Technical & Industrial Products), 627 Lake St. Kent, OH 44240, Phone: 215-256-6601, Fax: 330-677-3306, www.ametektip.com

Azbil North America, Inc., 9033 N. 24th Ave., Suite 5 Phoenix, AZ 85021, Phone: 888-262-4639, Fax: 602-216-8213, us.azbil.com

Controls and Meters Inc., 7613 Washington Ave. S. Minneapolis, MN 55439, Phone: 952-944-3666, www.controlsandmeters.com

Davis/Inotek, Bustleton Business Center, 2200 Michener St., Suite 23 Philadelphia, PA 19115, Phone: 800-548-9409, Fax: 410-358-0252, www.davis.com

Dwyer Instruments, Inc., Hwy. 212 at 12 P.O. Box 373 Michigan City, IN 46361, Phone: 219-879-8000, 800-872-9141 (toll free), Fax: 219-872-9057, www.dwyer-inst.com

Eastern Controls, Inc., 3866 Providence Rd. Edgemont, PA 19028, Phone: 610-325-4600, Fax: 610-325-2476, www.easterncontrols.com

Emerson Process Management, Rosemount Div., 8200 Market Blvd. Chanhassen, MN 55317, Phone: 952-906-8888, 800-685-8254 (toll free), Fax: 952-906-8889, www2.emersonprocess.com/en-us/brands/ros…

Endress+Hauser, 2350 Endress Place Greenwood, IN 46143, Phone: 888-363-7377, 888-ENDRESS (toll free), Fax: 317-535-8498, www.us.endress.com

Enpro, Inc., 121 S. Lombard Rd. Addison, IL 60101, Phone: 630-629-3504, 800-323-2416 (toll free), Fax: 630-629-3512, www.enproinc.com

Entegris, Inc., 300 Old Greenwood Rd. Decatur, TX 76234, Phone: 940-627-2121, 800-433-5547 (toll free), www.entegris.com

FCX Performance, 3000 E. 14th Ave. Columbus, OH 43219, Phone: 614-253-1996, 800-253-6223 (toll free), Fax: 614-253-2033, www.fcxperformance.com/

FMC Technologies Inc., Measurement Solutions, 1602 Wagner Ave. P.O. Box 10428 Erie, PA 16514, Phone: 814-898-5000, 800-477-6375 (toll free), Fax: 814-899-8927, www.fmctechnologies.com/en/MeasurementSo…

FTI Flow Technology Inc., 8930 S. Beck Ave. Suite 107 Tempe, AZ 85284, Phone: 480-240-3400, Fax: 480-240-3401, www.ftimeters.com

G Instruments, 13240 Evening Creek Dr. Suite 304 San Diego, CA 92128, Phone: 858-668-3088, Fax: 858-876-1963, www.ginstruments.com

Honeywell Sensing and Control, 1985 Douglas Dr. North, MN10-192B Golden Valley, MN 55422, Phone: 763-954-4818, 800-446-6555 (toll free), sensing.honeywell.com/index.cfm/ci_id/15…

Industrial Dynamics, Inc., 405 Gordon Ave. Exton, PA 19341, Phone: 610-363-8100, Fax: 610-363-8101, www.industrialdynamics.com

Intek, Inc., 751 Intek Way Westerville, OH 43082-9057, Phone: 614-895-0301, 888-569-3569 (toll free), Fax: 614-895-0319, www.intekflow.com

Keco Engineered Controls, 1200 River Ave. Bldg. 3A Lakewood, NJ 08701-5657, Phone: 732-901-5900, Fax: 732-901-5904, www.kecocontrols.com

KROHNE, Inc., 7 Dearborn Rd. Peabody, MA 01960, Phone: 978-535-6060, 800-356-9464 (toll free), Fax: 978-535-1720, www.krohne.com/index.php?id=12076

Liquid Measurement & Controls, Inc., 43 W. Third Ave., P.O. Box 26130 Collegeville, PA 19426, Phone: 610-489-0900, Fax: 610-489-1722, www.lmnc.net

Matheson, 166 Keystone Dr. Montgomeryville, PA 18936, Phone: 215-641-2700, 800-416-2505 (toll free), 877-MTG4GAS (toll free), Fax: 215-619-0458, www.mathesongas.com/index.aspx

MSC Industrial Supply Co., 75 Maxess Rd. Melville, NY 11747, Phone: 516-812-2000, 800-753-7937 (toll free), Fax: 516-812-1710, 800-255-5067 (toll free), www.mscdirect.com

MSR Magmeter, P.O. Box 32170 Edmonton, AB T6K4C2, Phone: 780-662-2171, Fax: 780-662-2146, www.mscdirect.com

Omron, STI Safety-Scientific Technologies, Inc., 6550 Dumbarton Circle Fremont, CA 94555, Phone: 510-608-3400, 888-510-4357 (toll free), Fax: 510-744-1442, www.sti.com

RCM Industries, Inc., 110 Mason Circle Suite D Concord, CA 94520-1214, Phone: 925-687-8363, 800-276-7915 (toll free), Fax: 925-671-9636, www.flo-gage.com

Siemens Corp., 527 Madison Ave. 8th Fl., New York, NY 10022-4611, Phone: 800-SIEMENS (toll free), Fax: 212-767-0580, www.usa.siemens.com/entry/en/

Sierra Instruments, 5 Harris Ct., Bldg. L, Monterey, CA 93940, Phone: 831-373-0200, 800-866-0200 (toll free), Fax: 831-373-4402, www.sierrainstruments.com

Steven Engineering, 230 Ryan Way South San Francisco, CA 94080, Phone: 866-974-0170 (toll free), Fax: 800-314-0716, 800-314-0716 (toll free), stevenengineering.rtrk.com/?scid=1975164…

Taco, Inc., 1160 Cranston St. Cranston, RI 02920, Phone: 401-942-8000, Fax: 401-942-2360, www.taco-hvac.com

Toshiba International Corp., 13131 W. Little York Rd. Houston, TX 77041, Phone: 713-466-0277, Fax: 713-466-8773, www.toshiba.com/ind/

V-F Controls, Inc., 8619 Tyler Blvd. Mentor, OH 44060-1008, Phone: 440-974-7757, 800-752-2033 (toll free), Fax: 440-974-6766, www.v-fcontrols.com

Zepeda Instruments, 615-T Aloha St. Seattle, WA 98109, Phone: 206-284-8519.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

57

Ultrasonic Flowmeters

Hans-Peter Vaterlaus
Rittmeyer Ltd.

Thomas Hossle
Rittmeyer Ltd.

Paolo Giordano
Rittmeyer Ltd.

Christophe Bruttin
Rittmeyer Ltd.

57.1 Introduction **57**-1
57.2 Transit Time Flowmeter **57**-2
Principle of Operation • Sensors • Axial • Radial • Field Mounting • Clamp on • Open Channels and Special Applications • Measurement of Flow in Closed Conduits • Single Path with Circular and Rectangular Cross Sections • Multipath Integration in Circular Cross Sections
57.3 Measurement of Flow in Open Channels **57**-9
Single Path • Multipath
57.4 Application: Penstock Leak Detection with Surge Chambers..... **57**-13
Instrumentation and Manufacturers/Distributors
References..... **57**-15
Further Readings **57**-15

57.1 Introduction

Flow is one of the most important physical parameters measured in industry and water management. There are various kinds of flowmeters available, depending on the requirements defined by the different market segments. For many years, differential pressure types of flowmeters have been the most widely applied flow measuring device for fluid flows in pipes and open channels that require accurate measurement at reasonable cost. In markets like waterpower, water supply, and irrigation, however, flow must be measured without any head losses or any pressure drop. This means there are no moving parts, no secondary devices, nor any restrictions allowed. Two types of flowmeters presently fulfill this requirement: electromagnetic and ultrasonic flowmeters. Whereas *ultrasonic flowmeters* can be applied in nearly any kind of flowing liquid, *electromagnetic* flowmeters require a minimum electric conductivity of the liquid for operation. In addition, the cost of ultrasonic flowmeters is nearly independent of pipe diameter, whereas the price of electromagnetic flowmeters increases drastically with pipe diameter.

There are various types of ultrasonic flowmeters in use for discharge measurement: (1) *Transit time*: This is today's state-of-the-art technology and most widely used type and will be discussed in this chapter section. This type of ultrasonic flowmeter makes use of the difference in the time for a sonic pulse to travel a fixed distance, first against the flow and then in the direction of flow. Transmit time flowmeters are sensitive to suspended solids or air bubbles in the fluid. (2) *Doppler*: This type is more popular and less expensive, but is not considered as accurate as the transit time flowmeter. It makes use of the Doppler frequency shift caused by sound reflected or scattered from suspensions in the flow path and is therefore more complementary than competitive to transit time flowmeters. (3) *Cross correlation*: Two measuring sections are installed with a certain distance to each other. Both measure the energy absorption of the ultrasonic signal. A cross correlation calculates the flow velocity. (4) *Phase shift*: The phase position of the transmitting and receiving signal is measured in the direction of the flow and against it. The resulting phase shift angle is directly proportional to the flow velocity. (5) *Drift*: The drift of an ultrasonic signal crossing the flow is measured by signal attenuation.

57.2 Transit Time Flowmeter

57.2.1 Principle of Operation

The acoustic method of discharge measurement is based on the fact that the propagation velocity of an acoustic wave and the flow velocity are summed vectorially. This type of flowmeter measures the difference in transit times between two ultrasonic pulses transmitted upstream t_{21} and downstream t_{12} across the flow, as shown in Figure 57.1. If there are no transverse flow components in the conduit, these two transmit times of acoustic pulses are given by

$$t_{12} = \frac{L_w}{c + v_a \cos\phi} \quad \text{and} \quad t_{21} = \frac{L_w}{c - v_a \cos\phi} \tag{57.1}$$

where

L_w is the distance in the fluid between the two transducers
c is the speed of sound at the operating conditions
ϕ is the angle between the axis of the conduit and the acoustic path
$\overline{v}_a$ is the axial low velocity averaged along the distance L_w

Since the transducers are generally used both as transmitters and receivers, the difference in travel time can be determined with the same pair of transducers. Thus, the mean axial velocity $\overline{v}_a$ along the path is given by

$$\overline{v}_a = \frac{L_w}{2\cos\phi}\left(\frac{1}{t_{21}} - \frac{1}{t_{12}}\right) = \frac{D}{2\cos\phi\sin\phi}\left(\frac{1}{t_{21}} - \frac{1}{t_{12}}\right) \tag{57.2}$$

The following example shows the demands on the time measurement technique: assuming a closed conduit with diameter D = 150 mm, angle ϕ = 60°, flow velocity $\overline{v}_a$ = 1 m·s^{-1}, and water temperature = 20 °C. This results in transmit times of about 116 µs and a time difference $\Delta t(\Delta t = t_{12} - t_{21})$ on the order of 78 ns. To achieve an accuracy of 1% of the corresponding full-scale range, Δt has to be measured with a resolution of at least 100 ps (1×10^{-10} s).

Standard time measurement techniques are not able to meet such requirements so that special techniques must be applied. The advantage of the approach of state-of-the-art real digital measurement is to process the measured value directly by a microcomputer. The most difficult problem is to reach the required resolution and to cope with the jitter of digital logic gates. It is well known in the measurement technique that, if signals are sampled multiple times (N_n), the resolution increases with the number of samples ("oversampling") [1]. This knowledge is not only applied for analog signals but also for transit time signals and is used in today's technology of flowmeters. The propagation time $t(t_{12} + t_{21})$, depending

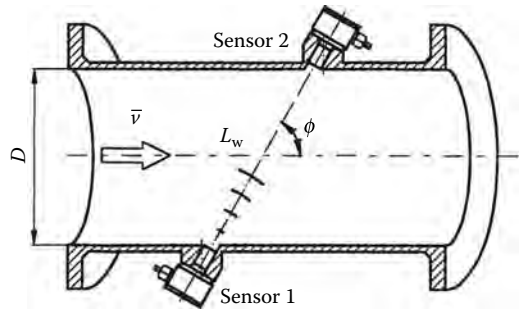


FIGURE 57.1 Principle of transit time flowmeters. Transmitting an ultrasonic pulse upstream and downstream across the flow: the liquid is moving with velocity $\overline{v}_a$ and with angle ϕ to the ultrasonic pulse.

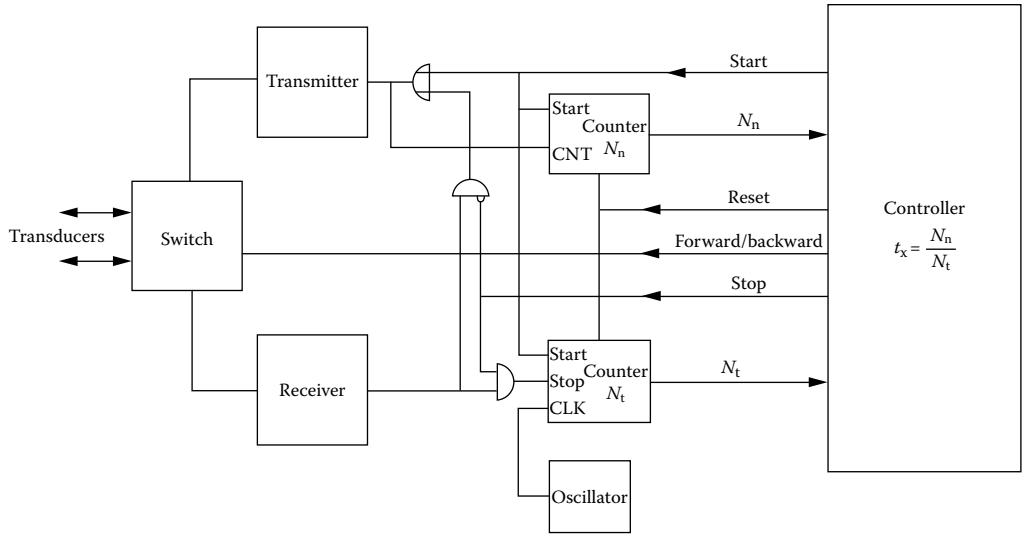


FIGURE 57.2 Block diagram of a transit time ultrasonic flowmeter using oversampling for higher resolution.

on the distance the sound pulse has to travel through the fluid, is measured several times. Due to this fact, one obtains the following relation [2]:

$$t = \frac{1}{N_n} \int_0^{\tau} \frac{1}{f} dt \tag{57.3}$$

where

τ is the integration time
f is the frequency

According to the block diagram in Figure 57.2, two counters are used. One counter, N_t, is clocked during the measuring period by a stable quartz oscillator; the other one counts the number of samples N_n. The measurement stops after a certain period of some milliseconds and after having reached an integer value for N_n. The two counts of N_t and N_n are used to calculate propagation time t_{12} or t_{21} by dividing N_t by N_n. The resolution r is calculated by

$$r = \frac{1}{N_n \cdot f} \tag{57.4}$$

Advantages of this measurement method for transit time flowmeters are

1. The resolution of the velocity measurement is constant (typical value 0.8 mm s^{-1}).
2. The accuracy depends almost only on the stability and the temperature coefficient of the quartz oscillator.
3. Due to the multiple sampling, the jitter of the digital logic is averaged.

57.2.2 Sensors

The transducer comprises the piezoelectric element that converts electric energy to acoustic energy and the basic structure for supporting the piezoceramic and providing electric connections. Transducer design entails choice of the piezoelectric element, determination of suitable dimensions and resonant

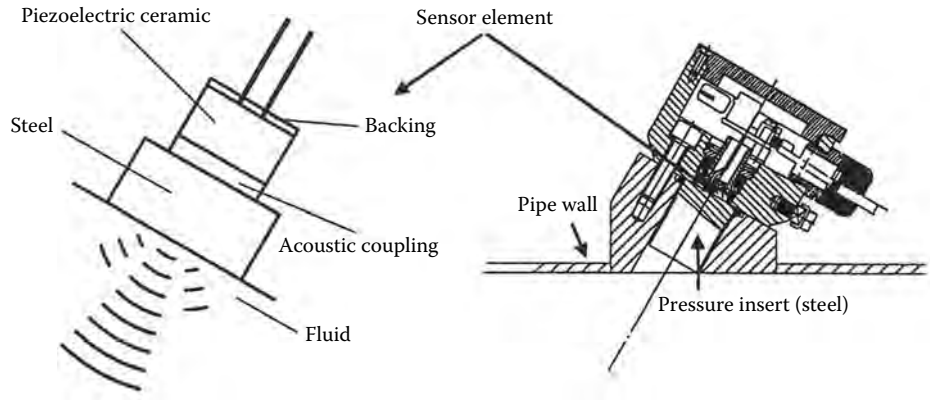


FIGURE 57.3 Ultrasonic transducer in principle and as an example of an existing version. The sensor element can be changed even under pressure.

frequency, and construction to withstand thermal and mechanical stress (see Figure 57.3). Considering the transmission line model [3,4], the optimal electroacoustic response, and the best matching of acoustic impedances $Z = \rho c$ of the different transducer elements is important. Not only must suitable waveforms to be detected by the electronics be obtained but also the energy loss must be minimized when crossing several interface boundaries. Because of the wide versatility of pipe sizes and flow conditions, there are a number of different sensor configurations for transit time flowmeters.

57.2.3 Axial

Due to the small differences in small-diameter pipes, it is necessary to pass the sonic pulse directly down the axis of the pipe to ensure that there is sufficient path length. Figure 57.4 shows an example of an axial sensor pipe section. The upper pipe size limit for this kind of sensor is about 0.075 m [5].

57.2.4 Radial

Many manufacturers supply complete metering sections with built-in sonic transducers on either side of a spool section. Such sensors, shown in Figure 57.5, are generally called "radial" because of the transducer placement. A lower pipe size limit of radial-type sensors is about 100 mm.

57.2.5 Field Mounting

Radial-type sensors are often used to instrument an existing line, where it is desirable to make the installation without cutting out a section of the pipe. To meet this requirement, field-mountable transducers, cemented, drilled, or welded into an existing pipe section, can be used. Metering sections with diameters up to 13 m or more can be achieved in this way.

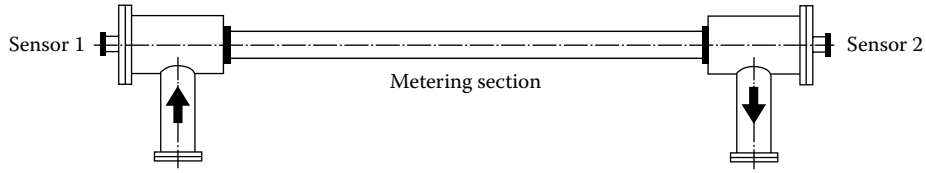


FIGURE 57.4 Axial sensor type. The ultrasonic pulse passes directly down the axis of the pipe.

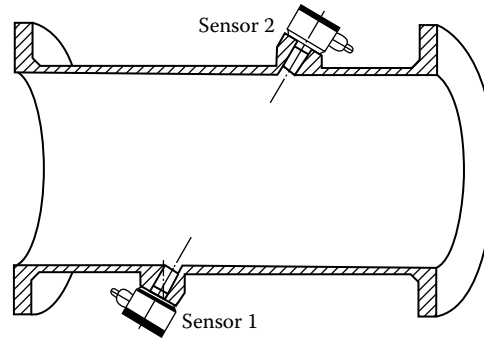


FIGURE 57.5 Radial sensor type. Manufacturers provide them as complete metering sections or as field-mounting sensors for existing conduits.

57.2.6 Clamp on

If there is a need for an installation where the pipe wall is not penetrated by the transducers, clamp-on systems are the right choice. Achieving a lower accuracy and being somewhat more complex to calibrate, clamp-on systems have their entitlement in applications where an easy movement of the metering section is an important requirement or where an existing process cannot be interrupted. The transducers are mounted on a calibration device and acoustically coupled to the pipe wall with grease and/or epoxy.

57.2.7 Open Channels and Special Applications

In open channels, the transducers are normally mounted on or dug into the channel walls. Figure 57.6 shows an example of an open-channel sensor. The piezoceramic element is placed on a sphere to achieve a wide range of mounting possibilities. Sometimes, existing pipe sections are completely enclosed with rock or concrete. In these cases, the transducers can be fixed in the wall of the conduit. For small diameters, the resulting protrusion of the transducers into the metering section must be taken into account.

57.2.8 Measurement of Flow in Closed Conduits

The most important issue in applying ultrasonic flowmeters is the understanding of the effects of the velocity profile of the flowing fluid within the conduit. The flow profile depends on the fluid, the Reynolds number Re, the relative roughness and shape of the conduit, upstream and downstream disturbances, and other factors. Transit time flowmeters give an average flow velocity v along the sonic path. The acoustic flow rate Q_{ADM} is therefore calculated through $Q = \overline{v}A$, with $\overline{v}$ the area-averaged flow velocity

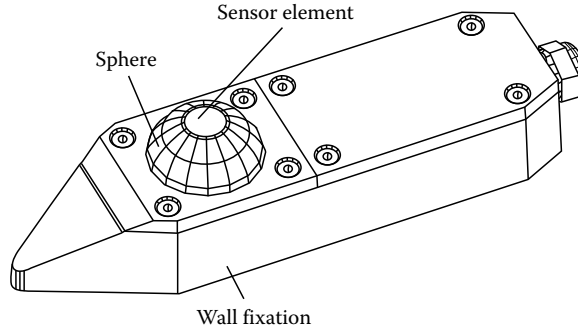


FIGURE 57.6 Open-channel sensor with sensor element placed on a movable sphere. Alignment of two sensors can be executed by laser equipment.

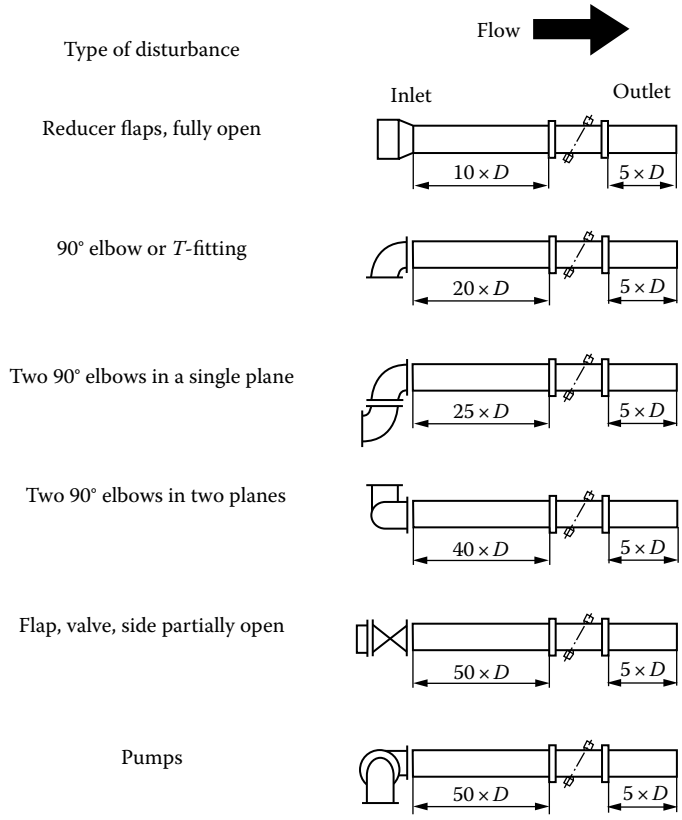


FIGURE 57.7 Minimum straight run requirements for a 1% accuracy of a single-path transit time flowmeter.

and A the cross section of the conduit. In order to obtain the area-averaged flow velocity $\bar{v}$, the measured velocity $\bar{v}_a$ must be corrected by a hydraulic coefficient k_h that depends on the type of the conduit and the Reynolds number. In order to achieve maximum performance and accuracy of ultrasonic flowmeters, one has to keep to sufficient straight run requirements as shown for some examples in Figure 57.7. By doing so, a typical accuracy of 1% of reading or better can be achieved, even when applying a single-path measurement system. Reduced straight runs lead to reduced accuracy. In some applications, this reduced accuracy is acceptable; if not, a multipath ultrasonic flowmeter must be installed. These flowmeters provide averaging of the various error-producing flow components. Accuracy of 0.5% of reading can be achieved even under nonideal conditions or insufficient straight runs. Figure 57.8 shows four examples of possible sonic path arrangements in a closed conduit.

57.2.9 Single Path with Circular and Rectangular Cross Sections

The Reynolds number Re is given by

$$\mathrm{Re} = \frac{\bar{v} \cdot D}{\nu} \tag{57.5}$$

where

$\bar{v}$ is the mean velocity over the cross section [m s^{-1}]
D is the pipe diameter
ν is the temperature-dependent cinematic viscosity

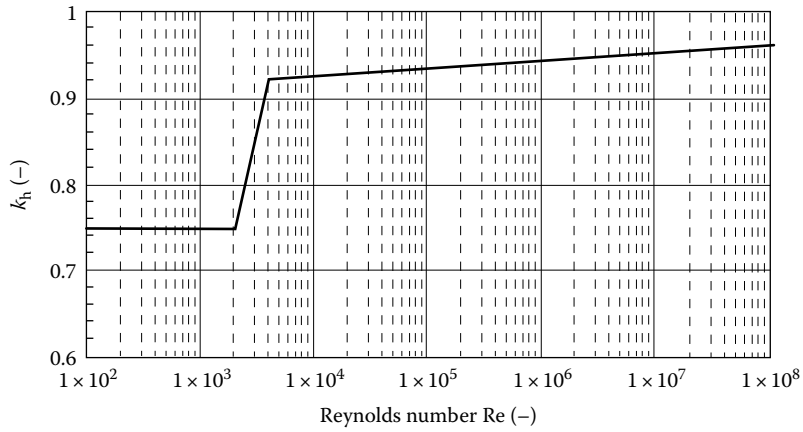


FIGURE 57.8 Dependence of the k_h actor for single-path measurement on the Reynolds number. Between a Reynolds number of 2000 and 4000, the transitional flow regime occurs.

In normal piping, a laminar flow exists as long as the Reynolds number is below about 2600. The shape of the velocity profile conforms to a parabola, and the velocity of a point on the profile is given by

$$v(r) = v_{\max}\left(1 - \left(\frac{r}{R}\right)^2\right) \tag{57.6}$$

where
r is the variable radius
R is the pipe radius

Between a Reynolds number Re of 2600 and 4000, the transitional flow regime with continuous switching between laminar and turbulent velocity profile comes into existence. When the Reynolds number exceeds 4000, the velocity profile enters the turbulent flow regime. Nikuradse [6] showed that the turbulent velocity profile of an axis-symmetrical flow in a closed conduit without swirl and sufficient inlet and outlet sections and smooth walls can be expressed by

$$v(r) = v_{\max}\left(\frac{R-r}{R}\right)^{1/n} \tag{57.7}$$

where, according to Nikuradse, n is a Reynolds number-dependent exponent given by

$$n = \frac{1}{(0.2525 - 0.0229 \times \log(\mathrm{Re}))} \tag{57.8}$$

With the definition of a hydraulic corrective coefficient k_h given by

$$k_h = \frac{\overline{v}}{\overline{v}_a} \tag{57.9}$$

where
$\overline{v}$ is the mean velocity over the cross section
$\overline{v}_a$ is the mean velocity along the sonic path

and integrating over the cross section according to

$$k_h = \frac{1/A\int_0^R \int_0^{2\pi} v(r) r \mathrm{d}\theta}{1/2R\int_{-R}^{R} v(r)\mathrm{d}r} \tag{57.10}$$

one obtains for laminar flow a hydraulic corrective coefficient of $k_h = 0.75$.

In the turbulent flow regime, according to Nikuradse [6], the hydraulic corrective coefficient k_h, dependent on the Reynolds number Re, can be expressed by

$$k_h = \frac{1}{1.125 - 0.011 \times \log(\mathrm{Re})} \tag{57.11}$$

for a circular cross section and

$$k_h = 0.79 + 0.02 \times \log(\mathrm{Re}) \tag{57.12}$$

for a rectangular cross section.

In hydropower applications, the Reynolds number Re generally exceeds the value of 4000. Figure 57.8 shows the dependence of the k_h factor in closed conduits with circular cross section on the Reynolds number. The Reynolds number not only changes its value as a function of the flow velocity v for a given diameter D but also is strongly dependent on the temperature-dependent cinematic viscosity $\nu(T)$. Not taking into account for correct value of Re can easily lead to errors of the k_h factor and thus to the flow rate Q on the order of 2%–3%. Modern transit time meters with microprocessors update the k_h factor at a rate of four times a second and by measuring the temperature T of the fluid at the same rate. A correct k_h factor is obtained and hence a temperature and Reynolds number-compensated flow rate Q.

57.2.10 Multipath Integration in Circular Cross Sections

In reality, however, the straight run requirements as defined in Figure 57.7 cannot always be kept. In addition, cross flow errors occur when nonaxial components of velocity in a pipe alter the transit times of a pulse between the sensors. Nonaxial velocities are caused by such disturbances in closed conduits as bends, asymmetric intake flows, and discontinuities in the pipe wall or pumps. It has become an accepted practice to eliminate the sensitivity of an acoustic flowmeter to velocity distributions by increasing the number of acoustic paths n. Additional paths not only decrease substantially the velocity distribution error but also reduce the numerical integration error as well as errors due to path misalignment. This suggests that for accurate measurements for which a few tenths of a percent error are significant, the cost of installing more sensors for a four-path arrangement according to Figure 57.9

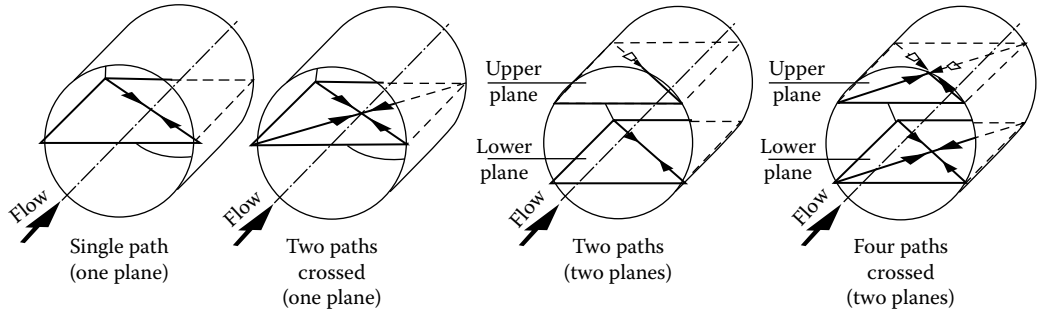


FIGURE 57.9 Possible sonic path arrangements for transit time flowmeters in closed conduits. Using multiple paths leads to reduced straight run requirements.

can be justified. In a multipath measurement system, according to the integration method described by IEC 41, Appendix J1 [7], the flow rate Q_{ADM} can be expressed by

$$Q_{ADM} = k\frac{D}{2}\sum_{i=1}^{n} \overline{v}_{ai} W_i L_i \sin\phi_i \tag{57.13}$$

where

$\overline{v}_{ai}$ is the velocities along the acoustic path i
W_i is the corresponding weighting factor
L_i is the corresponding path length
ϕ_i is the corresponding path angle

The method described in IEC 41 needs a very accurate transducer positioning due to the fact that the weighting factors W_i obtained by mathematical analysis are only valid when the sensors are positioned at the correct locations. Misalignment of the acoustic paths in conjunction with fixed weights can lead to considerable errors.

57.3 Measurement of Flow in Open Channels

Open-channel flow measurement is used in many applications, including water supply networks, hydrography, allocation of water for irrigation and agricultural purposes, and sewage treatment plants. Discharge measurements in rivers and open channels are often computed by means of a rating curve, used to convert records of water level readings into flow rates. The rating curve is developed using a set of discharge measurements and water level in the stream and must be checked periodically to ensure that the level–discharge relationship has remained constant; many phenomena can cause the rating curve to change so that the same recorded water level produces a different discharge. This is the case of open channels under changeable hydraulic conditions due, for example, to backwater effects and gates and where a univocal stage–discharge relationship does not exist. In this context, acoustic flowmeter application is extremely interesting and is currently experiencing wide success in water management. In fact, while different flow rate values can correspond to a given water level in relation to the hydraulic characteristics, there is always univocal relationship between the acoustic wave propagation velocity in a flowing fluid and the flowing fluid itself.

By means of the ultrasonic technique, discharge through open channels can be determined using single- and multipath technology. In a single-path configuration, particular attention must be paid to define the vertical velocity distribution (Figure 57.10) in order to achieve a good level of accuracy by a single "line" velocity reading $\overline{v}_{az}$ at the distance z above the bottom. On the other hand, in a multipath configuration, the mean line velocity profile is well described. In this case, special attention must be paid to the integration method used to determine the flow rate from the acoustic path readings.

57.3.1 Single Path

In single-path measurements, using the area–velocity method, the flow rate Q is calculated through $Q = \overline{v}A$, with $\overline{v}$ (ms^{-1}) the area-averaged flow velocity and A the cross section. To obtain the velocity average $\overline{v}$ over the entire cross section, the mean path velocity $\overline{v}_{az}$, measured by the acoustic flowmeter at a given depth z, must be corrected by a dimensionless hydraulic corrective coefficient k_h according to the h relation $\overline{v} = k_h\overline{v}_{az}$. In general, the coefficient k_h reflects the influence of the horizontal and vertical velocity profile. It mainly depends on the water level, on the cross section shape, and on the boundary

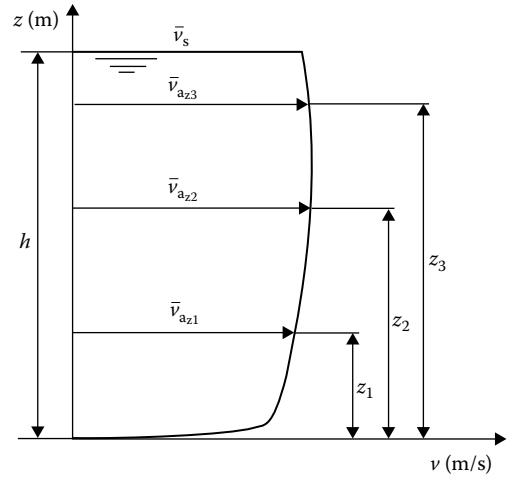


FIGURE 57.10 Open channel: possible vertical velocity distribution.

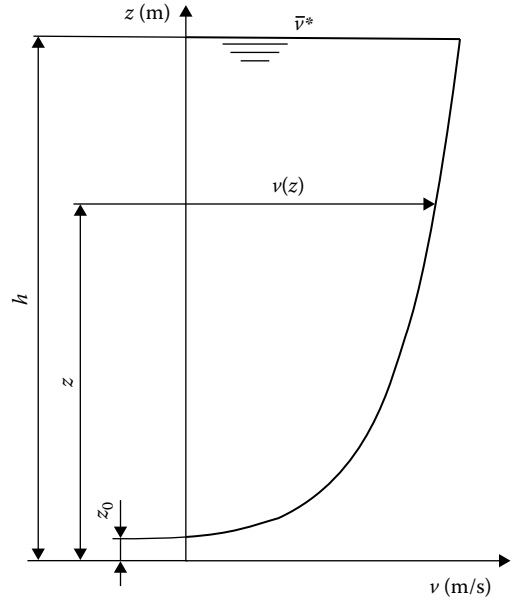


FIGURE 57.11 Logarithmic law describing open-channel velocity profile.

roughness. The mean vertical velocity profile can be described by the well-known logarithmic law (Figure 57.11) given by [8]

$$\bar{v}(z) = \left(\frac{\bar{v}^*}{k}\right) \ln \frac{z}{z_0} \tag{57.14}$$

where

$\bar{v}(z)$ is the mean velocity a distance z above the bottom
k is the von Kàrmàn's turbulence constant
z is the distance above the bottom
$\bar{v}^*$ is the shear velocity
z_0 is the constant of integration, dependent on the boundary roughness

When the boundary surface is hydraulically rough, z_0 has been found to depend solely on the roughness height k_s according to the relation $Z_0 = k_s/30$. Integrating Equation 57.14 over the total water height h and substituting for $\bar{v}$, the following equation is obtained:

$$k_h = \frac{\ln(h/k_s) - 1}{\ln(z/k_s)} \tag{57.15}$$

assuming z_0 is negligible with respect to h.

In Figure 57.12, the logarithmic k_h factor is represented for two different types of bottom roughness. However, in many practical applications, it is often difficult to define correct values for k_s. For this reason, another k_h model has been developed on the basis of the power law [9,10]. In this formula, the exponent $1/m$ is not constant anymore, but depends on the roughness and the hydraulic radius to take into account the influence of the channel shape. The mean vertical velocity profile $\bar{v}(z)$, according to the power law, can be expressed by the following formula (Figure 57.13):

$$\bar{v}(z) = \bar{v}_s \left(\frac{z}{h}\right)^{1/m} \tag{57.16}$$

where

h is the current water level
$\bar{v}$ is the mean line velocity at the free surface (maximum value)
$1/m$ is the exponent

From the integration of Equation 57.16 over the total water depth, one obtains the following expression for the k_h factor:

$$k_h = \frac{m}{m+1}\left(\frac{h}{z}\right)^{1/m} \tag{57.17}$$

The value of m, depending on the roughness, can be expressed using the dimensional friction factor f of the Darcy–Weisbach formula [8], according to

$$m = k\sqrt{\frac{8}{f}} \tag{57.18}$$

where k is the Von Kàrmàn constant, varying from 0.2 to 0.4.

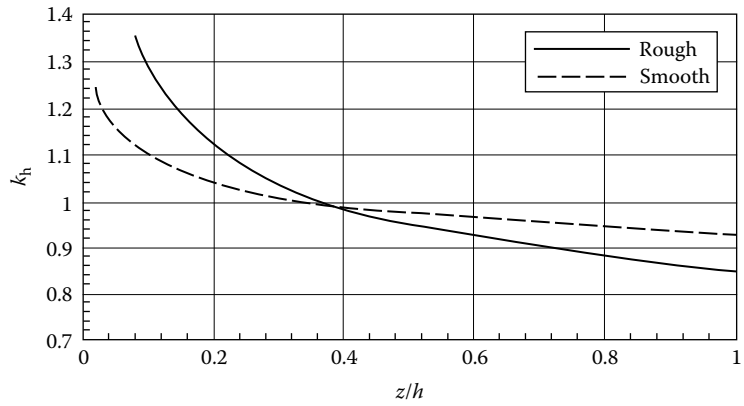


FIGURE 57.12 Logarithmic k_h factor in open channels for two different roughnesses k_s.

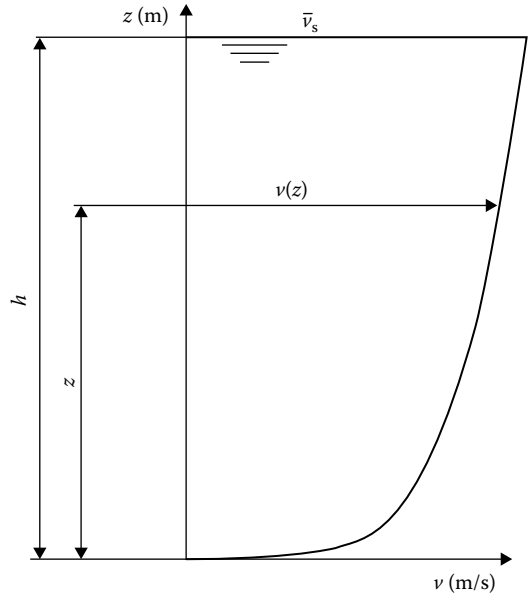


FIGURE 57.13 Power law describing open-channel velocity profile.

k depends on the suspended load (low values for high turbidity) or can be expressed using the Manning formula by the relation [11]

$$m = \left(\frac{k}{\sqrt{g}} \right) \frac{R_h^{1/6}}{n} \tag{57.19}$$

where

g is the gravity acceleration
n is the Manning's roughness coefficient
R_h is the hydraulic radius

In Figure 57.14, the corrective factor is represented for smooth and rough surfaces. By means of this relation, an easier field application has been obtained due to wide familiarity with the Manning formula in open-channel flow computation because of its simplicity of form, high versatility, and satisfactory results.

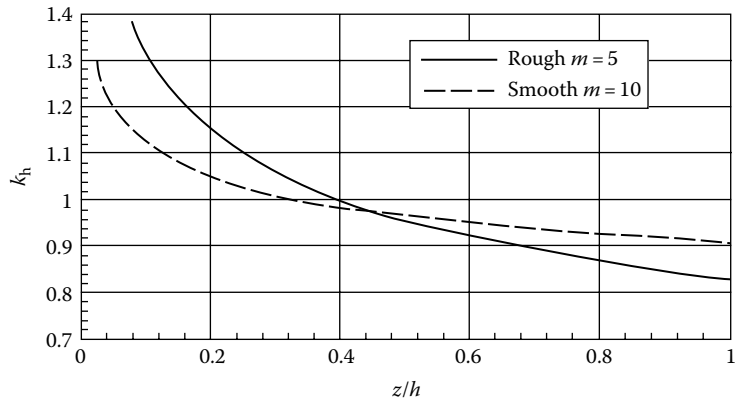


FIGURE 57.14 Power law k_h factor in open channels for two different roughnesses.

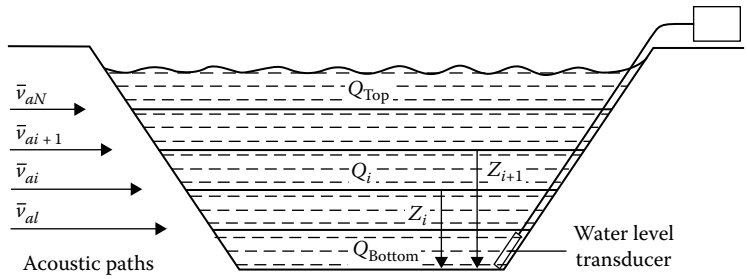


FIGURE 57.15 Multipath measurement in open channels by the mean section method. The flow velocity is measured by several levels.

57.3.2 Multipath

The foregoing equations for the mean vertical velocity profile in open channels predict that the maximum mean velocity occurs at the free surface. Field and laboratory measurements, however, demonstrate that the maximum mean velocity occurs below the free surface, strongly depending on the ratio B/h, with B being the channel width. These observations show that a 1-D velocity distribution law cannot always completely describe flow profiles in open channels. Therefore, to reduce uncertainties in velocity profile description and to achieve high accuracies even under unfavorable hydraulic conditions, a multipath configuration must be used.

In multipath measurement, using the "mean section method," the flow velocity is measured at several levels between the free surface and the channel bottom (Figure 57.15). The total discharge Q_{ADM} is performed by the relation [12]

$$Q_{ADM} = Q_b + Q_t + \sum_{i=1}^{n}\left(\frac{\bar{v}_{ai} + \bar{v}_{ai+1}}{2}[A(z_{i+1}) - A(z_i)]\right) \tag{57.20}$$

where

Q_b is the flow rate in the bottom section with the bottom velocity obtained from the lowest path velocity by correction for bottom friction
Q_t is the flow rate at the highest active section with velocity v_{top}, interpolated from the velocity profile
$\bar{v}_{ai}$ is the mean velocity along the ith acoustic path
$A(z_i)$ is the cross section below the ith path

Modern microprocessor-controlled ultrasonic flowmeters can cope with either single-path or multipath measurement in open channels, using the logarithmic or the power law for single-path measurement. In addition, up-to-date completely modular systems are able to use the major part of the equipment for measuring both in open channels and closed conduits.

57.4 Application: Penstock Leak Detection with Surge Chambers

Penstock leak detection is a typical acoustic flowmeter application [13]. It is used for immediate recognition of pipe rupture and leak losses in penstocks. Two flowmeters are installed at opposite ends of the penstock: one for measuring Q_{up} as near as possible to the intake and the other for measuring Q_{down} at the powerhouse entrance. In this description, the upstream flow is compared to the downstream flow and the flow difference. $\Delta Q = Q_{up} - Q_{down}$ is calculated and supervised. If the difference ΔQ exceeds a given threshold, the system enunciates alarms or valve closure contacts.

In hydropower applications, another problem arises due to the presence of surge chambers (Figure 57.16). This hydraulic structure causes mass oscillations due to load changes of the turbine.

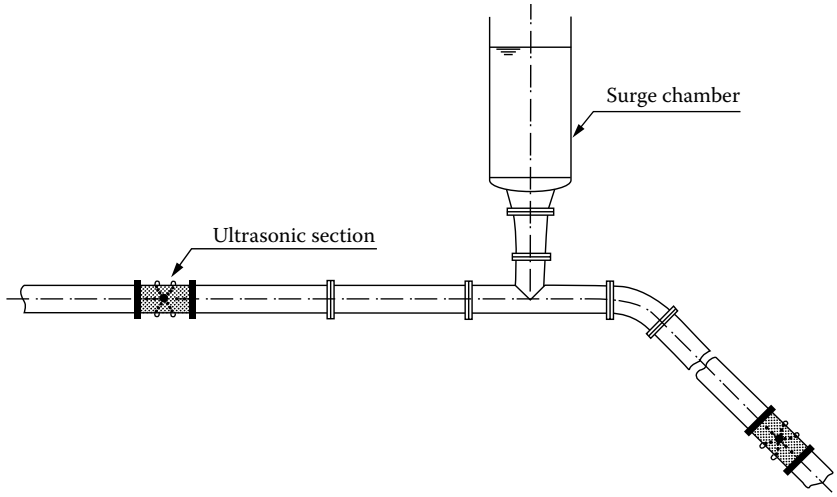


FIGURE 57.16 Penstock leak detection by transit time flowmeter. The penstock contains a surge chamber that causes mass oscillation in case of load changes of the turbine.

Up to now, if there was a surge tank in a penstock, one had two possibilities. On the one hand, one could divide the penstock into two parts, one before the surge chamber and one after it, and protect them separately. This solution is expensive and very often impractical because the surge chamber is built into the rock and inaccessible. On the other hand, one could develop something like a "leak detection algorithm," which includes the oscillatory behavior of a penstock. Such a system, however, needs extensive field tests and the knowledge for exact settings of resonant frequency, damping factor, and thresholds. The new technology in ultrasonic flowmeters, combined with accurate water level sensors, can offer a solution to this problem, by applying the formula

$$\Delta Q = Q_{\text{up}} - Q_{\text{down}} - \frac{\Delta V_{\text{chamber}}}{\Delta t} \tag{57.21}$$

TABLE 57.1 Companies Manufacturing and Distributing Ultrasonic Transit Time Flowmeters

Rittmeyer AG
P.O. Box 2143, CH-6302 Zug,
Switzerland, Tel: (+4141)-767-1000, Fax: (+4141)-767-1075,
instrumentation@rittmeyer.ch, http://www.rittmeyer.com/

Krohne Messtechnik GmbH & Co. KG
Postfach 10 08 62, Ludwig-Krohne-Strasse 5, 4100 Duisburg 1,
Germany, http://www.krohne.com/

Danfoss
A/S DK-6430, Nordborg, Denmark,
Tel: (+45) 74 88 22 22

Panametrics, Inc.
221 Crescent Street,
Waltham, MA 02254, Tel: (617) 899-2719, http://www.
panametrics.com/

Ultraflux le technoparc, 17, rue Charles
Edouard Jeanneret, 78306 Poissy Cedex,
France, Tel: 33(1)39 79 26 40

Accusonic Division, ORE International, Inc.
P.O. Box 709, Falmouth, MA 02541,
Tel: (508) 548-5800, http://www.ore.com/

Fuji Electric Co., Ltd.
12-1 Yurakucho 1-chome, Chiyoda-ku, Tokyo 100, Japan,
Tel: Tokyo 211-7111

Crouzet SA
Division "Aérospatial" 25, rue Jules-Védrines,
26027 Valence Cedex, France. Tel: 75 79 85 11

Nusonics Inc.
11391 E. Tecumseh St., Tulsa, OK 74116-1602,
Tel: (918) 438-1010

Changes of the volume of the surge chamber due to mass oscillations are taken into account, leading to a better and more realistic behavior of penstock leak detection in the presence of surge tanks.

57.4.1 Instrumentation and Manufacturers/Distributors

Table 57.1 gives an overview of some companies manufacturing and distributing transit time flowmeters. Prices of transit time flowmeters have a very wide range due to the wide versatility of different applications and are therefore difficult to list accurately. Generally, ultrasonic flowmeters are seldom sold as a "pre-packed" instrument. For this reason, the price for a metering section, including installation, varies from a few $1,000 to nearly $100,000.

References

1. M. Barmettler and P. Gruber, Anwendung von Oversampling-Verfahren zur Erhöhung der Auflösung digital erfasster Signale, *Technisches Messen*, Oldenbourg Verlag, 1992.
2. D. Hoppe, Kombinierte Zählung und Abstandsbestimmung von Impulssignalen, *Technisches Messen*, Oldenbourg Verlag, 10, 1991.
3. R. Krimholtz, D. Leedom, and G. Matthaei, New equivalent circuits for elementary piezoelectric transducers, *Electron. Lett*., 6, 398, 1970.
4. P. D. Edmonds, *Methods of Experimental Physics, Ultrasonics*, New York: Academic Press, 1981.
5. D. W. Spitzer, *Flow Measurement, Practical Guides for Measurement and Control*, Research Triangle Park, NC: Instrument Society of America, 1991.
6. J. Nikuradse, Gesetzmässigkeiten der turbulenten Strömung in glatten Rohren, *VDI Verlag GmbH*, 1932.
7. F. L. Brand, Akustische Verfahren zur Durchflussmessung, *Messen, Prüfen Automatisieren*, April 1987.
8. *International Standard IEC 41*, 3rd edn., 1991.
9. R. H. French, *Open Channel Hydraulics*, New York: McGraw-Hill, 1985.
10. Chen-Iung Chen, *J. Hydraulic Eng*., 379, 117, 1990.
11. M. F. Karim and J. F. Kennedy, *J. Hydraulic Eng*., 162, 113, 1987.
12. G. Grego, M. Baldin, et al., *Application of an Acoustic Flowmeter for Discharge Measurement in the Po*.
13. G. Grego and M. Baldin, *Energia Elettrica*, 1, 52, 72, 1995.
14. H. P. Vaterlaus and H. Gabler, A new intelligent ultrasonic flowmeter for hydropower applications, *Int. Water Power & Dam Construction*, 46, 84–88, 1994.

Further Reading

Introduction to Ultrasonic Doppler Flowmeters, http://www.omega.com/prodinfo/ultrasonicflowmeters.html, accessed on July 26, 2012.

John Flood, Acoustic/Ultrasound, Ultrasonic Flowmeter Basics, http://www.sensorsmag.com/sensors/acoustic-ultrasound/ultrasonic-flowmeter-basics-842, accessed on July 26, 2012.

Ultrasonic Flowmeter, Portable Flowmeter, Triton Process Technologies, http://www.ultrasonicflowmeter.in/index.html, accessed on July 24, 2012.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

58

Vortex-Shedding Flowmeters

58.1 Introduction **58**-1
58.2 Principle of Operation **58**-2
58.3 Calculation of Mass Flow and Standard Volume **58**-4
58.4 Flowmeter Construction **58**-4
Flow Tube • Transmitter
58.5 Application Considerations **58**-9
Meter Selection • Meter Installation • Meter Configuration
58.6 Recent Developments **58**-11
Multisensor • Single Sensor
Further Information **58**-12
Partial List of Vendors and Manufacturers **58**-12

Wade M. Mattar
The Foxboro Company

James H. Vignos
The Foxboro Company

58.1 Introduction

The *vortex-shedding flowmeter* first emerged 25–30 years ago and has steadily grown in acceptance since then to be a major flow measurement technique. Its appeal is due, in part, to the fact that it has no moving parts yet produces a frequency output that varies linearly with flow rate over a wide range of Reynolds numbers. The vortex meter has a very simple construction, provides accuracy (1% or better) comparable to higher-priced and/or more maintenance-intensive techniques, and works equally well on liquids and gases. In addition, it is powered primarily by the fluid and lends itself more readily than other linear flow devices to two-wire operation. Comparing the vortex-shedding flowmeter to an orifice plate, the former has higher accuracy and rangeability, does not require complex pressure impulse lines, is less sensitive to wear, and, for volumetric flow measurement, does not require the need to compensate for fluid density.

Industrial vortex-shedding flowmeters are normally available in pipe sizes ranging from 15 to 300 mm (1/2–12 in.), with some manufacturers offering sizes up to 400 mm (16 in.). Flow ranges covered depend on fluid properties and meter design. Typical ranges for a 15 mm meter are

- Water at 21 °C (70 °F), 0.06–2.2 L s^{-1} (1–35 gal min^{-1})
- Air at 16 °C (60 °F) and 101 kPa (14.7 psia), 1.1–15.7 L s^{-1} (140–2000 ft^3 h^{-1})
- Dry saturated steam at 689 kPa (100 psig), 4.5–225 kg h^{-1} (10–500 lb h^{-1})

Typical ranges for a 300 mm (12 in.) meter are

- Water at 21 °C (70 °F), 5.4–5400 L s^{-1} (85–8500 gal min^{-1})
- Air at 16 °C (60 °F) and 101 kPa (14.7 psia), 157–12,500 L s^{-1} (20,000–1,600,000 ft^3 h^{-1})
- Dry saturated steam at 689 kPa (100 psig), 1,240–124,000 kg h^{-1} (2,750–275,000 lb h^{-1})

Temperature capability ranges from cryogenic temperatures up to 427 °C (800 °F). Pressure capability as high as 20.7 MPa (3000 psig) is available.

58.2 Principle of Operation

Probably the first time, ages ago, that anyone placed a blunt obstacle in a flowing fluid, he or she observed the whirlpools or vortices that naturally form and shed downstream. In everyday life, examples of vortex shedding are numerous. The undulation of a flag is due to vortex shedding from the pole, and the singing of telephone wires in a strong wind is due to shedding from the wires. Analysis by Theodore von Karman in 1911 described the stability criterion for the array of shed vortices. Consequently, when stable array of vortices form downstream from an obstacle, it is often referred to as the von Karman vortex street (Figure 58.1).

Very early on, it was noted that, for a large class of obstacles, as the velocity increased, the number of vortices shed in a given time (or frequency of vortex shedding) increased in direct proportion to the velocity. The dimensionless Strouhal number, St, is used to describe the relationship between vortex-shedding frequency and fluid velocity and is given by

$$\mathrm{St} = \frac{f \times d}{U} \tag{58.1}$$

where

f is the vortex-shedding frequency
d is the width of shedding body
U is the fluid velocity

Alternatively,

$$U = \frac{f \times d}{\mathrm{St}} \tag{58.2}$$

Although early studies were conducted in unconfined flow, it was later observed that vortex shedding also occurred in confined flow, such as exists in a pipe (see Figure 58.2). For this case, the average

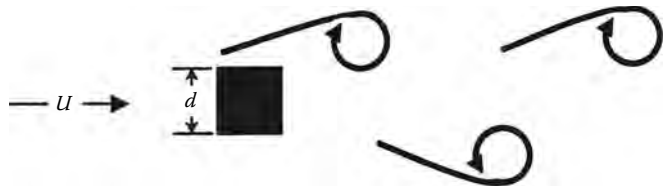


FIGURE 58.1 Von Karman vortex street.

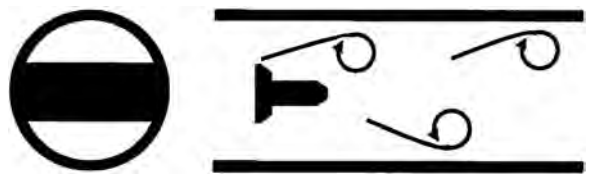


FIGURE 58.2 Vortex shedding in a pipe.

fluid velocity, $\bar{U}$, and the meter Strouhal number, St′, replace the fluid velocity and Strouhal number, respectively, in Equation 58.2 to give

$$\bar{U} = \frac{f \times d}{\text{St}'} \tag{58.3}$$

Since the cross-sectional area, A, of the pipe is fixed, it is possible to define a flowmeter K factor, K, that relates the volumetric flow rate (Q) to the vortex-shedding frequency. Given that

$$Q = A \times \bar{U} \tag{58.4}$$

from Equation 58.3, one obtains

$$Q = \left(\frac{A \times d}{\text{St}'} \right) \times f \tag{58.5}$$

defining

$$K = \frac{\text{St}'}{A \times d} \tag{58.6}$$

resulting in

$$Q = \frac{f}{K} \tag{58.7}$$

Vortex-shedding frequencies range from less than 1 Hz to greater than 3000 Hz, the former being for large meters at low velocities and the latter for small meters at high velocities.

For a vortex-shedding flowmeter, an obstacle is chosen that will produce a constant K factor over a wide range of pipe Reynolds numbers. Thus, simply counting the vortices that are shed in a given amount of time and dividing by the K factor will give a measurement of the total volume of fluid that has passed through the meter. A typical K factor vs. Reynolds number curve is shown in Figure 58.3.

The variation in K factor over a specified Reynolds number range is sometimes referred to as *linearity*. For the example in Figure 58.3, it can be seen that between Reynolds numbers from 15,000 to 2,000,000, the K factor is the most linear. This is referred to as the *linear range* of the shedder. The wider the linear range a shedder exhibits, the more suitable the device is as a flowmeter.

At Reynolds numbers below the linear range, linearization is possible but flowmeter uncertainty can increase.

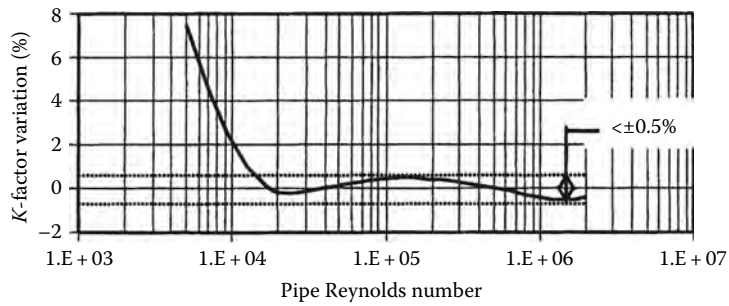


FIGURE 58.3 Typical K factor curve.

58.3 Calculation of Mass Flow and Standard Volume

Although the vortex flowmeter is a volumetric flowmeter, it is often combined with additional measurements to calculate or infer mass flow or standard volume.

To determine mass flow, $\dot{M}$,

$$\dot{M} = \rho_f \times Q = \rho_f \times \frac{f}{K} \tag{58.8}$$

where ρ_f is the fluid density at flowing conditions.

It is often desirable to know what the volumetric flow rate would be at standard process conditions with respect to pressure and temperature. This is referred to as *standard volume*. In different parts of the world and for different industries, the standard temperature and pressure can be different. The fluid density at standard conditions is referred to as the *base density*, ρ_b.

To calculate *standard volume*, Q_v,

$$Q_v = \left(\frac{\rho_f}{\rho_b}\right) \times \frac{f}{K} \tag{58.9}$$

58.4 Flowmeter Construction

The vortex-shedding flowmeter can be described as having two major components: the *flow tube* and the *transmitter*. Both are described as follows.

58.4.1 Flow Tube

The flow tube is composed of three functional parts: the flowmeter *body*, which contains the fluid and acts as housing for the hydraulic components; the *shedder*, which generates the vortices when the fluid passes by; and the *sensor*(*s*), which by some transducing means detects the vortices and produces a usable electric signal.

58.4.1.1 Flowmeter Body

The pressure-containing portion of the vortex flowmeter, sometimes referred to as the *flowmeter body*, is available in two forms: wafer or flanged (see Figure 58.4). The wafer design, which is

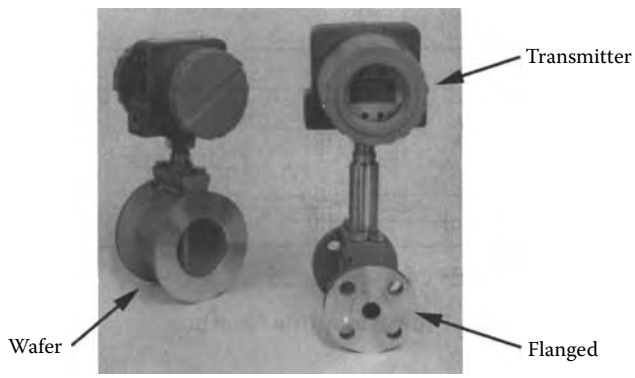


FIGURE 58.4 Vortex flowmeter construction.



FIGURE 58.5 Shedder cross sections.

sandwiched between flanges of adjacent pipe, is generally lower in cost than the flanged design, but often its use is limited by piping codes.

58.4.1.2 Shedder

The *shedder* spans the flowmeter body along the diameter and has a constant cross section along its length. Typical cross sections are shown in Figure 58.5.

58.4.1.3 Sensors

When shedding is present, both the pressure and velocity fields in the vicinity of the shedder will oscillate at the vortex-shedding frequency. *Pressure or velocity sensors* are used to transform the oscillating fields to an electric signal, current, or voltage, from which the vortex-shedding frequency can be extracted.

Figure 58.6 shows an example of a piezoelectric differential pressure sensor located in the upper portion of the shedder, which converts the oscillating differential pressure that exists between the two sides of the shedder into an electric signal. Some other sensing means utilized to detect vortex shedding are capacitive, thermal, and ultrasonic. Often times, the flowmeter electronics are mounted remotely from the flow tube. This might require a local preamplifier to power the sensor or boost its signal strength.

Since the sensor is the most likely mechanical component in a vortex flowmeter to fail, most designs have a provision for replacement of the sensors. In some cases, they can be replaced without removing the flowmeter from the pipeline. An example of an isolation manifold for sensor replacement under process conditions is shown in Figure 58.7.

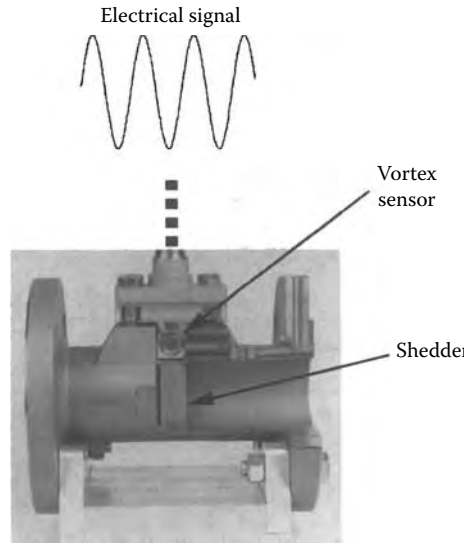


FIGURE 58.6 Example of vortex sensor.

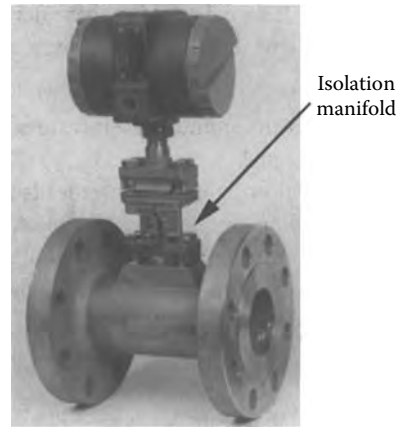


FIGURE 58.7 Isolation manifold for sensor replacement.

58.4.2 Transmitter

Vortex flowmeter *transmitters* can be classified into two broad groups: analog and intelligent (or smart). Communication with the analog transmitter is carried out via local manual means, including switches and jumper wires. Communication with the newer intelligent device is carried out via digital electronic techniques. Both types of transmitters are two-wire devices. One of the more important features of the intelligent transmitter is that it allows application-specific information to be loaded into the transmitter. This information is then used internally to automatically tailor the transmitter to the application, including calibration of the 4–20 mA output. Before describing these devices, it is useful to consider the three most common forms of flow measurement signals provided by vortex transmitters.

58.4.2.1 Measurement Output Signals

The three most common ways for a transmitter to communicate measurement information to the outside world are 4–20 mA, digital, and pulse signals.

The *4–20 mA signal* is the DC current flowing in the power leads to the transmitter. This current is directly proportional to the vortex-shedding frequency and, hence, is also linear with flow (see Figure 58.8). Four milli amperes corresponds to zero flow and 20 mA to the maximum flow rate,

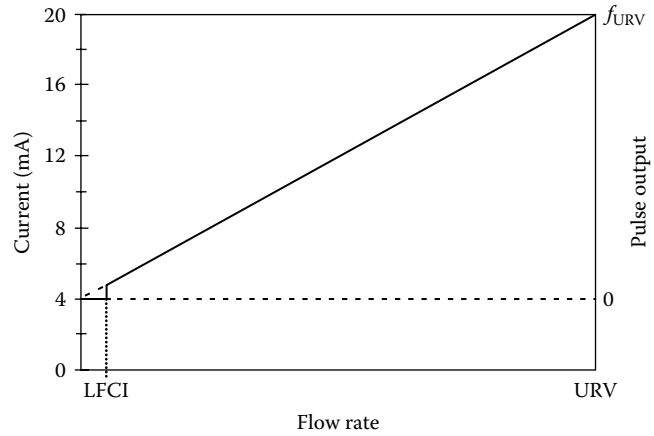


FIGURE 58.8 4–20 mA and pulse outputs.

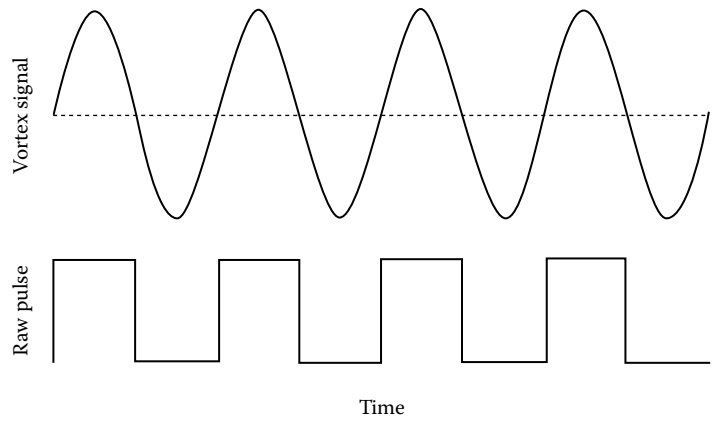


FIGURE 58.9 Raw pulse output.

that is, the upper range value (URV) of the meter. A frequency-to-analog converter in the analog meter and a digital-to-analog converter in the intelligent meter produce this output.

The *digital signal* is a digitized numeric value of the measured flow rate in engineering units transmitted over the two wires powering the meter.

The *pulse signal* is a squared-up version of the raw vortex signal coming from the sensor (see Figure 58.9) and is accessible via a pair of electric terminals inside the transmitter housing. The frequency of the pulse signal is either identical to the vortex-shedding frequency (raw pulse) or some multiple thereof (scaled pulse). As discussed in Section 58.2, in either case, the frequency of the pulse signal is linearly proportional to flow rate, going from zero to the frequency at the URV, f_{URV} (see Figure 58.8).

As shown in Figure 58.8, at a low but nonzero flow rate, the frequency and mA signals drop to 0 Hz and 4 mA, respectively. The flow rate at which this abrupt change takes place is normally referred to as the low-flow cut-in (LFCI) or cutoff. The reason for this forced zero is to avoid erroneous flow measurements at low flow, which result from process noise, including hydrodynamic fluctuations, mechanical vibration, and electrical interference. The digital signal also drops to zero below the LFCI flow rate.

58.4.2.2 Analog Transmitter

Originally, *analog transmitters* were constructed entirely of analog electronic components. Today, they are built around a combination of analog and digital electronic components. In either case, the measurement output is in the form of a raw pulse and/or a 4–20 mA signal. Depending on the particular transmitter, one or more of the following functions are available for tailoring the device via mechanical means to the specific application:

1. Signal output selection: If the transmitter provides both raw pulse and 4–20 mA signals, but not simultaneously, a means is available for selecting the one desired.
2. 4–20 mA calibration: Use of this signal requires that 20 mA corresponds to the desired URV. This is accomplished by inputting, via a signal or pulse generator, a periodic signal whose frequency corresponds to the upper range frequency (URF) and adjusting the output current until it reads 20 mA. The URF, which is the frequency corresponding to the vortex-shedding frequency at the desired URV, is calculated using the equation $\text{URF} = K \times \text{URV}$.

   In order to achieve the accuracy specified by the manufacturer, the K used in the preceding calculation must be corrected for process temperature and piping effects according to the

manufacturer's instructions. The temperature effect is a result of thermal expansion of the flow tube and is described by

$$\Delta K(\%) = -300 \times \alpha \times (T - T_0) \tag{58.10}$$

where

α is the thermal expansion coefficient of the flow tube material
T_0 is the fluid temperature at which the meter was calibrated

If the shedder and meter body materials are different, α must be replaced by $(2\alpha_1 + \alpha_2)/3$, where α_1 is the thermal expansion coefficient of the meter body material and α_2 that of the shedder.

Piping disturbances also affect the *K* factor because they alter the flow profile within the flow tube. This will be discussed in more detail in Section 58.5.2.2.

3. LFCI: For optimum measurement performance, the LFCI should be set to fit the specific application. The goal is to set it as low as possible, while at the same time avoiding an erroneous flow measurement output.
4. Filter settings: To reduce noise present on the signal from the sensor, electronic filters are built into the transmitter. Normally, means are provided for adjusting these filters, that is, setting the frequencies at which they become active. By attenuating frequencies outside the range of the vortex-shedding frequency, which varies from one application to another, better measurement performance is achieved.

58.4.2.3 Intelligent Transmitters

Intelligent transmitters, which are microprocessor-based digital electronic devices, have measurement outputs that usually include two or more of the following: raw pulse, scaled pulse, 4–20 mA, and digital. With regard to the digital output, there is at present no single, universally accepted protocol for digital communication; however, a number of proprietary and nonproprietary protocols exist.

The presence of a microprocessor in the intelligent transmitter allows for improved functionality and features compared to the analog transmitter, including

- Elimination of the need for 4–20 mA calibration
- Automatic setting of LFCI
- Automatic setting of filters
- Adaptive filtering (active filters that track the vortex frequency)
- Digital signal conditioning
- *K* factor correction for temperature and piping disturbances
- Correction for nonlinearity of *K* factor curve, including the pronounced nonlinearity at low Reynolds numbers (see Figure 58.3)
- Integral flow totalization
- Digital measurement output in desired engineering units

Configuring—that is, tailoring the transmitter to a specific application—is carried out by one or more of the following digital communicators:

- Local configurator: a configurator, built into a transmitter, that has a display and keypad
- Handheld terminal: a palm-size digital device programmed for configuration purposes
- PC configurator: a personal computer containing configuration software
- System configurator: a digital measurement and control system with embedded configuration software

Using one of these configurators, the dataset of parameters that defines the configuration can be modified to fit the application in question. The details of this dataset vary, depending on the specific transmitter; however, the general categories of information listed as follows apply:

- Flow-tube parameters (e.g., tube bore, *K* factor, serial no.)
- User identification parameters (e.g., tag no., location)
- Transmitter options (e.g., measurement units, function selections)
- Process fluid parameters (e.g., fluid density and viscosity, process temperature)
- Application parameters (e.g., *K* factor corrections, URV, LFCI level)
- Output options (e.g., measurement output modes, damping, fail-safe state)

58.5 Application Considerations

58.5.1 Meter Selection

From a safety viewpoint, it is essential that the vortex flowmeter satisfies the appropriate electrical safety requirements and be compatible with the process, that is, be able to withstand the temperature, pressure, and chemical nature of the process fluid. From a mechanical viewpoint, it must have the proper end connections and, if required for critical applications, have a sensor that can be replaced without shutting down the process. Meter size and measurement output signal type are also very important selection factors.

58.5.1.1 Size

Contrary to what one might expect, the required *meter size* is not always the same as the nominal size of the piping in which it is to be installed. In some applications, selecting the size based on adjacent piping will not allow the low end of the required flow range to be measured. The appropriate criterion for selecting meter size is that the meter provides a reliable and accurate measurement over the entire required flow range. This could dictate a meter size that is less than the adjacent piping.

Pressure drop is a competing sizing criteria to that described previously. This drop is given by

$$\Delta P = C \times \rho_f \times \frac{Q^2}{D^4} \tag{58.11}$$

where

C is a constant dependent on meter design
D is the bore diameter of the flow tube

The tendency is to pick a flow tube with the same nominal diameter as the adjacent piping to eliminate the extra pressure drop introduced by a smaller-sized meter. However, in the majority of cases, this added drop is of little consequence.

The meter manufacturer can provide the needed information for making the proper selection. In some cases, sizing programs from manufacturers are available on the Internet in either an interactive or downloadable form.

58.5.1.2 Measurement Output Options

As mentioned previously, three types of *measurement outputs* are in current use: a 4–20 mA analog signal, a pulse train, and a digital signal. Some vortex meters will provide all three of these outputs, but not always simultaneously. It is essential that the meter has the output(s) required by the application.

58.5.2 Meter Installation

The performance specifications of a vortex flowmeter are normally established under the following conditions: (1) the flow tube is installed in a pipeline running full with a single-phase fluid; (2) the piping adjacent to the flow tube consists of straight sections of specified schedule pipe (normally schedule 40), typically a minimum of 30 PD (pipe diameters) in length upstream and 5 PD downstream of the flow tube with no flow-disturbing elements located within these sections; and (3) the meter is located in a vibration-free and electrical interference–free environment. As a consequence, certain constraints are placed on where and how the meter is installed in process piping if published performance specifications are to be achieved. These constraints are discussed later. Because meters from different suppliers differ in their sensitivity to the aforementioned influences, the statements made are of a qualitative nature. The specific manufacturer should be consulted for more quantitative information.

58.5.2.1 Location

The flowmeter should be located in a place where vibration and electrical interference levels are low. Both of these influences can decrease the signal-to-noise ratio at the input to the transmitter. This reduction can degrade the ability of the meter to measure low flows.

The meter should not be installed in a vertical line in which the fluid is flowing down because there is a good possibility that the pipe will not be full.

58.5.2.2 Adjacent Piping

Recommended practice is to mount the flowmeter in the process piping according to the manufacturer's stated upstream and downstream minimum straight-length piping requirements. These are typically 15–30 and 5 PD, respectively. Piping elements such as elbows or reducers upstream of the meter normally affect its *K* factor, but not its linearity. This allows a bias correction to be applied to the *K* factor. Many manufacturers provide bias factors for common upstream piping arrangements. Some who offer intelligent flowmeters make the corrections internally once the user has selected the appropriate configuration from a pick list. For piping elements and arrangements where the bias correction is not available, an in situ calibration should be run if the manufacturer's specified uncertainty is to be achieved. If this is not possible, calibration in a test facility with an identical configuration should be run.

The same situation as earlier applies if the pipe schedule adjacent to the meter differs from that under which the meter was calibrated.

To avoid disturbance to the flow, flange gaskets should never protrude into the process fluid.

The following recommendations apply if a control valve is to be situated near a vortex flowmeter. In liquid applications, the control valve should be located a minimum of 5 PD downstream of the flowmeter. This not only prevents disturbance to the flow profile in the flow tube but also aids in preventing flashing and cavitation (see as follows). In gas applications, the control valve should be installed upstream of the meter, typically a minimum of 30 PD upstream of the meter to ensure an undisturbed flow profile. Having the pressure drop across the valve upstream of the meter results in a decreased density and subsequent increased velocity at the flowmeter. This helps in achieving good measurements at low flows. For condensable gases, such as steam, it also helps to reduce the amount of condensate that might otherwise be present at the flowmeter.

58.5.2.3 Orientation

In general, meter orientation is not an issue for vortex flowmeters, particularly for vertical pipe installations. However, for meters having electronics at the flow tube, it is recommended in high-temperature horizontal pipe applications that the flow tube be oriented with the electronics beneath the meter. Although vortex flowmeters are not recommended for multiphase applications, they do operate with somewhat degraded performance with dirty fluids (i.e., small amounts of gas bubbles in liquid, solid particles in liquid, or liquid

droplets in gas). The degree of degradation in horizontal pipe applications depends to some extent on the specific meter design. Orienting the flow tube according to manufacturer's recommendations for the dirty fluid in question can help to alleviate this problem.

58.5.2.4 Pressure and Temperature Taps

The placement of pressure and temperature taps for determining gas densities, if required, is also an important consideration. Recommendations for location of these taps vary, depending on the manufacturer. The temperature probe is inserted typically 6 PD downstream of the flow tube. This prevents any flow disturbance in the meter and at the same time gets the probe as close to the meter as possible. The pressure tap is made typically 4 PD downstream of the meter. Although a pressure tap does not significantly affect the flow, its placement is critical for achieving an accurate density measurement.

58.5.2.5 Process Conditions

Flashing and cavitation can occur in a liquid application just downstream of the shedder if the pressure drop across the meter results in the downstream pressure being below the vapor pressure of the liquid. These phenomena lead to undefined measurement errors and possibly to structural damage and hence should be avoided. This is usually accomplished by increasing the inlet pressure or inserting a back-pressure valve downstream of the meter. To avoid flashing and cavitation, the downstream pressure after recovery (approximately 5 PD downstream) must be equal to or greater than P_{dmin}, where

$$P_{dmin} = c_1 \times \Delta P + c_2 \times P_{vap} \tag{58.12}$$

where

P_{dmin} is the minimum absolute downstream pressure after recovery
P_{vap} is the vapor pressure of the liquid at the flowing temperature
ΔP is the overall pressure drop
c_1, c_2 is the empirical constants for a specific meter (normally available from the meter manufacturer)

Pulsating flow can also in some circumstances lead to measurement errors. It is best to avoid placing the meter in process lines where noticeable pulsation exists.

58.5.3 Meter Configuration

It is important when installing an analog or intelligent vortex flowmeter that it be configured for the specific application (see Section 58.4). This is often done by the supplier prior to shipping if the user supplies the relevant information at the time the order is placed. If this is not the case, the user must carry out the configuration procedures provided by the manufacturer.

58.6 Recent Developments

Recent efforts have been made to make the vortex flowmeter into a real-time mass flow measurement device. As was demonstrated in Section 58.2, the output of the meter, based on the frequency of vortex shedding, is related to actual volumetric flow (see Equation 58.7). In intelligent transmitters, the flowing density (the density at flowing conditions) and the base density can be entered into the transmitter's database. Based on these values, mass flow or standard volumetric flow can be computed (see Equations 58.8 and 58.9). This procedure is valid if the flowing density does not vary in time. If this is not the case, an on-line, real-time measure of the density must be provided. Two different approaches have been used. One (multisensor) employs sensors in addition to the vortex sensor; the other (single sensor) relies on additional information being extracted from the vortex-shedding signal.

58.6.1 Multisensor

In this method, temperature and pressure measurements are made in addition to the vortex frequency. This approach is similar to that used in orifice-d/p mass flowmetering, in which case temperature and pressure ports are located in the pipe normally downstream of the orifice plate. However, for the multisensor vortex, the temperature and pressure sensors are incorporated into the flowmeter rather than located in the adjacent piping. Using these two additional measurements, the flowing density is calculated from the equation of state for the process fluid.

58.6.2 Single Sensor

This approach takes advantage of the fact that, in principle, for a force- or pressure-based vortex-shedding sensor, the amplitude of the vortex-shedding signal is directly proportional to the density times the square of the fluid velocity; that is,

$$\text{Signal amplitude} \propto \rho_f \times U^2 \tag{58.13}$$

The fluid velocity can be determined from the vortex frequency; that is,

$$\text{Frequency} \propto U \tag{58.14}$$

Hence,

$$\frac{\text{Signal amplitude}}{\text{Frequency}} \propto \rho_f \times U \propto \text{Mass flow} \tag{58.15}$$

This approach, in principle, is independent of the process fluid and requires no additional sensors.

Further Information

ASME MFC-6M, *Measurement of Fluid Flow in Closed Conduits Using Vortex Flowmeters*, Washington, DC: American Society of Mechanical Engineers, 1998.

DeCarlo, J. P., *Fundamentals of Flow Measurement*, Research Triangle Park, NC: Instrument Society of America, 1984, Chapter 8.

Gotthardt, W. C., Oscillatory flowmeters, in *Practical Guides for Measurement and Control: Flow Measurement*, D. W. Spitzer (ed.), Research Triangle Park, NC: Instrument Society of America, 1991, Chapter 12.

Miller, R. W., *Flow Measurement Engineering Handbook*, 3rd edn., New York: McGraw-Hill, 1996, Chapter 14.

Vortex Flowmeters, http://saba.kntu.ac.ir/eecd/Ecourses/instrumentation/projects/reports/Flowmeter/vortex.htm (accessed on July 12, 2012).

Vortex Shedding Flowmeter, UFM, http://www.flowmeters.com/ufm/index.cfm?task=vortex_fluidic, (accessed on July 11, 2012).

Partial List of Vendors and Manufacturers

BDC, Inc., 436 Anglum Road Hazelwood, MO 63042, Toll Free: 800-432-5810, Phone: 314-993-5810, Fax: 314-993-4107, www.gobdc.com

Ives Instrument Corp., 601 Croton Rd., King of Prussia, PA 19406-3111, Phone: 610-265-3071, 877-768-1600 (toll free), Fax: 610-768-1605, www.ivesequipment.com

Malema Flow Sensors, 1060 S. Rogers Cir., Boca Raton, FL 33487-2815, Phone: 561-702-7273, www.malema.com

Sierra Instruments, 5 Harris Ct., Bldg. L, Monterey, CA 93940, Phone: 831-373-0200, 800-866-0200 (toll free), Fax: 831-373-4402, www.sierrainstruments.com

Vortex Instruments, 8475 W I-25 Frontage Rd., Suite 300, Longmont, CO 80504, Phone: 866-901-4141 (toll free), Fax: 303-682-4368, vortekinst.rtrk.com/?scid=2127745&rl_alt…

Yokogawa Filed Instruments Division, 2 Dart Rd., Newnan, GA 30265, Phone: 770-254-0400, Fax: 770-254-4913, www.yokogawa.com/us/ia/fieldinstruments/…



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

59

Thermal Anemometry

Jugal K. Agarwal
TSI, Inc.

John G. Olin
Sierra Instruments, Inc.

59.1 General Description **59**-1
Commercial Thermal Anemometers • Research Thermal Anemometers
59.2 Principle of Operation **59**-5
First Law of Thermodynamics • Temperature Compensation • Flow Calibration
59.3 Measurements **59**-13
Point Velocity • Turbulence • Thermal Anemometer Instruments
References **59**-17

59.1 General Description

A thermal anemometer measures the velocity at a point in a flowing fluid—a gas or a liquid. Figure 59.1 shows a typical thermal anemometer used to monitor velocity in gas flows. It has two sensors—a velocity sensor and a temperature sensor—which automatically correct for changes in gas temperature. Both sensors are reference-grade platinum resistance temperature detectors (RTDs). The electric resistance of RTDs increases as temperature increases. For this reason, they are one of the most commonly used sensors for accurate temperature measurements. The electronics circuit passes current through the velocity sensor, thereby heating it to a constant-temperature differential ($T_v - T_a$) above the gas temperature T_a, and measures the heat q_c carried away by the cooler gas as it flows past the sensor. Hence, it is called a "constant-temperature thermal anemometer."

Since the heat is carried away by the gas molecules, the heated sensor directly measures gas mass velocity (mass flow rate per unit area), ρU. The mass velocity is typically expressed as U_s in engineering units of standard meters per second, or standard $m \cdot s^{-1}$, referenced to standard conditions of 20°C temperature and 1 atm pressure. If the fluid's temperature and the fluid's pressure are constant, then the anemometer's measurement can be expressed as actual meters per second, or $m \cdot s^{-1}$. When the average mass velocity is multiplied by the cross-sectional area of a flow channel, the mass flow rate through the channel is obtained. Mass flow rate, rather than volumetric flow rate, is the direct quantity of interest in most applications, such as chemical reactions, combustion, heating, cooling, drying, mixing, fluid power, human respiration, meteorology, and natural convection.

The thermal anemometer has some advantages and disadvantages when compared with the two other common point-velocity instruments—Pitot tubes and laser-Doppler anemometers. Compared with Pitot tubes, the thermal anemometer measures lower velocities, has a much wider dynamic range, and has better frequency response (i.e., it responds well to changing flows) but may cost slightly more. When thermal anemometers are compared with laser-Doppler anemometers, they have a much lower cost, do not require seeding the flow with particles, and can have a higher signal-to-noise ratio. On the other hand, laser-Doppler anemometers provide a fundamental measurement of velocity, independent of temperature and fluid properties. For this reason, they are often used to calibrate thermal anemometers.

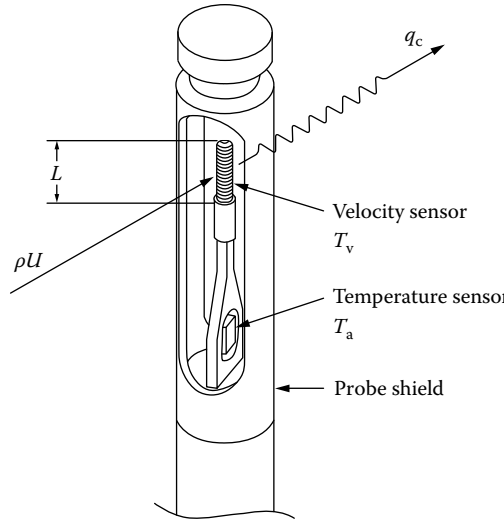


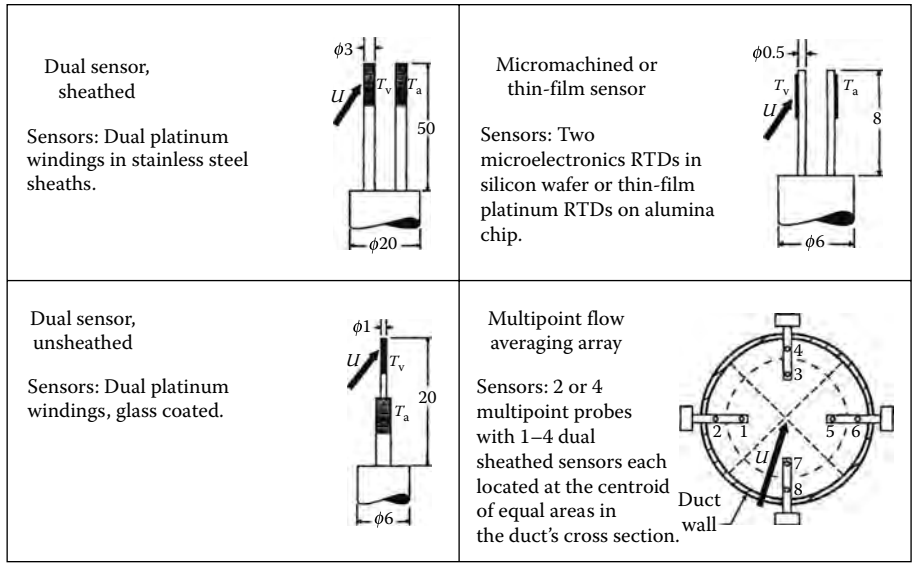
FIGURE 59.1 Principle of operation of a typical thermal anemometer. T_v is the temperature of the heated velocity sensor, T_a is the gas temperature measured by the temperature sensor, ρ is the gas mass density, U is the gas velocity, q_c is the heat carried away by the flowing gas stream, and L is the length of the heated tip of the velocity sensor. (With permission of TSI Incorporated, Shoreview, MN.)

Thermal anemometers are subdivided into three categories: heavy duty for industrial applications, medium-duty battery powered for commercial (building) applications, and research grade (delicate and very small) designed for fluctuating flow applications. Small thermal anemometers are also used in many medical devices. Figure 59.2 shows typical sensors of thermal anemometers. Industrial thermal anemometers are described in Chapter 63. Commercial and research thermal anemometers are discussed in this section.

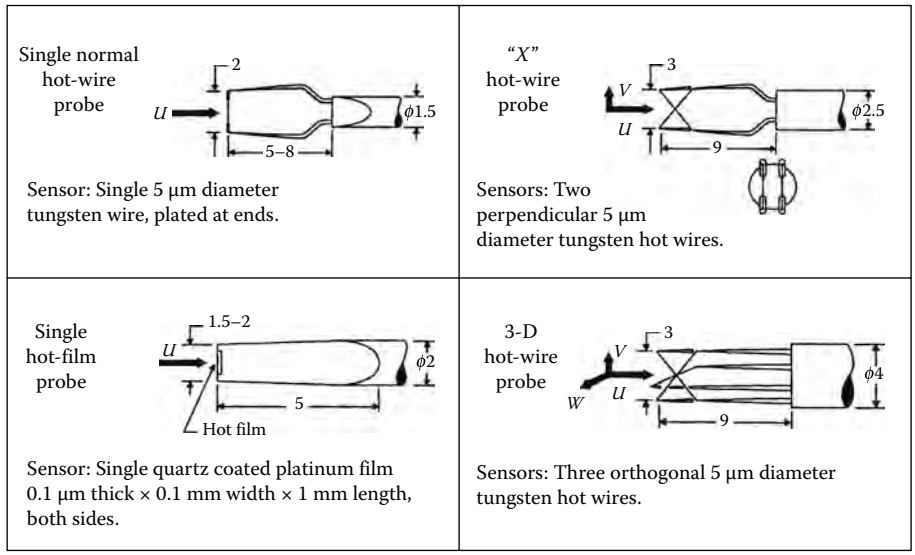
59.1.1 Commercial Thermal Anemometers

Thermal anemometers are used in many commercial applications to measure point velocity. Total flow through a duct or pipe can be calculated from a single-point sensor located at the center if the flow profile is well defined. When the flow profile is not well defined, multiple single-point sensors located at various points in the cross section are used. In general, thermal anemometers are accurate at very low velocities, have very large dynamic range, have very good frequency response, and maintain their accuracy for many years without the need for any recalibration or adjustments. Due to contamination problems, it is advisable to use thermal anemometers to measure air or gases that are clean enough to breathe. Thermal anemometers are not suitable to measure or monitor dirty unfiltered air.

In modern buildings, conditioned (heated or cooled) filtered air is supplied for human comfort and safety. Because of their ability to measure very low velocity, thermal anemometers have found many applications in the heating, ventilating, and air conditioning (HVAC) industry. Often, multipoint thermal anemometer arrays are used to monitor fresh air that is drawn into a building. Since conditioning the air is expensive, it is economical to bring in only the amount of fresh air that is needed and recirculate the remaining air from the occupied space. The air velocities are typically kept low at the fresh air intake to reduce noise; hence, multipoint thermal anemometer arrays are ideal for this application. Single-point installed thermal anemometers are used to monitor supply air for many processes such as drying of pills in the pharmaceutical industry.



(a)



(b)

FIGURE 59.2 (a) Typical commercial and (b) research thermal anemometer sensors. All dimensions are in millimeters. T_v indicates the heated velocity sensor, T_a indicates the temperature sensor, U is the major velocity component in the x direction, V is the transverse velocity component in the y direction, and W is the transverse velocity component in the z direction.

Battery-operated portable thermal anemometers are used to measure face velocities of devices like fume hoods, biological safety cabinets, and HEPA filters. In such applications, the velocity of interest is very low (0.5 $m \cdot s^{-1}$ or 100 fpm). Other point-velocity measurement techniques such as Pitot probe anemometry are not accurate at these low velocities. Sometimes, thermal anemometers are incorporated into flow measuring devices such as capture hoods, which are used to measure the total airflow exiting or entering an HVAC ducting system. Thermal anemometers are also used to measure air velocities inside an HVAC ducting system.

Another interesting application of thermal anemometry is the measurement of the very low differential pressure between two adjacent spaces, such as a hospital room and the corridor. The two adjacent spaces are connected through a tube. The differential pressure between the two spaces creates a flow. A thermal anemometer measures this flow velocity. This velocity is indicative of the differential pressure between the two adjacent spaces. Normally the pressure differential in such applications is too low to be measured by conventional differential pressure transducers.

Because of their good frequency response, thermal anemometers are widely used in medical applications to measure air, oxygen, nitrous oxide, or other gases that may be given to a patient. In particular, thermal anemometers are used to monitor air and oxygen flow inside ventilators that are used to assist the breathing of a patient.

The construction of thermal anemometers depends on their applications. All thermal anemometers consist of a velocity sensor and a temperature sensor. The velocity sensor is heated to several degrees centigrade higher than the fluid temperature. The resulting temperature difference is often referred as "overheat." The temperature sensor is not heated and measures the fluid temperature. Higher overheat is preferred for better temperature compensation. Since higher overheat increases power consumption, it is less desirable for battery-operated instruments.

Heated thermistors have been used for multipoint applications, but platinum RTDs are more common for both velocity and temperature sensors. Typically, the velocity sensors are in the 10–20 Ω range, whereas the temperature sensors are in the 100–1000 Ω range. Portable air velocity meters typically use sensors manufactured by winding a platinum wire around a ceramic mandrel. Such wire-wound sensors are normally covered by a high-temperature glass coating. The sensors used in medical applications are made by depositing a thin film of platinum onto a very small diameter glass cylinder, about 50 μm in diameter and 4 mm long.

59.1.2 Research Thermal Anemometers

Research thermal anemometers are used to measure the point velocity and/or turbulence of clean gases or liquids for fundamental fluid flow and heat-transfer research and related product development. Because of their more fragile nature, they are not suitable for commercial applications. Typically, the gas is ambient air. Constant-temperature, filtered, and degasified water is the primary liquid application, but the technique has also been applied to clean hydrocarbon liquids. As shown in Figure 59.2, the research anemometer's velocity sensor is either a hot wire or a hot film. Hot-wire sensors have a high-frequency response and, therefore, are excellent for turbulence measurements in air and other gases. They are seldom used in liquid flows because they are susceptible to fouling and contamination. Hot-film sensors trade off lower frequency response for increased ruggedness and are used in gas and liquid flows. For liquid flows, hot-film sensors are designed to shed lint and other fouling or contaminating materials. Bruun [1] is an excellent reference source for the theory and applications of hot-wire and hot-film anemometers. Another comprehensive source is Fingerson and Freymuth [2].

Typical applications for hot-wire and hot-film anemometers include 1-D, 2-D, and 3-D flow and turbulence studies; validation of computational fluid dynamics codes; environmental and micrometeorological measurements; turbomachinery; internal combustion engines; biological studies; heat-transfer research; boundary-layer measurements; supersonic flows; two-phase flows; and vorticity measurements. Freymuth [3] describes the 80-year history of research thermal anemometers. Laser-Doppler anemometry and particle image velocimetry are also widely used for these applications.

The typical hot-wire sensor is a fine platinum or tungsten wire welded at each end to miniature prongs designed to minimize their influence on the wire's flow field. A hot-film sensor is made by coating a glass rod with a platinum film. The hot-film sensors are also attached to miniature prongs. A short length at each end of the hot-film sensor is usually gold plated to define an active sensor length away from the two prongs. For work in water, the film is quartz coated to prevent electrolysis or electrical shorting, but in this case, cracking of the coating can occur. A typical tungsten wire has a diameter of 4–5 μm, an active

length of 1–3 mm, and an electrical resistance of 2–6 Ω. Because it oxidizes above 350 °C in air, tungsten hot wires are usually operated at a temperature not exceeding 300 °C. Platinum, 90% platinum + 10% rhodium and 80% platinum + 20% iridium, wires also are used. They can be soldered onto the prongs but are weaker than tungsten. In cases where the fluid temperature T_a changes enough to cause measurement errors, a separate sensor is used to measure T_a and make temperature corrections. The temperature sensor is either a hot-wire- or larger-wire-wound RTD mounted either on a separate probe or integrally on the same probe stem as the velocity sensor. As shown in Figure 59.2, for 2-D or 3-D flow studies, probes with two perpendicular wires in an "X" pattern or three orthogonal wires are used, respectively. Special subminiature probes and probes with the prongs displaced from the probe stem are used for near-wall, boundary-layer work and small flow passages. Gibbings et al. [4,5] describe hot-wire probes for use in near-wall, turbulent boundary-layer studies.

As shown in Figure 59.2, the typical hot-film sensor for water applications is a wedge-tipped or cone-tipped quartz rod with a thin 0.1 μm thick platinum film plated on its tip via cathode sputtering. The platinum film usually is coated with a 1–2 μm layer of quartz for protection and to avoid electrical shorting or electrolysis in water flows. Because hot-film sensors have a much larger mass than hot-wire sensors, their frequency response is not as good as hot wires; hence, they are not quite as applicable to high-frequency turbulence measurements. It also has been observed by Mikulla [6] that the shape of some hot-film sensors can suppress response to the turbulent velocity component normal to its surface. On the other hand, hot-film sensors have less breakage and a more stable geometry than hot-wire sensors. Other configurations of hot-film sensors include cylindrical quartz rods (approximately 50–150 μm in diameter), one or more split-film cylindrical sensors for multidimensional measurements, and flush-mounted sensors for wall heat-transfer measurements.

59.2 Principle of Operation

59.2.1 First Law of Thermodynamics

Figure 59.3 shows the first law of thermodynamics applied to a control volume consisting of the velocity sensor of a thermal anemometer, such as shown in Figure 59.1. The application of the first law to thermal anemometer sensors provides the basis for determining point velocity. Applied to Figure 59.3, the first law states that the energy into the control volume equals the energy out plus the energy stored.

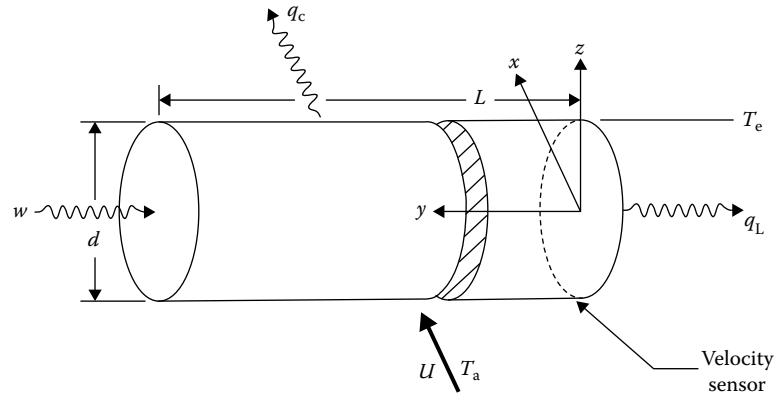


FIGURE 59.3 First law of thermodynamics applied to a thermal anemometer velocity sensor. The term w is the electric power (W) supplied to the sensor, q_c is the heat convected away from the sensor by the flowing fluid having a velocity U and temperature T_a, q_L is the conductive heat lost, T_e is the average surface temperature of the sensor over its length L, and d is the sensor's outside diameter.

Making the practical simplifying assumptions of steady-state operation (i.e., no energy stored) and no heat transfer via radiation, one obtains

$$w = q_c + q_L \tag{59.1}$$

The heat transfer q_c due to natural and forced convection normally is expressed in terms of the heat-transfer coefficient h as

$$q_c = hA_v(T_e - T_a) \tag{59.2}$$

where $A_v = \pi dL$ is the external surface area of the velocity sensor. The electric power w usually is expressed as

$$w = \frac{E_v^2}{R_v} \tag{59.3}$$

where

E_v is the voltage across the sensor
R_v is its electric resistance

For the velocity sensor shown in Figure 59.1, q_L is the heat conducted from the end of the heated velocity sensor of length L to the remainder of the sensor's length. Most of this heat is convected away by the flowing fluid, and a small fraction is conducted to the probe stem. In the case of research hot-wire or cylindrical hot-film sensors, q_L is conducted to the two prongs, of which a major fraction is convected away and a minor fraction enters the probe stem. In well-designed velocity sensors, q_L is at most 10%–15% of w, a fraction that decreases as velocity increases.

For research hot-wire sensors, the surface temperature T_e is identical to the wire temperature T_v. However, the surface temperature T_e of other sensors with some kind of coating is slightly less than the temperature T_v of the platinum because a temperature drop is required to pass the heat q_c through the intervening "skin." This is expressed as

$$T_e = T_v - q_c R_s \tag{59.4}$$

where R_s is the thermal skin resistance in units of kW^{-1}. R_s is a constant for a given sensor and is the sum of the thermal resistances of the coating layers. For research anemometer velocity sensors, $R_s = 0$ and $T_e = T_v$ in Equation 59.4. As evidenced by Equation 59.4, the effect of skin resistance increases as velocity (i.e., q_c) increases. The effect is almost negligible at low velocity; but at high velocity, it is responsible for the characteristic droop in power vs. velocity curves.

Because the flow around cylinders in cross flow is confounded by the boundary-layer separation and vortex shedding, it has defied analytical solution. Therefore, the film coefficient h in Equation 59.2 is found using empirical correlations. Correlations for h are expressed in terms of the following nondimensional parameters:

$$\mathrm{Nu} = \Im(\mathrm{Re}, \mathrm{Pr}, \mathrm{Gr}, \mathrm{M}, \mathrm{Kn}) \tag{59.5}$$

where

Nu = hd/k is the Nusselt number (the heat-transfer parameter)
Re = $\rho Vd/\mu$ is the Reynolds number (the ratio of dynamic to viscous forces)
Pr = $\mu C_p/k$ is the Prandtl number (the fluid properties parameter)
M is the Mach number (the gas compressibility parameter)
Kn is the Knudsen number (the ratio of the gas mean free path to d)

In the aforementioned, d is the outside diameter of the sensor, ρ is the fluid density, V is average velocity, k is the fluid's thermal conductivity, μ is its viscosity, and C_p is its coefficient of specific heat at constant pressure. If one takes the practical case where (1) natural convection is embodied in Re and Pr, (2) the velocity is less than one-third the fluid's speed of sound (i.e., <100 $m \cdot s^{-1}$ in ambient air), and (3) the flow is not in a high vacuum, then one can ignore the effects of Gr, M, and Kn, respectively. Thus,

$$\mathrm{Nu} = \Im(\mathrm{Re}, \mathrm{Pr}) \tag{59.6}$$

Over the years, many attempts have been made to find universal correlations for the heat transfer from cylinders in cross flow. For an isothermal fluid at constant pressure, King [7] expresses Equation 59.6 as

$$\mathrm{Nu} = A + B\,\mathrm{Re}^{0.5} \tag{59.7}$$

where A and B are empirical calibration constants that are different for each fluid and each temperature and pressure. Kramers [8] suggests the following correlation:

$$\mathrm{Nu} = 0.42\,\mathrm{Pr}^{0.2} + 0.57\,\mathrm{Pr}^{0.33}\,\mathrm{Re}^{0.50} \tag{59.8}$$

This correlation accounts for the variation in fluid properties (k, μ, and Pr) with temperature. Kramers [8] evaluates these properties at the so-called "film" temperature $(T_v + T_a)/2$, rather than at T_a itself. Another comprehensive correlation is given by Churchill and Bernstein [9]. Several other correlations are similar to Equation 59.8 but have exponents for the Reynolds number ranging from 0.4 to 0.6. Others have 0.36 and 0.38 for the exponent of the Prandtl number. Equations 59.7 and 59.8 are strictly valid only for hot-wire sensors with very high L/d ratios, in which case q_L and R_s are zero. The following universal correlation is suggested for real-world velocity sensors with variable fluid temperature and nonzero q_L and R_s:

$$\mathrm{Nu} = A + B\,\mathrm{Pr}^{0.33}\,\mathrm{Re}^{n} \tag{59.9}$$

where constants A, B, and n are determined via flow calibration. Equation 59.9 is applicable to most commercial, industrial, and research velocity sensors.

Combining Equations 59.1 through 59.3, and 59.9, and recognizing that $h = kNu/d$, one obtains

$$\frac{E_v^2}{R_v} = (Ak + Bk\,\mathrm{Pr}^{0.33}\,\mathrm{Re}^{n})(T_v - T_a) \tag{59.10}$$

where A and B are new constants. A, B, and n are determined via flow calibration and account for all nonidealities, including end conduction and skin resistance. Equation 59.10 is applicable to most commercial, industrial, and research velocity sensors. The manufacturers of industrial thermal anemometers can add other calibration constants to Equation 59.10 to enhance its correlation with flow-calibration data. The presence of end conduction means that the temperature of the velocity sensor varies with the axial coordinate y in Figure 59.3. The temperature T actually sensed by the velocity sensor is the average temperature over length L, or

$$T_v = \left(\frac{1}{L}\right)\int_0^L T_v(y)\,\mathrm{d}y \tag{59.11}$$

Bruun [1] presents an analytical solution for $T_v(y)$ for hot-wire sensors. Equation 59.11 is the correct expression for T_v in Equation 59.10 and is so defined hereafter.

For fluid temperatures less than 200 °C, the electric resistance of the RTD velocity and temperature sensors is usually expressed as

$$R_v = R_{v0}\left[1 + \alpha_v (T_v - T_0)\right] \tag{59.12}$$

$$R_T = R_{T0}\left[1 + \alpha_T (T_T - T_0)\right] \tag{59.13}$$

where R_{v0} and R_{T0} are, respectively, the electric resistances of the velocity sensor and the temperature sensor at temperature T_0 (usually 0 °C or 20 °C), and α_v and α_T are the temperature coefficients of resistivity at temperature T_0. Additional terms are added to Equations 59.12 and 59.13 when fluid temperatures exceed 200 °C. When evaluated at the fluid temperature T_a, the resistance R_a of the velocity sensor is

$$R_a = R_{v0}\left[1 + \alpha_v (T_a - T_0)\right] \tag{59.14}$$

For applications with wide excursions in fluid temperature, additional terms are added to Equations 59.12 through 59.14. At 20 °C, α_v and α_T are approximately 0.0036 °C^{-1} for tungsten wire, 0.0038 °C^{-1} for pure platinum wire, 0.0016 °C^{-1} for 90% platinum + 10% rhodium wire, 0.0024 °C^{-1} for platinum film, and 0.0040 °C^{-1} for tungsten film. R_v and R_a are called the "hot" and "cold" resistances of the velocity sensor, respectively. The ratio R_v/R_a is called the "overheat ratio." For gas flows, commercial thermal velocity sensors are operated at overheat ratios from 1.1 to 1.4 ($T_v - T_a$ = 30 °C–100 °C). For gas flows, the overheat ratio of tungsten hot-wire and hot-film sensors is usually set to approximately 1.8 ($T_v - T_a$ = 200 °C–300 °C) and 1.4 ($T_v - T_a$ = 150 °C–200 °C), respectively. For water flows, the overheat ratio of hot-film sensors is approximately 1.05–1.10 ($T_v - T_a$ = 10 °C–20 °C). Mikulla [6] shows the importance of the effect of overheat ratio on frequency response.

Combining Equations 59.12 and 59.14, one obtains

$$T_v - T_a = \frac{R_v - R_a}{\alpha_v R_{v0}} \tag{59.15}$$

Inserting this into Equation 59.10, one obtains

$$\frac{E_v^2}{R_v (R_v - R_a)} = Ak + Bk\mathrm{Pr}^{0.33}\mathrm{Re}^n \tag{59.16}$$

where new constants A and B have absorbed the constants α_v and R_{v0}.

Figures 59.4 through 59.6 show three typical electronic drives for thermal anemometer sensors. Figure 59.4 shows the commonly used constant-temperature anemometer Wheatstone bridge circuit described by Takagi [10]. Figure 59.5 is similar but is controlled and operated via a personal computer (PC). In the constant-temperature mode, the hot resistance R_v, and hence the velocity sensor's temperature, remains virtually constant, independent of changes in velocity. With the addition of the temperature sensor shown in Figure 59.4, the bridge circuit also compensates for variations in fluid temperature T_a, as described later. Another common analog sensor drive is the constant-current anemometer. In this mode, a constant current is passed through the velocity sensor, and the sensor's temperature decreases as the velocity increases. Because the entire mass of the sensor must participate in this temperature change, the sensor is slower in responding to changes in velocity. Because the constant-temperature anemometer has a flatter frequency response and an excellent signal-to-noise ratio [2] and is easier to use, it is favored over constant-current anemometers by most researchers and manufacturers for velocity and turbulence measurements. The constant-current anemometer with a very low overheat ratio is often used as the temperature sensor. Subsequently, references made herein to sensor electronics will be based on the constant-temperature anemometer.

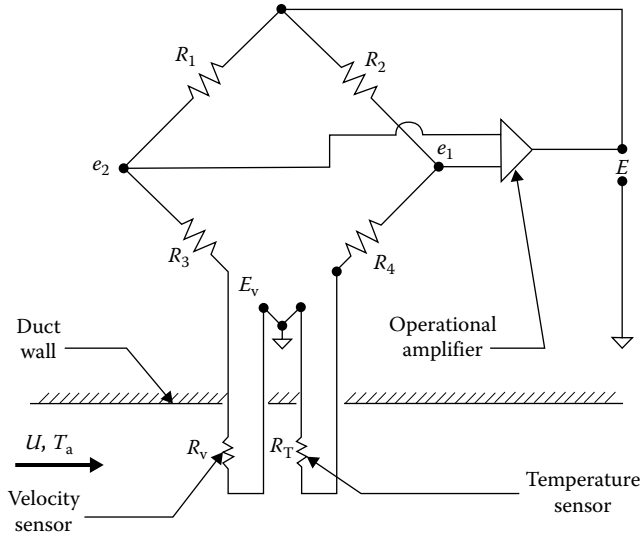


FIGURE 59.4 Constant-temperature thermal anemometer bridge circuit with automatic temperature compensation. R_1, R_2, and R_4 are fixed resistors selected to achieve temperature compensation, R_3 is the probe and cable resistance, R_v is the velocity sensor's resistance, R_T is the temperature sensor's resistance, and E is the bridge voltage output signal. For research anemometers operating in isothermal flows, the temperature sensor is eliminated and replaced with a variable bridge resistor. Some temperature-compensation circuits have an additional resistor in parallel with R_T.

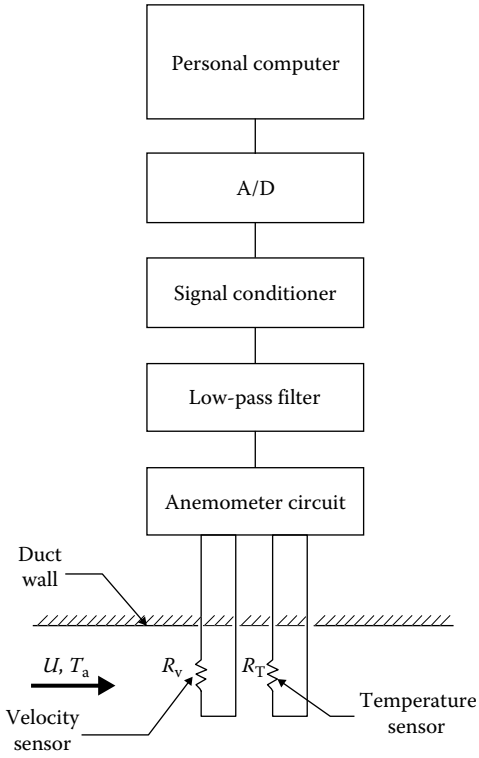


FIGURE 59.5 PC-based digital thermal anemometer system. The signal conditioner matches the anemometer circuit's output to the ADC. For isothermal flows, the temperature sensor is eliminated.

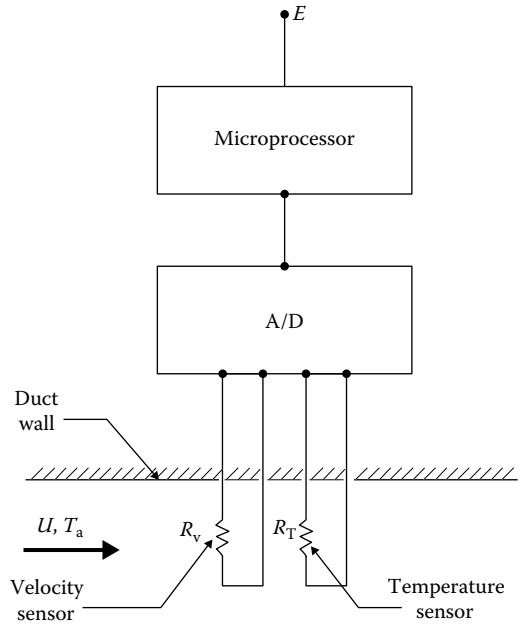


FIGURE 59.6 Microprocessor-based digital thermal anemometer. This system digitally maintains a constant-temperature difference ($T_v - T_T$) and automatically corrects for the variation in fluid properties with temperature. The manufacturer provides a probe-mounted electronics package delivering an analog output signal E and/or a digital RS485 signal linearly proportional to gas mass velocity. (With permission of Sierra Instruments, Inc., Monterey, CA.)

In the constant-temperature anemometer drive shown in Figure 59.4, the resistances R_1 and R_2 are chosen to (1) maximize the current on the velocity-sensor side of the bridge so it becomes self-heated and (2) minimize the current on the temperature-sensor side of the bridge so it is not self-heated and is independent of velocity. Additionally, the temperature sensor must be sufficiently large in size to avoid self-heating. The ratio R_2/R_1 is called the "bridge ratio." A bridge ratio of 5:1–20:1 is normally used, but for optimum frequency response and compensation for long cable length, a bridge ratio of 1:1 can be used. In Figure 59.4, the operational amplifier, in a feedback control loop, senses the error voltage ($e_2 - e_1$,) and feeds the exact amount of current to the top of the bridge necessary to make ($e_2 - e_1$,) approach zero. In this condition, the bridge is balanced; that is,

$$\frac{R_1}{R_v + R_3} = \frac{R_2}{R_T + R_4} \tag{59.17}$$

or

$$R_v = \frac{R_1}{R_2}(R_T + R_4) - R_3 \tag{59.18}$$

From Equation 59.18, one sees that R_v is a linear function of R_T. This relationship forms the basis for analog temperature compensation.

Expressing the voltage E_v across the velocity sensor in terms of the bridge voltage E, one obtains

$$E_v = \frac{ER_v}{R_1 + R_3 + R_v} \tag{59.19}$$

Inserting this into Equation 59.16, one arrives at the generalized expression for the first law of thermodynamics for the thermal anemometer velocity sensor:

$$E^2 = G\left[Ak + Bk\left(\frac{\rho_s}{\mu}\right)^n \mathrm{Pr}^{0.33} U_s^n\right] \tag{59.20}$$

where $G = (R_1 + R_3 + R_v)^2 (R_v - R_a)/R_v$ and where A and B again are new constants. In Equation 59.20, one recognizes that conservation-of-mass considerations require that $\rho U = \rho_s U_s$, where ρ and U are referenced to the actual fluid temperature and pressure and ρ_s and U_s are referenced to standard conditions of 20 °C temperature and 1 atm pressure. To write Equation 59.20 in terms of U, one simply replaces ρ_s by ρ and U_s by U.

59.2.2 Temperature Compensation

The objective of temperature compensation is to make the bridge voltage E in Equation 59.20 independent of changes in the fluid temperature T_a. This is accomplished if (1) the term G in Equation 59.20 is independent of T_a and (2) compensation is made for the change in fluid properties (k, μ, and Pr) with T_a. Since these fluid properties have weaker temperature dependence than G in Equation 59.20, for small temperature changes (less than ±10 °C) in gas flows, only G requires compensation.

The two-temperature method is a typical procedure for compensating for both G and fluid properties. In this method, fixed-bridge resistors R_1, R_2, and R_4 in Figure 59.4 are selected so that E is identical at two different temperatures but at the same mass flow rate. This procedure is accomplished with flow calibration at two different temperatures and has variations among manufacturers.

The two-temperature method adequately compensates for temperature variations less than approximately ±50 °C. In higher temperature gas flow applications, such as the flow of preheated combustion air and stack gas, temperature variations typically are higher. The microprocessor-based digital sensor drive in Figure 59.6 provides temperature compensation for temperature variations ranging from ±50 °C to ±150 °C. This sensor drive has no analog bridge. Instead, it has a virtual digital bridge that maintains $(T_v - T_a)$ constant within 0.1 °C and has algorithms that automatically compensate for temperature variations in k, m, and Pr. For this digital sensor drive, the first law of thermodynamics is found from Equation 59.10 as

$$w = \left[Ak + Bk\left(\frac{\rho_s}{\mu}\right)^n \mathrm{Pr}^{0.33} U_s^n\right]\Delta T \tag{59.21}$$

where $\Delta T = (T_v - T_a)$ is now a known constant.

59.2.3 Flow Calibration

Figure 59.7 shows a typical flow-calibration curve for the digital electronics drive shown in Figure 59.4. The curve is nonlinear in nature. In the past, this inherent nonlinearity was considered a disadvantage because it required linearization circuitry that was expensive to implement using analog circuits. Now, digital circuits are universally used in instrumentation, and the nonlinearity is no longer a disadvantage. The steep curve at low velocities is advantageous because it provides excellent resolution at low velocities. This increases the dynamic range up to 1000:1 for a single sensor. Additionally, the high-level output of several volts provides excellent repeatability and requires no amplification other than that for spanning. Since the critical dimensions of thermal anemometer sensors are small, current manufacturing technology is incapable of maintaining tolerances sufficient to ensure sensor reproducibility.

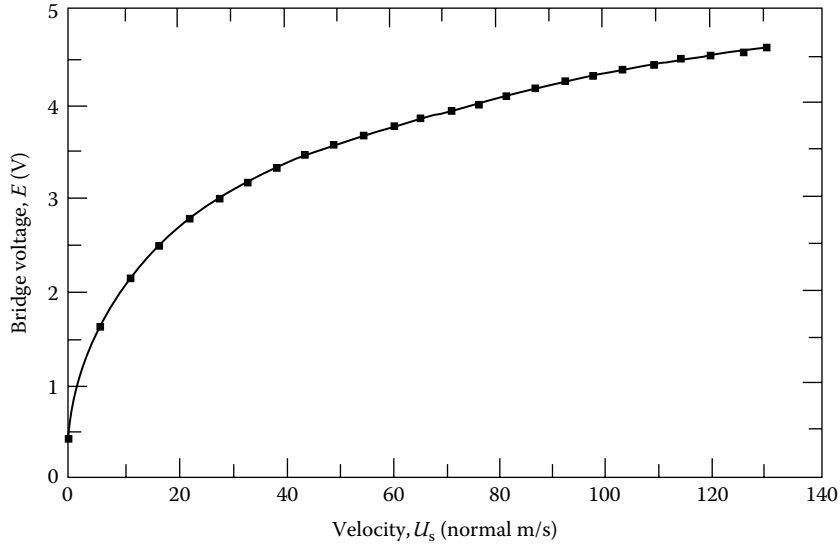


FIGURE 59.7 Typical flow-calibration curve for an industrial thermal anemometer. The electronics drive is that shown in Figure 59.6. The constant-temperature differential ($T_v - T_T$) is 50.0 °C. The cold resistances R_{v0} and R_{T0} of the velocity and temperature sensors at 20 °C are approximately 20 and 200 W, respectively. (With permission of Sierra Instruments, Inc., Monterey, CA.)

Therefore, each thermal anemometer must be calibrated by exposing it to a known flow, for example, as in Figure 59.7, over its entire velocity range, either at the exact fluid temperature of its usage or over the range of temperatures it will encounter if it is to be temperature compensated. A 10–20-point-velocity calibration is required to accurately determine the calibration constants *A*, *B*, and *n* in Equation 59.20. A least-squares curve-fitting procedure usually is applied. Bruun [1] and Gibbings et al. [4] provide more insight into curve fitting.

Proper flow calibration requires two critical elements: (1) a stable, reproducible, flow-generating facility and (2) an accurate velocity transfer standard. A typical flow-calibration facility consists of (1) a flow source such as a fan, pump, elevated tank, or compressed gas supply; (2) a flow-quieting section, such as a plenum with flow straighteners, screens, or other means to reduce swirling, turbulence, or other flow nonuniformities; (3) a nozzle to accelerate the flow and further flatten, or make uniform, the velocity profile; (4) a test section or free jet into which the thermal anemometer probe is inserted; (5) a means for holding and sealing the thermal anemometer probe; and (6) a velocity transfer standard. The test section or free jet must have a velocity profile that is uniform within approximately 0.5%–1.0% in its central portion, turbulence intensity less than about 0.5%, and an area large enough so that the projected area of the velocity probe is less than 5%–10% of the cross-sectional area. In the free-jet type of flow calibrator, the flow velocity is often determined by measuring the pressure drop across the nozzle. In a wind-tunnel type of flow calibrator, the flow velocity is determined by inserting a Pitot tube into the test section. Gibbings [4] describes a water box displacement rig for flow calibration at very low velocities in the range of 0.1–4 m s^{-1}.

Pitot tubes and laser-Doppler anemometers are the two most common velocity transfer standards used to calibrate thermal anemometers. Both have detailed descriptions earlier in this handbook. The Pitot tube usually has the classical "L" shape and an outside diameter of about 3 mm. Its tip is located in the same plane in the test section as the thermal anemometer probe and is no closer than approximately 3 cm from the wall. The measuring volume of the laser-Doppler anemometer is similarly located. The Pitot tube is far less expensive and easier to operate but is difficult to use if air velocities are less than about 3 m·s^{-1}. A Pitot-tube-based flow transfer standard should have its calibration recertified every

6 months by an accredited standards laboratory. On the other hand, the laser-Doppler anemometer is a first-principle standard that accurately measures air velocity from approximately 0.5 to 100 $m \cdot s^{-1}$. Its primary disadvantages are high expense and complications associated with properly seeding the flow with particles.

Small wind tunnels are used for calibrating thermal anemometers in manufacturing facilities and in accreditation laboratories. Such wind tunnels are calibrated using Pitot-tube and/or laser-Doppler anemometry. In operation, the wind-tunnel velocity at the test section is found by measuring the difference between the total pressure and the static pressure with a differential pressure transducer. This method is easier to use than other methods and has an excellent repeatability. Generating a uniform and stable low velocity (lower than 2.5 $m \cdot s^{-1}$) is a challenge. In one commercially available benchtop wind tunnel [11], this is accomplished by using a nozzle plate with multiple equally spaced holes downstream of the test section. The manufacturer claims uniform, stable low velocities down to 0.15 $m \cdot s^{-1}$ in the test section.

59.3 Measurements

59.3.1 Point Velocity

Based on the first law of thermodynamics expressed by Equation 59.20, one now can solve for the desired quantity—either the actual point velocity U ($m \cdot s^{-1}$) or the point mass velocity U_s (standard $m \cdot s^{-1}$). Here, one assumes that the velocity vector is normal to the flow sensor. The 2-D and 3-D velocity measurements are discussed later. In the following, A, B, and n are constants but are different for each case.

The simplest case is isothermal flow with a hot-wire sensor having a very high length-to-diameter ratio (L/d). In this case, the exponent n in Equation 59.20 is 0.5, as shown in Equation 59.7. The applicable first law and velocity expressions are

$$E^2 = A + BU^{0.5} \tag{59.22}$$

and

$$U = \left[\frac{E^2 - A}{B}\right]^2 \tag{59.23}$$

In the case of a real-world sensor in an isothermal flow having either end loss only or both end loss and skin resistance, one obtains

$$E^2 = A + BU^n \tag{59.24}$$

and

$$U = \left[\frac{E^2 - A}{B}\right]^{1/n} \tag{59.25}$$

Often, Equation 59.25 is replaced with a polynomial of the form $U = F(E^2)$, where the function $F(\ )$ is a fourth-order polynomial whose coefficients are determined from flow-calibration data using least-squares curve-fitting algorithms. For the same case as earlier, but with non-isothermal flow, the first law is expressed by (59.20), and the velocity is expressed as

$$U_s = \frac{\mu}{\rho_s}\left[\frac{E^2/G - Ak}{Bk\,\mathrm{Pr}^{0.33}}\right]^{1/n} \tag{59.26}$$

For the digital sensor drive of Figure 59.6, the first law is given by Equation 59.21 and the velocity by

$$U_s = \frac{\mu}{\rho_s}\left[\frac{w/\Delta T - Ak}{Bk\,\mathrm{Pr}^{0.33}}\right]^{1/n} \tag{59.27}$$

Current commercial thermal anemometer systems have temperature-compensation and "linearization" electronics (or lookup table in software) that automatically calculate U_s as a linear function of E or w, based on the foregoing relationships.

59.3.2 Turbulence

Turbulence measurements are the second most common application of research thermal anemometers. This measurement requires the high-frequency response of hot-wire and hot-film research anemometers operated in the constant-temperature mode. The vast majority of fluid flows are turbulent. Only flows with very low Reynolds numbers are nonturbulent, or laminar. Turbulent flows are time variant and usually are separated as follows into time-mean and fluctuating parts:

$$\begin{aligned} U(t) &= \bar{U} + u \\ V(t) &= \bar{V} + v \\ W(t) &= \bar{W} + w \\ T_a(t) &= \bar{T}_a + \theta \\ E &= \bar{E} + e \end{aligned} \tag{59.28}$$

where $U(t)$, $V(t)$, and $W(t)$ are the orthogonal components in the x, y, and z directions, respectively, such as shown in Figure 59.2 for the 3-D hot-wire probe. $T_a(t)$ is the fluid temperature, and $E(t)$ is the bridge voltage. $\bar{U}$, $\bar{W}$, $\bar{V}$, $\bar{T}_a$, and $\bar{E}$ are the time-mean parts, and $u(t)$, $v(t)$. $w(t)$, $\theta(t)$, and $e(t)$ are the time-dependent fluctuating parts. The time-mean parts are averaged sufficiently long to become independent of turbulent fluctuations, yet respond to changes with time in the main flow. In most of this text and for the purpose of simplicity, we have used the symbols U and T_a for the time-mean velocity and the time-mean temperature, respectively. In the study of turbulence, one is primarily interested in the time average of the product of two fluctuating velocity components (turbulence correlations) because these terms appear in the time-averaged Navier–Stokes equation. Two important turbulence correlations are $\overline{u^2}$ and $\overline{uv}$. The correlation $\left(\sqrt{\overline{u^2}}\right)/\bar{U}$ is called the *turbulence intensity*. The manufacturers of research anemometer systems provide electronics for automatically computing turbulence correlations.

For a fluid with changes in temperature sufficiently small that fluid properties are essentially constant, one can write Equation 59.20 in the following form:

$$E^2 = (A + BU^n)(T_v - T_a) \tag{59.29}$$

where A, B, and n are constant and where R_v is virtually constant because the anemometer is in the constant-temperature mode. Elsner [12] shows that the fluctuating voltage e is found by taking the total derivative of Equation 59.29, as follows:

$$e = S_u u + S_\theta \theta \tag{59.30}$$

where

$$S_u = \frac{\delta E}{\delta U} = \frac{nBU^{n-1}}{2}\left[\frac{(T_v - T_a)}{A + BU^n}\right]^{1/2} = \text{Velocity sensitivity} \tag{59.31}$$

$$S_\theta = \frac{\delta E}{\delta T_a} = -\frac{1}{2}\left[\frac{(A+BU)^n}{T_v - T_a}\right]^{1/2} = \text{Temperature sensitivity} \tag{59.32}$$

It is seen from Equations 59.31 and 59.32 that increasing $(T_v - T_a)$, that is, operating the sensor as hot as possible, maximizes the velocity sensitivity and minimizes the sensitivity to temperature fluctuations. This is why tungsten hot wires are operated at high temperatures (typically 200 °C–300 °C).

The fluctuating components of velocity have a broad frequency spectrum, ranging from 10^2 to 10^5 Hz and sometimes even higher. Therefore, it is imperative that the frequency response of constant-temperature research anemometers have a flat frequency response, that is, minimized attenuation and phase shift at higher frequencies. Blackwelder [13] and several other investigators have studied the frequency response of hot-wire anemometers. For turbulence measurements, Borgos [14] describes commercial research anemometer systems with features such as low-pass filters to decrease electronics noise, a sub-circuit for determining and setting overheat ratio, a square-wave generator for frequency response testing, and two or more controls to optimize the frequency response to fast fluctuations. Recent systems have electronics that compensate for frequency attenuation. When used with 5 μm diameter hot-wire sensors in air, commercial systems are capable of nearly flat frequency response and very small phase lag from 0 Hz to approximately 10^4 Hz. As reported by Nelson and Borgos [15], wedge and conical hot-film sensors in water have a relatively flat response from 0 to 10 Hz for velocities above 0.3 m s^{-1}. The compensation for attenuation to get high-frequency response is a more involved process for hot-film sensors than for hot-wire sensors.

Two- and three-component velocity and turbulence measurements are made using hot-wire or hot-film research anemometers, such as shown in Figure 59.2. As described by Müller [16], hot-wire or cylindrical hot-film probes in the "X" configuration are used to measure the *U* and *V* velocity components. In a three-sensor orthogonal array, they measure *U*, *V*, and *W*. Döbbeling et al. [17] and other investigators have developed four-wire arrays for measurement of *U*, *V*, and *W*. Olin and Kiland [18] describe an orthogonal array of three cylindrical split hot-film sensors. Each of the three sensors in this array has two individually operated hot films separated by two axial splits 180° apart along its entire length. The two split films take advantage of the nonuniform heat-transfer distribution around a cylinder in cross flow.

In multisensor arrays, the velocity vector is not necessarily normal to a cylindrical sensor. If the discussion is limited to isothermal flows, the first law expressed by Equation 59.20 becomes

$$E^2 = A + BV_e^n \tag{59.33}$$

where V_e is the effective velocity sensed by a single cylindrical sensor in the array and *A*, *B*, and *n* are constants. Jörgenson [19] describes V_e as follows:

$$V_e^2 = U_N^2 + a^2U_T^2 + b^2U_B^2 \tag{59.34}$$

where

U_N is the velocity component normal to the sensor
U_T is the tangential component
U_B is the component perpendicular to both U_N and U_T (i.e., binormal)

The constants *a* and *b* in Equation 59.34 are referred to as the sensor's yaw and pitch coefficients, respectively, and are determined via flow calibration. Typical values for *a* and *b* for a plated hot-wire sensor are 0.2 and 1.05, respectively. Inserting Equation 59.34 into Equation 59.33, we get the following expression for the output signal of a single sensor in the array:

$$E^2 = A + B\left(U_N^2 + a^2U_T + b^2U_B^2\right)^{n/2} \tag{59.35}$$

Expressions like this, or similar ones such as given by Lekakis et al. [20], are written for all sensors in the array. These expressions and the geometry of the sensor configuration are then used to solve for the components of velocity U, V, and W in the x, y, and z spatially fixed reference frame.

59.3.3 Thermal Anemometer Instruments

Table 59.1 lists examples of typical thermal anemometer instruments offered by manufacturers. Table 59.2 lists the major manufacturers. A typical portable thermal anemometer includes sensor, probe, Wheatstone bridge drive circuits, linearization circuits, memory to store calibration information, battery, and digital display. Sensors and probes have been described in the previous sections. The electronics of a portable meter is enclosed in a small plastic case with a digital display. They are powered by dry cell batteries. In addition to the digital display, they have data storage and recall capabilities.

Research thermal anemometer systems usually are coupled with a PC, as shown in Figure 59.5. The PC provides system setup and control, as well as data display and analysis. Modern systems feature low-noise circuits, together with smart bridge optimization technology that eliminates tuning and automatically provides flat frequency response up to 300,000 Hz. Lower-cost units provide flat response up to 10,000 Hz. A built-in thermocouple circuit simplifies temperature measurement. The PC's windows-based software provides near real-time displays of velocity, probability distribution, and turbulence intensity. Postprocessing gives additional statistics, including mean velocity; turbulence intensity; standard deviation; skewness; flatness; normal stress for one-, two-, and three-component probes; as well as shear stress, correlation coefficients, and flow-direction angle for 2-D and 3-D probes. In addition, power spectrum, autocorrelations, and cross correlations can be displayed. The software automatically handles flow-calibration setup and calculates calibration velocity. Systems are available in 1-, 2-, 8-, and 16-channel versions. A relatively new approach, known as constant-voltage anemometry (CVA), has been aimed at making measurements in supersonic flows. Berson et al. [21] examine the behavior of this technique in high levels of turbulence.

Research thermal anemometer systems were first introduced in the early 1960s. The commercial anemometers were introduced in late 1970s. Nowadays, portable, handheld thermal anemometers are popular with professionals responsible for measuring air velocities in airflow applications. At first, thermal anemometers were not considered sufficiently durable for the rigors of industrial use. With the advent of

TABLE 59.1 Typical Thermal Anemometers

Product Description	Approximate 2010 U.S. Price
Portable battery-operated instruments	
Basic portable air velocity meter	$300–$900
Data-logging multiparameter portable air velocity meter	$700–1,500
Installed duct monitoring systems	
Single-point air velocity probe	$300–$800
Multipoint air-duct flow monitoring system	$2,000–$10,000
Research systems	
Single-channel hot-wire or hot-film anemometer system	$25,000
Three-component hot-wire anemometer system	$35,000

Note: Prices listed are the price range quoted by the manufacturers listed in Table 59.2. All portable air velocity meters include sensor, battery, and display. The multiparameter instruments typically measure air velocity, temperature, humidity, and differential pressure and have data-logging capabilities. The single-point air velocity instruments typically have analog or digital outputs, and the required voltage is supplied by an external power supply. The multipoint systems are typically used with a control system. Research anemometer systems have standard hot-wire probes and electronics and include ambient air calibrations.

TABLE 59.2 Manufacturers of Thermal Anemometer Instruments

Portable battery-operated instruments
TSI Incorporated, 500 Cardigan Road, Shoreview, MN 55164, Tel.: 1 800 874 2811, info@tsi.com
Dwyer Instruments, Inc., P. O. BOX 373, Michigan City, IN 46361, Tel.: 219 879 8000, info@dwyer-inst.com
Extech Instruments Corp., 285 Bear Hill Road, Waltham, MA 02451, Tel.: 781 890 7440, sales@extech.com
Testo, Inc., 40 White Lake Road, Sparta, NJ 07871, Tel.: 1 800 227 0729, info@testo.com
Kanomax USA, Inc., P. O. Box 372, 219 US Hwy. 206, Andover, NJ 07821, Tel.: 1 800 247 8887, info@kanomax-usa.com
TES Electrical Electronic Corp., 7F, No 31, Lane 513, Rui Guang Road, Neihu Dist. Taiwan, Republic of China, Tel.: 866 2 27999 3660, www.tes.com.tw/indexe.htm
Single-point duct monitoring systems
TSI Incorporated, 500 Cardigan Road, Shoreview, MN 55164, Tel.: 800 874 2811, info@tsi.com
Sierra Instruments, Inc., 5 Harris Court, Building L, Monterey, CA 93940, Tel.: 800 866 0200, info@sierrainstruments.com
Dwyer Instruments, Inc., P. O. BOX 373, Michigan City, IN 46361, Tel.: 219 879 8000, info@dwyer-inst.com
E+E Elektronik, Langwiesen 7, A-4209 Engerwitzdorf, Austria, Tel.: 43 7235 605-0, info@epluse.com
Multipoint duct monitoring systems
Ebtron, Inc., 1663 Highway 701 South, Loris, SC 29569, Tel.: 843.756.1828, www.ebtron.com
Air Monitor Corporation, 1050 Hopper Avenue, Santa Rosa, CA 95406, Tel.: 919 491 2127, amcsales@airmonitor.com
Research systems
TSI Incorporated, 500 Cardigan Road, Shoreview, MN 55164, Tel.: 800 874 2811, info@tsi.com
Dantec Dynamics A/S, Tonsbakken 16–18, DK-2740 Skovlunde, Denmark, Tel.: 45 44 57 80 00, www.dantecdynamics.com

stainless-steel sheathed sensors and microprocessor-based electronics, industrial thermal anemometers now enjoy the credibility formerly attributed to only traditional flowmeter approaches (see Chapter 63). Initial research systems required a high level of user knowledge and considerable involvement in operation. In contrast, current research systems have nearly flat frequency response, high accuracy, and easy-to-use controls providing the flexibility researchers require. Research systems based on PCs have graphical user interfaces that enhance both performance and simplicity of operation.

References

1. H. H. Bruun, *Hot-Wire Anemometry: Principles and Signal Analysis*, Oxford, U.K.: Oxford University Press, 1995.
2. L. M. Fingerson and P. Freymuth, Thermal anemometers, in R. J. Goldstein (ed.), *Fluid Mechanics Measurements*, Washington, DC: Hemisphere, 1983.
3. P. Freymuth, History of thermal anemometry, in N. P. Cheremisinoff and R. Gupta (eds.), *Handbook of Fluids in Motion*, Ann Arbor, MI: Ann Arbor Science Publishers, 1983.
4. J. C. Gibbings, J. Madadnia, and A. H. Yousif, The wall correction of the hot-wire anemometer, *Flow Meas. Instrum.*, 6(2), 127–136, 1995.
5. J. C. Gibbings, J. Madadnia, S. Riley, and A. H. Yousif, The proximity hot-wire probe for measuring surface shear in air flows, *Flow Meas. Instrum.*, 6(3), 201–206, 1995.
6. V. Mikulla, The measurement of intensities and stresses of turbulence in incompressible and compressible air flow, PhD Thesis, University of Liverpool, Liverpool, U.K., 1972.
7. L. V. King, On the convection of heat from small cylinders in a stream of fluid: Determination of the convection constants of small platinum wires with application to hot-wire anemometry, *Phil. Trans. Roy. Soc.*, A214, 373–432, 1914.
8. H. Kramers, Heat transfer from spheres to flowing media, *Physica*, 12, 61–80, 1946.
9. S. W. Churchill and M. Bernstein, A correlating equation for forced convection from gases and liquids to a circular cylinder in crossflow, *J. Heat Transfer*, 99, 300–306, 1997.

10. S. Takagi, A hot-wire anemometer compensated for ambient temperature variations, *J. Phys. E.: Sci. Instrum*., 19, 739–743, 1986.
11. Calibration of the CERTIFIER air velocity calibrator using a laser doppler velocimeter (LDV), TSI Incorporated, http://www.tsi.com/uploadedFiles/Product_Information/Literature/Application_Notes/TI-130.pdf (accessed on March 15, 2013).
12. J. W. Elsner, An analysis of hot-wire sensitivity in non-isothermal flow, *Proceedings of the Dynamics Flow Conference*, Marseille, France, 1972.
13. F. Blackwelder, Hot-wire and hot-film anemometers, in R. J. Emrich (ed.), *Methods of Experimental Physics: Fluid Dynamics*, New York: Academic Press, 18A, pp. 259–314, 1981.
14. J. A. Borgos, A review of electrical testing of hot-wire and hot-film anemometers, *TSI Quart*., VI(3), 3–9, 1980.
15. E. W. Nelson and J. A. Borgos, Dynamic response of conical and wedge type hot films: Comparison of experimental and theoretical results, *TSI Quart*., IX(1), 3–10, 1983.
16. U. R. Müller, Comparison of turbulence measurements with single, X and triple hot-wire probes, *Exp. Fluids*, 13, 208–216, 1992.
17. K. Döbbeling, B. Lenze, and W. Leuckel, Four-sensor hot-wire probe measurements of the isothermal flow in a model combustion chamber with different levels of swirl, *Exp. Thermal Fluid Sci*., 5, 381–389, 1992.
18. J. G. Olin and R. B. Kiland, Split-film anemometer sensors for three-dimensional velocity-vector measurement, *Proceedings of the Symposium on Aircraft Wake Turbulence*, Seattle, WA, 1970, pp. 57–79.
19. F. E. Jörgenson, Directional sensitivity of wire and fibre-film probes, *DISA Info*., (11), 31–37, 1971.
20. I. C. Lekakis, R. J. Adrian, and B. G. Jones, Measurement of velocity vectors with orthogonal and non-orthogonal triple-sensor probes, *Exp. Fluids*, 7, 228–240, 1989.
21. A. Berson, P. Blanc-Benon, and G. Comte-Bellot, A strategy to eliminate all nonlinear effects in constant-voltage hot-wire anemometry, *Rev. Sci. Instrum*., 80, 045102, 2009, DOI: 10.1063/1.3103948

60

Coriolis Effect Mass Flowmeters

60.1 Theory of Operation **60**-1
60.2 Construction **60**-4
60.3 Advantages **60**-4
60.4 Disadvantages **60**-5
60.5 Applications **60**-5
60.6 A Look Ahead **60**-5
Bibliography **60**-6
Partial List of Vendors and Suppliers **60**-6

Jesse Yoder
Flow Research, Inc.
Automation Research Corporation

Coriolis flowmeters were developed in the 1970s to fill the need for a flowmeter that measures mass directly, as opposed to those that only measure velocity or volume. Because they are independent of changing fluid parameters, Coriolis meters have found wide application. Many velocity and volumetric meters are affected by changes in fluid pressure, temperature, viscosity, and density. Coriolis meters, on the other hand, are virtually unaffected by these types of changes. By measuring mass directly as it passes through the meter, Coriolis meters make a highly accurate measurement that is virtually independent of changing process conditions. As a result, Coriolis meters can be used on a variety of process fluids without recalibration and without compensating for parameters specific to a particular type of fluid. Coriolis flowmeters are named after Gaspard G. Coriolis (1792–1843), a French civil engineer and physicist for whom the Coriolis force is named.

Coriolis meters have become widely used in industrial environments because they have the highest accuracy of all types of flowmeters. They measure mass directly, rather than inferentially. Coriolis meters do not have moving parts like turbine and positive displacement meters, which have parts that are subject to wear. Maintenance requirements for Coriolis meters are low, and they do not require frequent calibration. Wetted parts can be made from a variety of materials to make these meters adaptable to many types of fluids. Coriolis meters can handle corrosive fluids and fluids that contain solids or particulate matter. While these meters were used mainly for liquids when they were first introduced, they have since become adaptable for gas applications.

60.1 Theory of Operation

Coriolis meters typically consist of one or two vibrating tubes with an inlet and an outlet. While some are U shaped, most Coriolis meters have some type of complex geometric shape that is proprietary to the manufacturer. Fluid enters the meter in the inlet, and mass flow is determined based on the action of the fluid on the vibrating tubes. Figure 60.1 shows flow tube response to Coriolis acceleration.

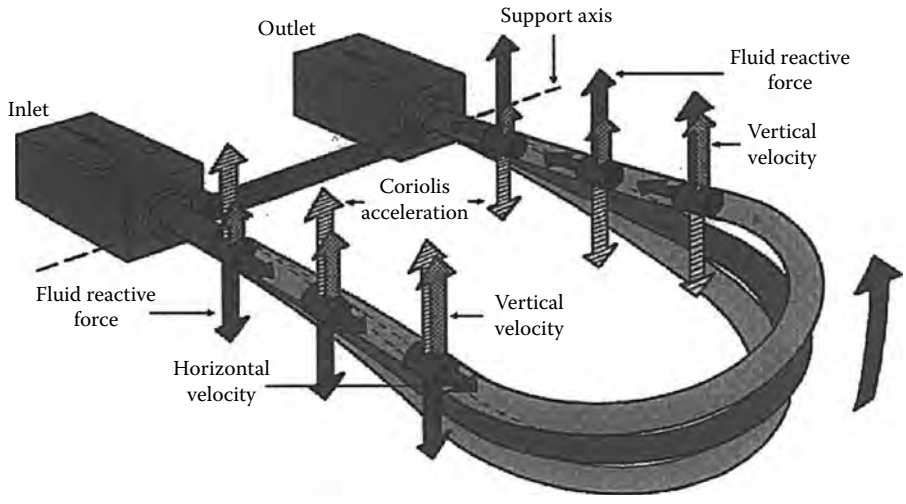


FIGURE 60.1 Flow tube response to Coriolis acceleration.

Common to Coriolis meters is a central point that serves as the axis of rotation. This point is also the peak amplitude of vibration. What is distinctive about this point is that fluid behaves differently, depending on which side of the axis of rotation or point of peak amplitude it is on. As fluid flows toward this central point, the fluid takes on acceleration due to the vibration of the tube. As the fluid flows away from the amplitude of peak vibration, it decelerates as it moves toward the tube outlet. On the inlet side of the tube, the accelerating force of the flowing fluid causes the tube to lag behind its no-flow position. On the outlet side of the tube, the decelerating force of the flowing fluid causes the tube to lead ahead of its no-flow position. As a result of these forces, the tube takes on a twisting motion as it passes through each vibrational cycle; the amount of twist is directly proportional to the mass flow through the tube. Figure 60.2 shows the Coriolis flow tube in a no-flow situation, and Figure 60.3 shows Coriolis tube response to flow.

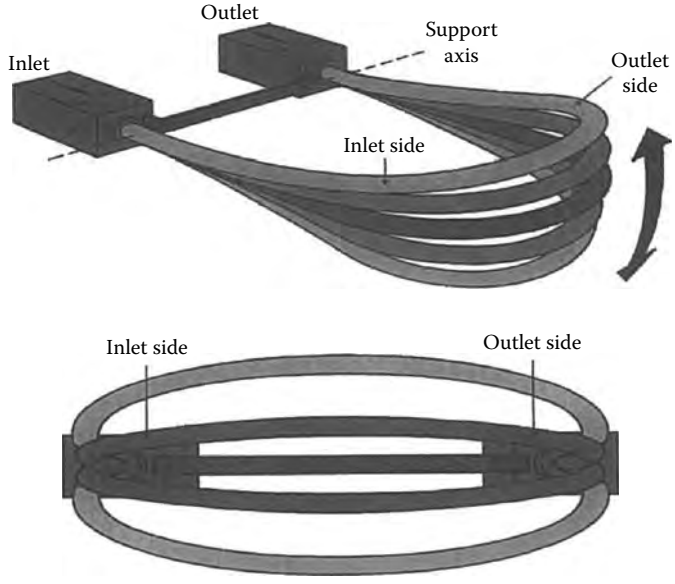


FIGURE 60.2 Two views of an oscillating flow tube with no flow.

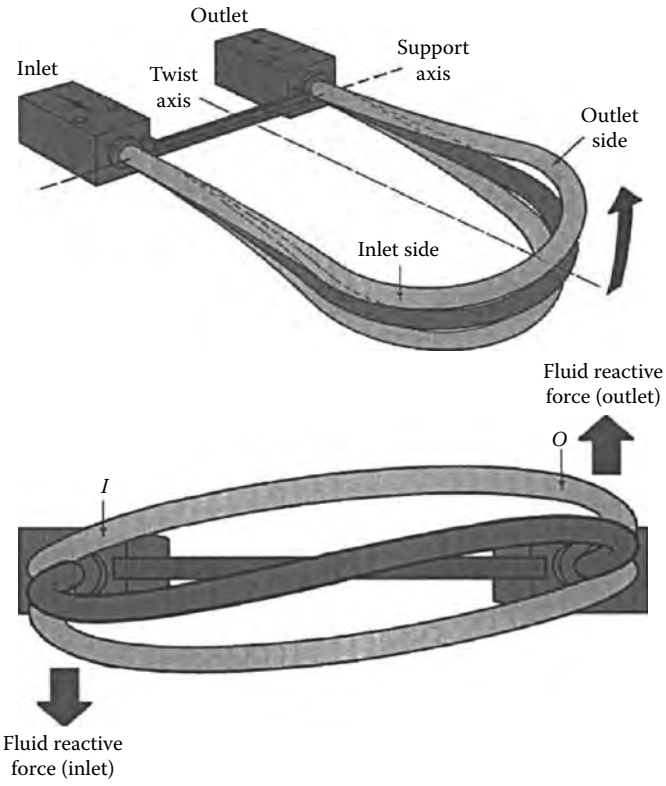


FIGURE 60.3 Two views of an oscillating flow tube in response to flow.

The Coriolis tube (or tubes, for multitube devices) is vibrated through the use of electromagnetic devices. The tube has a drive assembly and has a predictable vibratory profile in the no-flow position. As flow occurs and the tube twists in response to the flow, it departs from this predictable profile. The degree of tube twisting is sensed by the Coriolis meter's detector system. At any point on the tube, tube motion represents a sine wave. As mass flow occurs, there is a phase shift between the inlet side and the outlet side. This is shown in Figure 60.4.

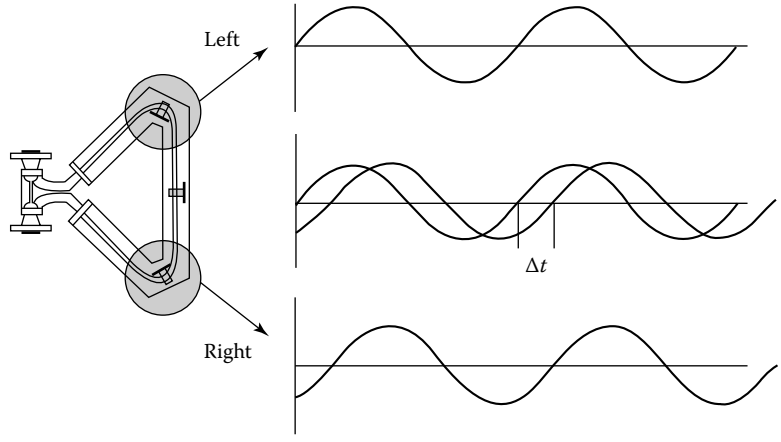


FIGURE 60.4 Phase shift between inlet side and outlet side.

The Coriolis force induced by flow is described by an equation that is equivalent to Newton's second law for rotational motion. This equation is as follows:

$$F = 2m\omega\overline{v} \tag{60.1}$$

where

F is the force
m is the mass to be applied to a known point at a distance L from the axis 0–0
ω is a vector representing angular motion
$\overline{v}$ is a vector that represents average velocity

The Coriolis force is sometime illustrated by likening it to a ball propelled through the air, a long distance in a straight line from the North Pole to a target on the equator. By the time the ball arrives at the equator, it will not land at the apparent target, since the target will have moved due to the rotation of the earth. This makes the path of the ball appear curved from the perspective of the person standing on the North Pole.

A similar example is used with a merry-go-round. If someone throws a ball directly at a horse on the merry-go-round, the ball will miss its target since the horse will have moved by the time the ball arrives at the place where the horse was located at the time the ball was thrown.

If both these examples are intended to illustrate the Coriolis "force," then it should be more appropriately called the Coriolis "effect." There is no force acting to push the ball in a curved direction in either case. Instead, it appears to be curved due to the rotating frame of reference underneath the ball. This Coriolis effect refers to the apparent motion of an object that is passing over a rotating frame of reference when it is viewed from the perspective of the thrown object's frame of reference.

Since Coriolis flowmeters do not have a rotating frame of reference, it is difficult to see in what way they illustrate either the Coriolis force or the Coriolis effect. Instead, they rely on the inertia created by fluid flowing through an oscillating tube, which causes the tubes to twist in proportion to flow rate. While it is unlikely that they will ever be renamed, their name does cause some confusion about their true operating principle.

60.2 Construction

The internal part of the Coriolis tube is the only part of the meter that is wetted. A typical material of construction is stainless steel. Other corrosion-resistant metals such as Hastelloy are used for tube construction. Some meters are lined with Teflon.

Some designs have thin-wall tubes as well as standard tubes. A thin-wall design makes the meter more useful for gas and low-velocity liquid applications, where the amount of twist by the tube is reduced. It is important to be aware of the extent to which the fluid degrades or attacks the tube wall or lining. If the fluid eats away at the wall, this can reduce the accuracy of the meter.

60.3 Advantages

The most significant advantage of Coriolis meters is high accuracy under wide flow ranges and conditions. Because Coriolis meters measure mass flow directly, they have fewer sources of errors. Coriolis meters have a high turndown, which makes them applicable over a wide flow range. This gives them a strong advantage over orifice plate meters, which typically have low turndown. Coriolis meters are also insensitive to swirl effects, making flow conditioning unnecessary. Flow conditioners are placed upstream from some flowmeters to reduce swirl and turbulence for flowmeters whose accuracy or reliability is affected by these factors.

Coriolis meters have a low cost of ownership. Unlike turbine and positive displacement meters, Coriolis meters have no moving parts to wear down over time. The only motion is due to the vibration

of the tube and the motion of the fluid flowing inside the tube. Because Coriolis flowmeters are designed not to be affected by fluid parameters such as viscosity, pressure, temperature, and density, they do not have to be recalibrated for different fluids. Installation is simpler than installation for many other flowmeters, especially orifice plate meters, because Coriolis meters have fewer components.

Coriolis meters can measure more than one process variable. Besides mass flow, they can also measure density, temperature, and viscosity. This makes them especially valuable in process applications where information about these variables reduces costs. It also makes it unnecessary to have a separate instrument to measure these additional variables.

60.4 Disadvantages

The chief disadvantage of Coriolis meters is their initial cost. While some small Coriolis meters have prices as low as $4000, the base price for most Coriolis meters is $6000 and up. The cost of Coriolis meters rises significantly as line sizes increase. The physical size of Coriolis meters increases substantially with the increase in line size. In the past, 150 mm (6 in.) was the upper-line size limit on Coriolis meters today. The large size of some Coriolis meters makes them difficult to handle and can also make installation difficult in some cases.

In the past 5 years, at least four suppliers have begun manufacturing Coriolis flowmeters in sizes from 8 to 16 in. The primary motivation for this change is that Coriolis suppliers want to address the expanding energy markets, especially oil and gas. Many of these large-size meters are designed for custody transfer. While most are targeting liquid applications, mainly hydrocarbon based, some are also being designed for gas applications.

These new meters are 5 ft and more in height, and many cost as much as $75,000. They are also very heavy and bulky. Even so, some end users are ordering them because of the high accuracy of Coriolis meters. While they are still not large enough to handle most pipeline applications, they are beginning to be used in oil production environments.

The lack of an established body of knowledge about Coriolis meters is a substantial disadvantage. Because Coriolis meters were recently invented, not nearly as much data are available about them as are for differential pressure–based flowmeters. This made it more difficult in the past for Coriolis meters to gain approvals from industry associations. However, the American Gas Association and the American Petroleum Institute have now issued standards for the use of Coriolis flowmeters in custody transfer applications.

60.5 Applications

Coriolis meters have no Reynolds number constraints and can be applied to almost any liquid or gas flowing at a sufficient mass flow to affect vibration of the flowmeter. Typical liquid applications include foods, slurries, harsh chemicals, and blending systems. The versatility of Coriolis meters in handling multiple fluids makes them very useful for plants where the flow of multiple fluid types must be measured.

There are an increasing number of gas applications for Coriolis meters. While gas applications are still very much in the minority, the use of this meter to measure gas is likely to increase as more is learned about its use for this purpose.

60.6 A Look Ahead

There are some important areas of research for Coriolis meters. While most Coriolis meters have been bent, several manufacturers have recently introduced straight-tube designs. Manufacturers will continue to fine tune the single-tube design versus double-tube design and to work on tube geometry. As noted earlier, the use of Coriolis meters for gas and also for steam applications is another area for future development.

Bibliography

Coriolis Flowmeter, Coriolis Flowmeter resource Center, http://www.coriolisflowmeter.net/seafloor-spreading/, accessed on August 8, 2012.

Coriolis Mass Flowmeters, UFM, http://www.flowmeters.com/ufm/index.cfm?task=coriolis, accessed August 8, 2012.

Doebelin, E. O., *Measurement Systems: Application and Design*, 4th edn., New York: McGraw-Hill, 1990, pp. 603–605.

Figliola, R. S. and Beasley, D. E., *Theory and Design for Mechanical Measurements*, 2nd edn., New York: John Wiley & Sons, 1995, pp. 475–478.

Plache, K. O., Coriolis/gyroscopic flowmeter, *Mechanical Engineering*, March 1979, 36–41.

Smith, L. and Ruesch, J. R., Mass flowmeters, in D. W. Spitzer (ed.), *Flow Measurement*, Research Triangle Park, NC: Instrument Society of America, 1996, pp. 221–247.

Yoder, J., The Coriolis effect: Unquestionable accuracy, but are Coriolis meters aptly named? *Flow Control*, November 2011, 14–18

Partial List of Vendors and Suppliers

ABB, Inc., 125 E. County Line Rd., Warminster, PA 18974, Phone: 215-674-6580, 800-829-6001 (toll free), Fax: 215-674-7180, www.abb.us/instrumentation

Endress+Hauser, 2350 Endress Place Greenwood, IN 46143, Phone: 888-363-7377, 888-ENDRESS (toll free), Fax: 317-535-8498, www.us.endress.com

FMC Technologies, Inc., Measurement Solutions, 1602 Wagner Ave., P.O. Box 10428, Erie, PA 16514, Phone: 814-898-5000, 800-477-6375 (toll free), Fax: 814-899-8927, www.fmctechnologies.com/en/MeasurementSo…

Keco Engineered Controls, 1200 River Ave., Bldg. 3A, Lakewood, NJ 08701-5657, Phone: 732-901-5900, Fax: 732-901-5904, www.kecocontrols.com

KROHNE, Inc., 7 Dearborn Rd., Peabody, MA 01960, Phone: 978-535-6060, 800-356-9464 (toll free), Fax: 978-535-1720, www.krohne.com/index.php?id=12076

Siemens Corp., 527 Madison Ave., 8th Fl., New York 10022-4611, Phone: 800-SIEMENS (toll free), Fax: 212-767-0580, www.usa.siemens.com/entry/en/

61

Drag Force Flowmeters

Rekha Philip-Chandy
Liverpool John Moores University

Roger Morgan
Liverpool John Moores University

Patricia J. Scully
Liverpool John Moores University

61.1 Design of the Flow Sensor **61**-1
61.2 Principle of Fluid Flow Measurement **61**-2
61.3 Implementation Using Resistive Strain Gages **61**-5
61.4 Optical Fiber Strain Gage Drag Force Flowmeter **61**-7
Fiber Optic Flow Sensor with Rubber Beam • Fiber Optic Flow Sensor without Rubber Beam
61.5 Conclusion **61**-10
References **61**-10
Partial List of Vendors and Suppliers **61**-10

In a *target flowmeter*, a solid object known as a *drag element* is exposed to the flow of fluid that is to be measured. The force exerted by the fluid on the drag element is measured and converted to a value for speed of flow.

The flow-sensing element has no rotating parts, and this makes the instrument suitable for conditions where abrasion, contamination, or corrosion makes more conventional instruments unsuitable. An important application of such flowmeters involves environmental monitoring in areas such as meteorology, hydrology, and maritime studies to measure speeds of air or water flow and turbulence close to the surface. In these applications, the fluid flows are sporadic and multidirectional.

A further advantage of the instrument is that it can be made to generate a measurement of flow *direction* in two dimensions, or even in three dimensions, as well as of flow speed. To implement this feature, the drag element must be symmetrical in the appropriate number of dimensions, and it is necessary to measure the force on it vectorially, again in the appropriate number of dimensions. Provided that the deflecting forces are independent in the sensing directions, the resulting outputs can be added vectorially to generate independent values for flow speed and direction. Target flowmeters using strain gage technology have been used by industry, utilities, aerospace, and research laboratories. They have been used to successfully measure the flow of unidirectional and bidirectional liquids (including cryogenic), gases, and steam (both saturated and superheated) for almost half a century.

Despite these advantages, the target flowmeter appears to have been neglected in favor of more complex and sophisticated devices. The authors have sought to remedy this neglect by developing a sensor suitable for measuring multidirectional flows in two dimensions, instead of measuring only bidirectional flows in a single dimension.

61.1 Design of the Flow Sensor

The sensor described in this section is ideally suited for environmental flow measurement. The operation is based on strain measurement of deformation of an elastic rubber cantilever, to which a force is applied by a spherically symmetrical drag element (Figure 61.1). This sensor has many advantages,

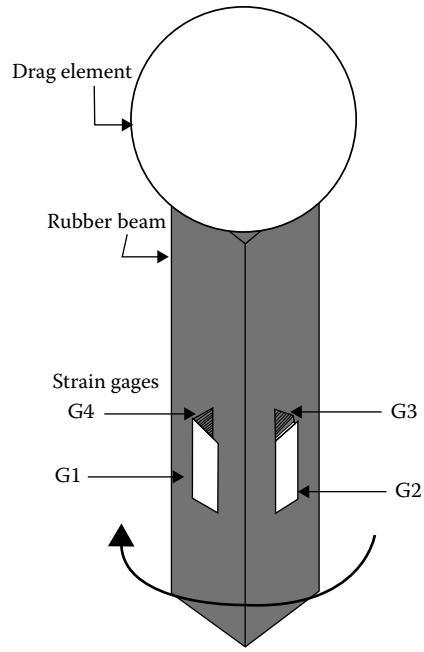


FIGURE 61.1 Schematic of the electric resistance strain gage drag force flow sensor.

including compactness and a simple construction requiring no infrastructure other than a rigid support, and it can cope with fluids containing solid matter such as sludge and slurries, provided that they do not tangle with the drag element or the rubber beam.

According to Clarke [1], the ideal drag element is a flat disk, because this configuration gives a drag coefficient independent of flow rate. Using a spherical drag element, which departs from the ideal of a flat disk [1], the drag coefficient is likely to vary with flow speed, and therefore the gage must be calibrated and optimized for the conditions of intended use. In this discussion, a gage is developed for air flows in the range normally encountered in the natural environment.

The strain measurement can be performed with conventional strain gages, but this limits the applications of the device to conditions where corrosion of the metal-resistive track of the strain gage can be avoided. Therefore, an optical fiber strain gage has been developed as an alternative.

61.2 Principle of Fluid Flow Measurement

The drag force F_D exerted by a fluid on a solid object exposed to it is given by the *drag equation*, which from incompressible fluid dynamics is

$$F_D = \frac{C_D \rho A V^2}{2} \tag{61.1}$$

where

ρ is the fluid density
V is the fluid's velocity at the point of measurement
A is the projected area of the body normal to the flow
C_D is the overall drag coefficient

C_D is a dimensionless factor, whose magnitude depends primarily on the physical shape of the object and its orientation relative to the fluid stream.

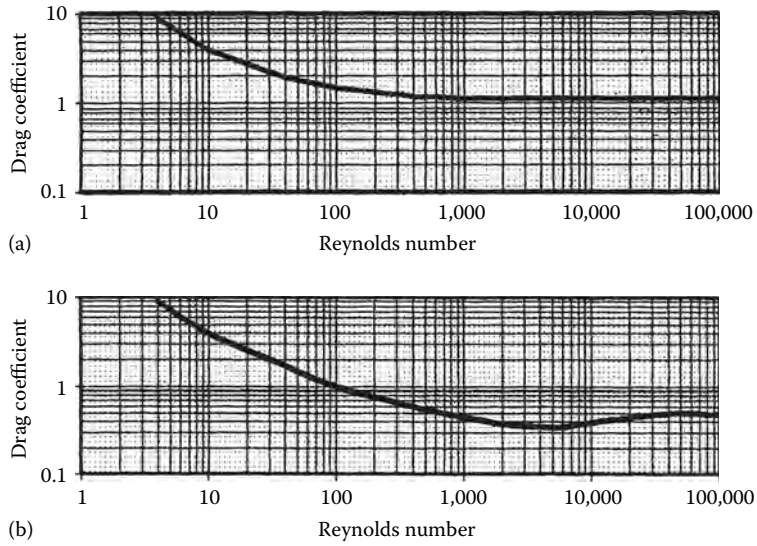


FIGURE 61.2 Graph shows drag coefficient C_D plotted against Reynolds number Re for (a) a flat disk and (b) a sphere.

The drag coefficient C_D for a sphere is related to the Reynolds number (Re), another dimensionless factor, given by

$$\mathrm{Re} = \frac{\rho VA}{\eta} \tag{61.2}$$

where η is the viscosity of the fluid.

A graph of C_D against Reynolds number (derived from Clarke [1]) is shown in Figure 61.2a for a flat disk and Figure 61.2b for a sphere, from which it is evident that the value of C_D, although not constant, is not subject to wide variation over the range of the graph.

Now consider the effect of the force F_D on an elastic beam to which the drag element is attached (as in Figure 61.1). If the mass of the beam can be ignored, the deflection of the beam will only be due to force exerted on the drag element by the fluid. From the theory of elasticity for a cantilever beam of length L with point load at its end [2], the shear force P will be constant along the beam. The bending moment, M, at a point x along the beam will be $P(L - x)$ so that it varies linearly from PL at $x = 0$ to 0 at $x = L$. The distance y is measured from the neutral plane, which for a rectangular section is the mid-plane.

The inverse of the radius of curvature is given by

$$\frac{1}{\rho} = \frac{M}{EI} \tag{61.3}$$

where

E is the modulus of elasticity or Young's modulus

I is the moment of inertia of the cross section where it is assumed that the force is applied perpendicular to the broad face of width b:

$$I = \frac{ba^3}{12} \tag{61.4}$$

Through the thickness a of the beam, the strain is ε_x:

$$\varepsilon_x = \frac{\sigma_x}{E} = \frac{Y}{\rho} \tag{61.5}$$

Substituting Equation 61.3 in Equation 61.5 gives

$$\varepsilon_x = \frac{YM}{EI} \tag{61.6}$$

The strain at the surfaces is

$$\varepsilon_x = \frac{a}{2\rho} = \frac{YM}{2EI} \tag{61.7}$$

Substituting Equation 61.4 and the bending moment into Equation 61.6 gives

$$\varepsilon_x = \frac{6P(L-x)}{Ea^2b} \tag{61.8}$$

or the shear force P is

$$P = \frac{\varepsilon Ea^2b}{6(L-x)} \tag{61.9}$$

Consequently,

$$\frac{C_D \rho A V^2}{2} = \frac{\varepsilon Ea^2b}{6(L-x)} \tag{61.10}$$

Therefore, the strain is

$$\varepsilon = \frac{3C_D \rho A V^2 (L-x)}{Ea^2b} \tag{61.11}$$

From Equation 61.11, the strain is a square law function of fluid speed.

For most fluid flows in the natural environment, a two-dimensional measurement is necessary as the flow in the natural environment is almost always two-dimensional. For a measurement of flow speed and direction, it is necessary to relate wind speed to strain measured in two orthogonal directions, on the assumption (which has been justified by experiment) that the velocity vector has a zero component along (parallel to) the rubber beam support (i.e., $U_z = 0$). The wind speed U has orthogonal components U_x and U_y that are proportional to the strain measured in the x (strain$_x$) and y directions (strain$_y$), respectively. Since U_x is proportional to $\sqrt{\text{strain}_x}$ and U_y is proportional to $\sqrt{\text{strain}_y}$, then the velocity magnitude $|U|$ can be written as

$$|U| = \sqrt{U_x^2 + U_y^2} \tag{61.12}$$

thus,

$$|U| \propto \sqrt{\text{strain}_x + \text{strain}_y} \tag{61.13}$$

Therefore, the magnitude of the velocity, $|U|$, is proportional to the square root of the sum of the orthogonal strain components, strain_x and strain_y. The velocity direction, θ, is calculated using the relation

$$\theta = \tan^{-1}\left[\frac{\sqrt{\text{strain}_y}}{\sqrt{\text{strain}_x}}\right] \tag{61.14}$$

61.3 Implementation Using Resistive Strain Gages

A sensor was constructed for measuring one-dimensional flows by bonding two strain gages onto the opposite sides of a square-sectioned elastic beam, 165 mm long and 13.5 mm square, made from polybutadiene polymer (unfilled vulcanized rubber) supplied by the Malaysian Rubber Products Research Association. The modulus of the beam, at approximately 1.2×10^6 N m^{-2}, was relatively low in order to achieve good sensitivity. The drag element was a table tennis ball of diameter 20 mm, glued to the end of the rubber beam. The strain gages were cupro-nickel alloy on a polyimide film and were bonded to the rubber beam with epoxy resin. The two gage elements were connected in a half-bridge configuration, and the output was taken to a digital strain gage indicator reading directly in microstrain units.

The sensor was mounted in a calibrated wind tunnel, in such a way that one gage underwent compression and the other gage underwent extension. The wind speed was set at different values, and the corresponding strain readings were noted. Results were plotted as wind speed against square root of strain (Figure 61.3). The graph obtained was linear with a correlation coefficient of 0.98. The linearity shows that the strain is a square-law function of speed, which confirms the earlier theory.

In this first prototype of the sensor, the range was limited by beam oscillations at speeds greater than 23 m s^{-1}, causing spurious signals. In principle, however, much greater speeds up to 50 m s^{-1} can be monitored by this sensor, if the construction is suitable.

A second sensor was constructed for measuring two-dimensional flows, in which a square-sectioned rubber beam similar to the previous one was set up with four strain gages attached to the four longitudinal surfaces of the beam. Each opposite pair of strain gage elements was connected in a half-bridge

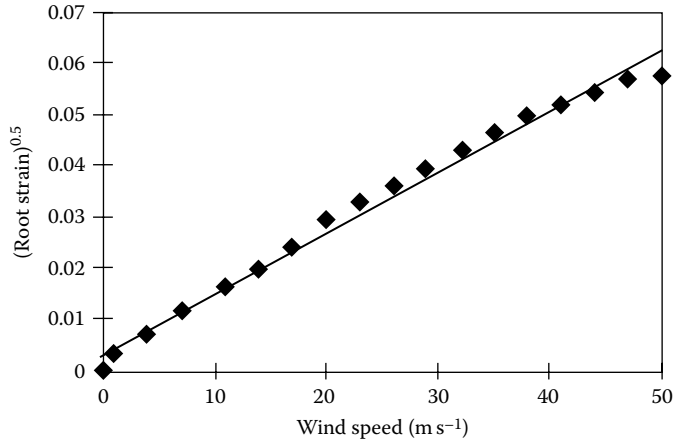


FIGURE 61.3 Root strain versus wind speed for one-dimensional air flow measurement.

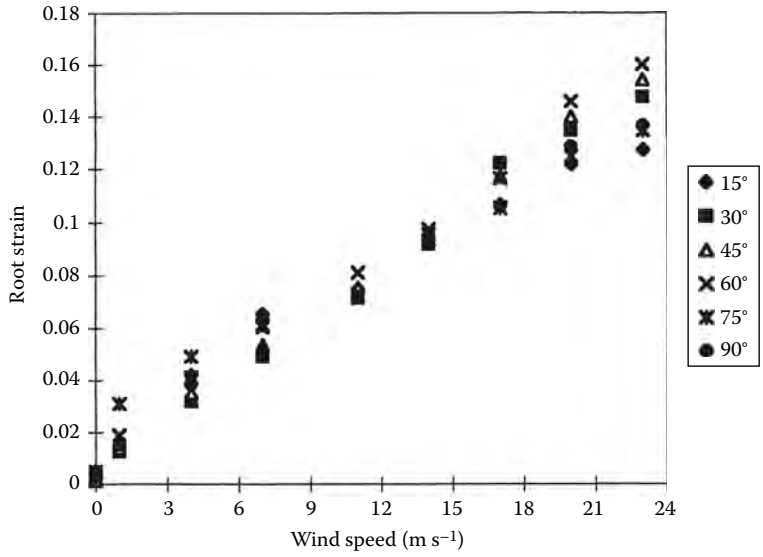


FIGURE 61.4 Root strain plotted against wind speed with the flowmeter oriented at various angles to wind flow.

configuration as before. The two outputs were taken to instrumentation amplifiers and then to a data acquisition board interfaced to the LabVIEW instrumentation software package (National Instruments Corporation). After appropriate signal processing by the software, the values of strain_x and strain_y from both the channels are added and then the square root of the absolute value of this sum is found.

The device was clamped on a turntable, and this was rotated about its longitudinal axis from 0° to 90° at 15° intervals. At each angle, the x and y strain gage outputs were recorded as a function of velocity, over a range from 0 to 23 m s^{-1}. The graph of speed versus $(\text{strain}_x + \text{strain}_y)^{1/2}$ for different angles is shown in Figure 61.4. The data presented in Figure 61.4 can be used to obtain a speed calibration curve for the sensor, which will be valid for multidirectional air flow.

The direction of wind flow was calculated according to Equation 61.14 for the different wind speeds. Figure 61.5 shows the plots of the wind flow direction at different wind speeds for various sensor positions. Several authors have published the inability of calculating the wind direction

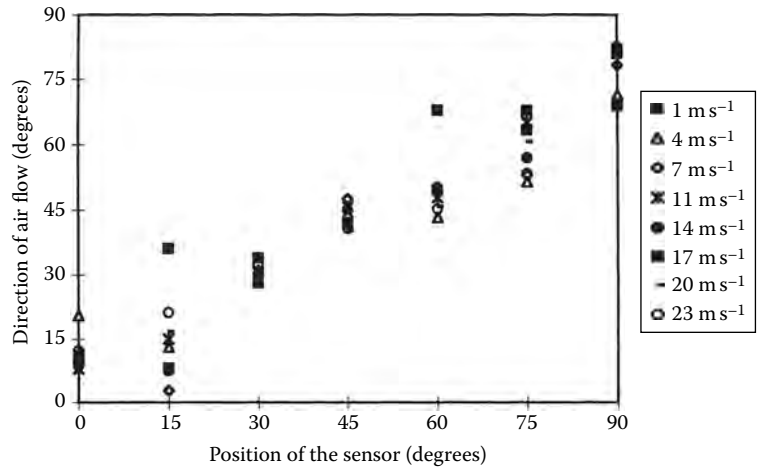


FIGURE 61.5 Calculated direction of wind velocity at different wind speeds.

accurately at low wind speeds. The situation is similar in this case, and the readings get more accurate as the wind speed increases, especially beyond 11 m s^{-1}.

For the measurement of gusts, the time response of the sensor becomes a critical parameter. Experiments were performed to measure the response time by dropping known weights on the free end of the sensor and using a specially written program in Labview to acquire and record the response. The sensor indicated a 95% response time of 50 ms and a time constant of 30 ms.

61.4 Optical Fiber Strain Gage Drag Force Flowmeter

Optical fibers have been applied to measurement of fluid flow, and strain measurement, using interferometric techniques and bending losses [3–5]. The instrument described in the previous sections can be adapted to use optical strain measurement instead of resistive strain gages. In this way, the advantages of the instrument can be preserved; in addition, the flowmeter has the usual benefits of optical fiber sensors such as immunity to electromagnetic interference and intrinsic safety in hazardous environments.

The principle of an optical fiber strain gage is based on the effect of one or more grooves inserted radially into a 1 mm diameter PMMA fiber, which cause a loss in light transmission. As the optical fiber is bent, as in a cantilever, the angle of the grooves varies. These changes of angle cause light to be attenuated at each groove. The intensity variation can be related to the change of the angle of the groove caused by the bending of the cantilever.

To develop an analogy between the optical strain measurement and the resistive strain measurement, the optical strain was calculated using the formula

$$\text{Strain}_{\text{opt}} = \frac{\text{Change in power output}}{\text{Original power output}} = \frac{\Delta P}{P} \tag{61.15}$$

Hence,

$$\text{Strain}_x = \frac{\Delta P_x}{P_x} \quad \text{and} \quad \text{Strain}_y = \frac{\Delta P_y}{P_y} \tag{61.16}$$

Therefore, the magnitude of the velocity, $|U|$ is given by

$$|U| = \sqrt{U_x^2 + U_y^2} = \sqrt{\text{strain}_x + \text{strain}_y}$$

Substituting Equation 61.16 in $|U|$, the optical root strain magnitude is

$$|U| = \sqrt{\left(\frac{\Delta P_x}{P_x} + \frac{\Delta P_y}{P_y}\right)} \tag{61.17}$$

and the direction of wind velocity, θ

$$\theta = \tan^{-1}\left[\frac{U_y}{U_x}\right] = \tan^{-1}\sqrt{\frac{\text{strain}_y}{\text{strain}_x}} \tag{61.18}$$

Substituting Equation 61.16 in θ, the optical strain direction is

$$\theta = \tan^{-1} \frac{\sqrt{\left(\Delta P_y / P_y\right)}}{\sqrt{\left(\Delta P_x / P_x\right)}} \tag{61.19}$$

Two types of the optical fiber flow sensor are described here. In the first type, the rubber beam is used as the deflected device with optical fiber strain gages attached to the deflected beam. The second version of this flow sensor uses an unsupported sensitized 1 mm diameter plastic optical fiber that undergoes deflection in the airflow.

61.4.1 Fiber Optic Flow Sensor with Rubber Beam

The sensor has been built using a rectangular cross-sectioned rubber beam and a spherical drag element with a 1 mm plastic optical fiber glued to the beam with epoxy. The fiber is looped inside the drag element to enhance the sensitivity of the device. Grooves have been made in the fiber surface that extends into the core of the fiber to increase the losses as a function of the bending of the rubber beam (Figure 61.6).

The grooves are normal to the rubber beam on its outside face, and their depths affect only the cladding of the fiber. As the cantilever bends due to the force exerted by the air flow, the angle of the grooves varies. The groove angle increases when the air flow is facing the sensor, and vice versa. These changes of groove angle cause an intensity modulation of the light transmitted through the fiber because light is lost at each groove. Changes in intensity can be related to changes in the angle caused by the force inducing the bending of the cantilever and therefore to the velocity of the fluid.

The orthogonal components of strain were measured by attaching two grooved optical fibers on adjacent sides of the rubber beam, orthogonal to each other, as illustrated in Figure 61.7. These two fibers measured the x and y components of optical strain.

Light from a 1 mW helium–neon laser of peak wavelength 633 nm was split into equal components using a cubic beam splitter and coupled to each fiber. The transmitted intensity through each of the two fibers was monitored using a power meter. The signals from the power meter were sent to a 486 DX2 laptop computer via a data acquisition card to be acquired and processed by Labview. Experiments indicated that this version of the flow sensor could measure wind speeds up to 30 m s^{-1} with a resolution of 1.3 m s^{-1}.

61.4.2 Fiber Optic Flow Sensor without Rubber Beam

This version of the sensor uses 1 mm diameter polymer fiber as the deflection element, with a core diameter of 0.980 mm, and a thin cladding layer of approximately 20 μm [6]. Multiple grooves were etched radially into the fiber surface to a depth of 0.5 mm, extending into the core of the fiber, using a hot scalpel, and a manufactured V-groove template to ensure uniformity of grooves. Six grooves were

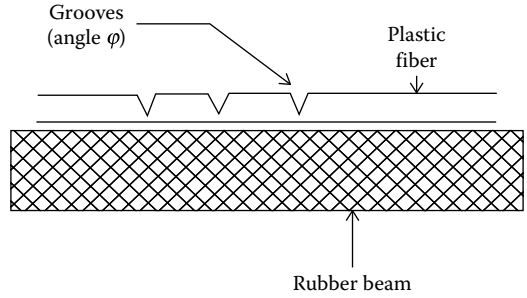


FIGURE 61.6 Principle of operation of the fiber optic drag force flow sensor.

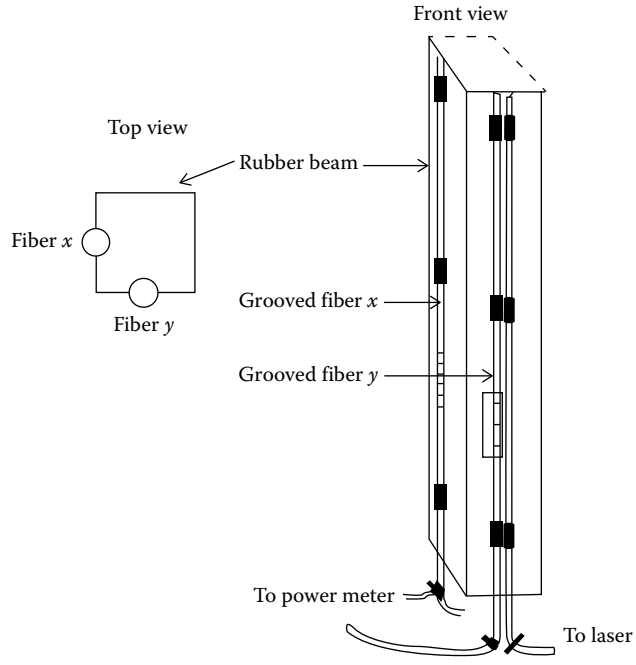


FIGURE 61.7 Front and top views of a section of the fiber optic drag force flow sensor used to measure the two-dimensional fluid flow.

determined as the optimum number to achieve a compromise between insertion losses and strain sensitivity, spaced 0.4 cm apart, over a length of 2.5 cm.

In order to measure strain in two orthogonal directions, perpendicular to the longitudinal axis of the fiber, two fibers were used, as shown in Figure 61.8, with the grooves oriented at 90° to each other. This was achieved by positioning the fibers on two adjacent faces of a beam of square cross section. This beam was short enough to prevent any restriction to the deflection and yet long enough to hold and support the optical fibers. The fibers were looped around so that the looped ends acted as drag elements. The grooved portion of the fiber was unsupported and free to deflect in the air stream. Wind tunnel calibration indicated that the sensor could measure two-dimensional flow up to 35 m s^{-1} with a resolution of 0.96 m s^{-1}.

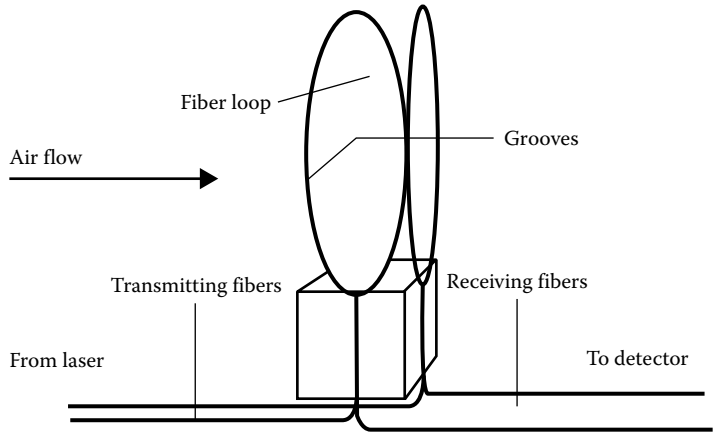


FIGURE 61.8 Optical fiber drag force flow sensor without rubber beam.

61.5 Conclusion

The measurement of one- and two-dimensional fluid velocity, using both optical fiber and conventional strain gages on a deflected beam, for a range of 0–30 m s^{-1}, has been demonstrated experimentally. Although flow visualization and modeling techniques are well advanced in engineering, there is still a need for real measurements, especially in the natural environment. The sensors described in this study are particularly suited for such measurements, which are almost always two-dimensional. The outputs of the sensor representing speed and direction of fluid flow are independent of each other, so the sensor is suitable for environmental applications such as wind measurement or river flow, where the fluid forms gusts and can change direction as well as speed. One noteworthy feature of this sensor is its quick response time of 50 ms, which easily enables the measurement of gusts. The dimensions and materials of the sensor must be chosen to suit the fluid. In terms of sensitivity and resolution, the resistive strain gage sensor is a better option, but replacing the conventional strain gages with optical fiber strain gages ensures electrical noise immunity and intrinsic safety for use in hazardous environments.

References

1. T. Clarke, Design and operation of target flowmeters, in *Encyclopedia of Fluid Mechanics*, Vol. 1, Houston, TX: Gulf Publishing Company, 1986.
2. F. H. Newman and V. H. L. Searle, *The General Properties of Matter*, London, U.K.: Edward Arnold, 1948.
3. S. Webster, R. McBride, J. S. Barton, and J. D. C. Jones, Air flow measurement by vortex shedding from multimode and monomode fibres, *Meas. Sci. Technol.*, 3, 210–216, 1992.
4. J. S. Barton and M. Saudi, A fibre optic vortex flowmeter. *J. Phys. E: Sci. Instrum*. 19, 64–66, 1986.
5. N. Narendran, A. Shukla, and S. Letcher, Optical fibre interferometric strain sensor using a single fibre, *Exp. Tech*. 16(2), 33–36, 1992.
6. R. Philip-Chandy, Fluid flow measurement using electrical and optical fibre strain gauges, PhD thesis, Liverpool John Moores University, Liverpool, U.K., 1997.

Partial List of Vendors and Suppliers

AMETEKTIP (Technical&Industrial Products), 627 Lake St., Kent, OH 44240, Phone: 215-256-6601, Fax: 330-677-3306, www.ametektip.com

Cooper Instruments, 400 Belle Air Lane, P.O. Box 3048, Warrenton, VA 20188, Phone: 540-349-4746, 800-344-3921 (toll free), Fax: 540-347-4755, www.cooperinstruments.com

Kistler North America, 75 John Glenn Dr., Amherst, NY 14228-2119, Phone: 716-691-5100, 888-KISTLER, 888-547-8537 (toll free), Fax: 716-691-5226, www.kistler.com

PCB Piezotronics, Inc., 3425 Walden Ave., Depew, NY 14043-2417, Phone: 716-684-0001, 800-828-8840 (toll free), Fax: 716-684-0987, www.pcb.com

Strainsert Co, 12 Union Hill Rd., West Conshohocken, PA 19428, Phone: 610-825-3310, 800-529-1502 (toll free), Fax: 610-825-1734, www.strainsert.com

Tekscan, 307 W. First St., South Boston, MA 02127, Phone: 617-464-4500, 800-248-3669 (toll free), Fax: 617-464-4266, www.tekscan.com

62

Pitot Probe Anemometer

62.1 Theory **62**-1
62.2 The Pitot Tube in Flow with Variable Density **62**-7
62.3 Volumetric Flow Measurements **62**-10
62.4 A Hybrid System **62**-12
62.5 Commercial Availability **62**-14
References **62**-15

John A. Kleppe
University of Nevada, Reno

62.1 Theory

It is instructive to review briefly the principles of fluid dynamics in order to understand *Pitot tube theory* and applications. Consider, for example, a constant-density fluid flowing steadily without friction through the simple device shown in Figure 62.1. If it is assumed that there is no heat being added and no shaft work being produced by the fluid, a simple expression can be developed to describe this flow:

$$\frac{p_1}{w} + \frac{v_1^2}{2g} + z_1 = \frac{p_2}{w} + \frac{v_2^2}{2g} + z_2 \tag{62.1}$$

where

p_1, v_1, z_1 are the pressure, velocity, and elevation at the inlet, respectively
p_2, v_2, z_2 are the pressure, velocity, and elevation at the outlet, respectively
$w = \rho g$ is the specific weight of the fluid
ρ is the density
$g = 9.80665\ \text{m s}^{-2}$

Equation 62.1 is the well-known Bernoulli equation. The following example will demonstrate the use of Equation 62.1 and lead to a discussion of the theory of Pitot tubes.

Example 62.1

A manometer [1] is used to measure the dynamic pressure of the tube assembly shown in Figure 62.2 [2]. The manometer fluid is mercury with a density of 13,600 kg m^{-3}. For a measured elevation change, Δh, of 2.5 cm, calculate the flow rate in the tube if the flowing fluid is (a) water, (b) air. Neglect all losses and assume STP conditions for the air flowing in the tube and $g = 9.81\ \text{m s}^{-2}$.

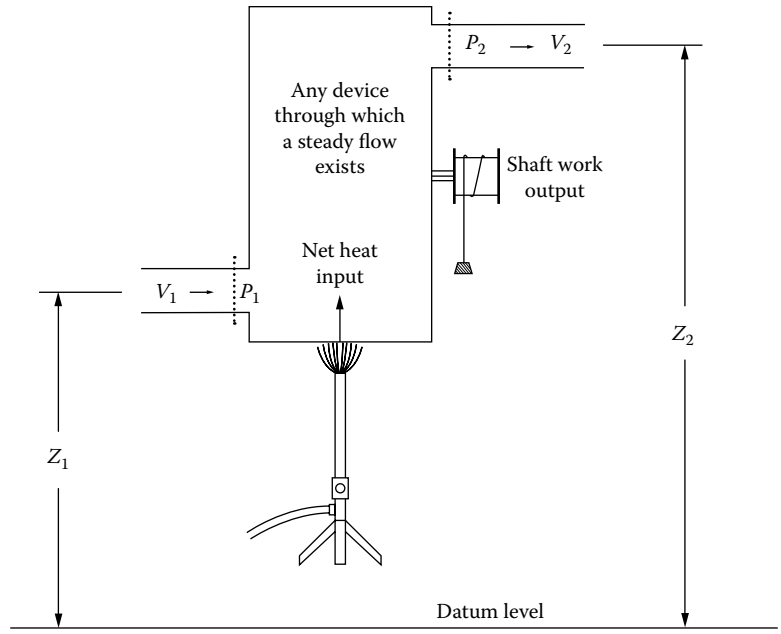


FIGURE 62.1 A device demonstrating Bernoulli's equation for steady flow, neglecting losses. (From Massey, B.S., *Mechanics of Fluids*, Van Nostrand, Princeton, NJ, 1968.)

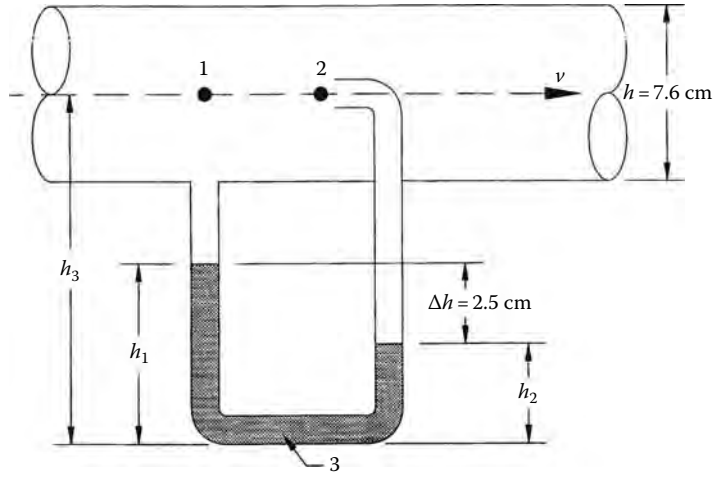


FIGURE 62.2 Using a manometer to measure a Pitot-static tube type assembly—Example 62.1.

Solution 62.1

Begin by writing expressions for the pressure at point 3

$$p_3 = h_1 w_{Hg} + (h_3 - h_1)w + p_1 \tag{62.2}$$

and

$$p_3 = h_2 w_{Hg} + (h_3 - h_2)w + p_2 \tag{62.3}$$

Subtracting these equations and rearrangement yields an expression for the pressure difference

$$p_2 - p_1 = \Delta h(w_{\text{Hg}} - w) \tag{62.4}$$

where w is the specific weight for water or air, etc.

Also using Equation 62.1, one can show that for $z_1 = z_2$ and $v_2 = 0$,

$$p_2 - p_1 = \frac{wv_1^2}{2g} \tag{62.5}$$

1. For water,

$$\begin{aligned} p_2 - p_1 &= \Delta h(w_{\text{Hg}} - w_{\text{H}_2\text{O}}) \\ &= 0.025[13{,}600(9.81) - 998(9.81)] \\ &= 3090.6\,\text{Pa} \end{aligned} \tag{62.6}$$

Then,

$$3090.6 = \frac{(998)(9.81)v_1^2}{2(9.81)} \tag{62.7}$$

or

$$v_1 = 2.5\,\text{m}\,\text{s}^{-1} \tag{62.8}$$

The flow Q is then calculated to be

$$Q = A_1 v_1 = \frac{\pi d^2 v_1}{4} = \frac{\pi(0.076)^2(2.5)}{4} = 0.011\,\text{m}^3\,\text{s}^{-1} \tag{62.9}$$

2. For air, one can use these same methods to show that

$$Q = 0.34\,\text{m}^3\,\text{s}^{-1} \tag{62.10}$$

A point in a fluid stream where the velocity is reduced to zero is known as a stagnation point [3]. Any nonrotating object placed in the fluid stream will produce a stagnation point, x, as seen in Figure 62.3. A manometer connected to point x would record the stagnation pressure of the fluid. From Bernoulli's equation (Equation 62.1), the quantity $p + \frac{1}{2}\rho v^2 + \rho gz$ is constant along a streamline for the steady flow of a fluid of constant density. Consequently, if the velocity v

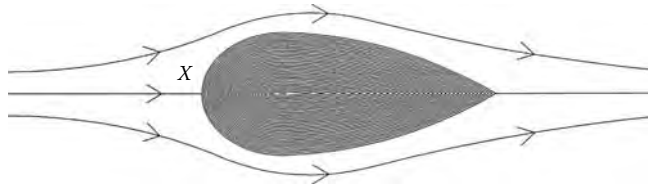


FIGURE 62.3 Flow around a nonrotating solid body. (From Massey, B.S., *Mechanics of Fluids*, Van Nostrand, Princeton, NJ, 1968.)

at a particular point is brought to zero, the pressure there is increased from p to $p + \frac{1}{2}\rho v^2$. For a constant-density fluid, the quantity $p + \frac{1}{2}\rho v^2$ is known as the stagnation pressure p_0 of that streamline, while the term $\frac{1}{2}\rho v^2$—that part of the stagnation pressure due to the motion—is termed the dynamic pressure. A manometer connected to point x would measure the stagnation pressure, and if the static pressure p were also known, then $\frac{1}{2}\rho v^2$ could be obtained. One can show that

$$p_t = p + p_v \tag{62.11}$$

where

p_t is the total pressure, which is the sum of the static and dynamic pressures that can be sensed by a probe that is at rest with respect to the system boundaries when it locally stagnates the fluid isentropically

p is the actual pressure of the fluid whether in motion or at rest and can be sensed by a probe that is at rest with respect to the fluid and does not disturb the fluid in any way

p_v is the dynamic or velocity pressure equivalent of the directed kinetic energy of the fluid

Using Equation 62.11, one can develop an expression that relates to the velocity of the fluid:

$$p_t = p + \frac{1}{2}\rho v^2 \tag{62.12}$$

or solving for v,

$$v = \sqrt{\frac{2(p_t - p)}{\rho}} \tag{62.13}$$

Consider as an example the tube arrangement shown in Figure 62.4. A right-angled tube, large enough to neglect capillary effects, has one end A facing the flow. When equilibrium is attained, the fluid at A is stationary and the pressure in the tube exceeds that of the surrounding stream by $\frac{1}{2}v^2$. The liquid is forced up the vertical part of the tube to a height

$$\Delta h = \frac{\Delta p}{w} = \frac{v^2}{2g} \tag{62.14}$$

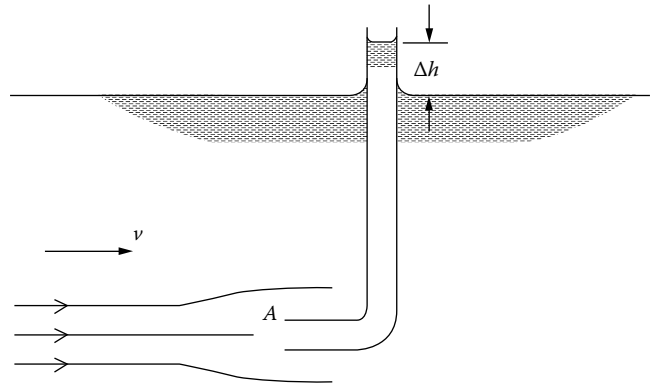


FIGURE 62.4 Right-angle tube in a flow system.

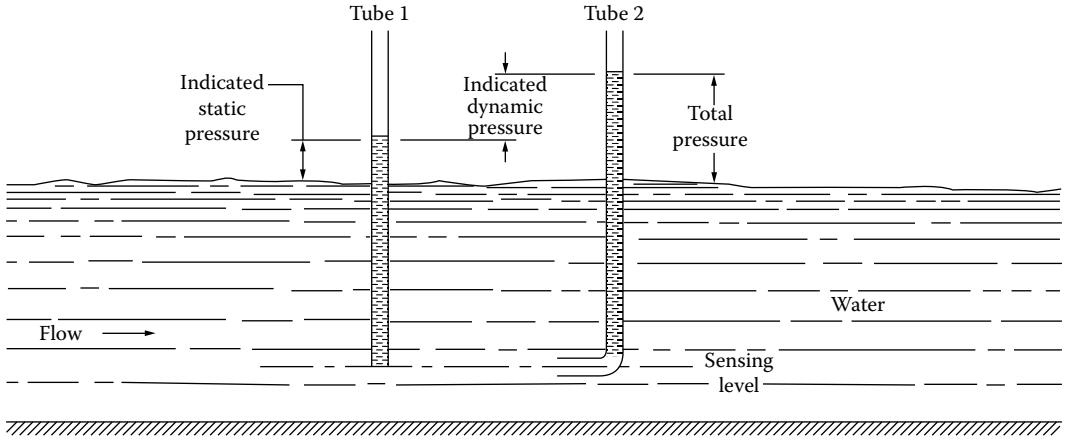


FIGURE 62.5 Basic Pitot tube method of sensing static, dynamic, and total pressure. (From Benedict, R.P., *Fundamentals of Temperature, Pressure and Flow Measurements*, 3rd edn., John Wiley & Sons, New York, 1984. With permission.)

This relationship was used in the example given earlier to solve for v. It must be remembered that the total pressure in a fluid can be sensed only by stagnating the flow isentropically, that is, when its entropy is identical at all points in the flow. Such stagnation can be accomplished by a Pitot tube, as first developed by Henri de Pitot in 1732 [4]. In order to obtain a velocity measurement in the River Seine (in France), Pitot made use of two tubes immersed in water. Figure 62.5 shows his basic Pitot tube method. The lower opening in one of the tubes was taken to be a measurement of the static pressure. The rise of fluid in the 90° tube was used as an indication of the velocity of the flow. For reasons to be discussed later, Pitot's method for measuring the static pressure was highly inadequate and would be considered incorrect today [4].

A modern-day Pitot-static tube assembly is shown in Figure 62.6 [5]. The static pressure is measured using "static holes" or pressure taps in the boundary. A pressure tap usually takes the form of a hole drilled in the side of a flow passage and is assumed to sense the "true" static pressure. When the fluid is moving past in the tap, which is usually the case, the tap will

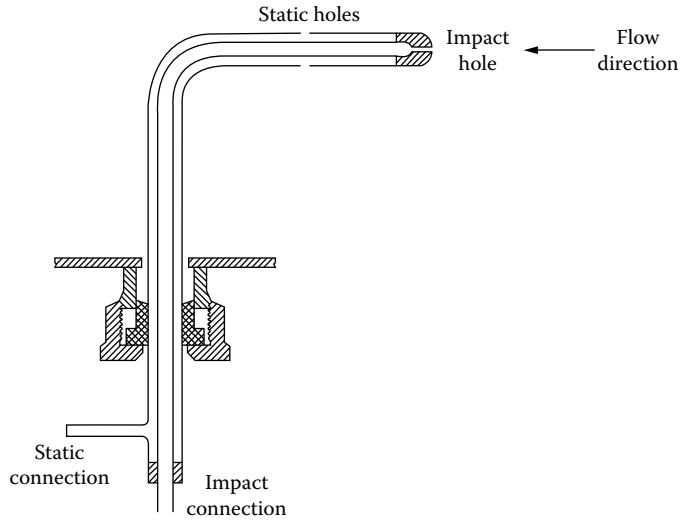


FIGURE 62.6 A modern Pitot-static tube assembly. (From ASME/ANSI PTC 19.2-2010, *Instruments and Apparatus, Part 2, Pressure Measurements*, ASME, Washington, DC, 1987. With permission.)

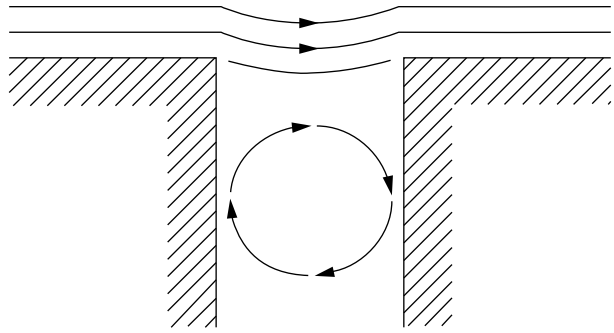


FIGURE 62.7 Pressure tap flow field.

not indicate the true static pressure. The streamlines are deflected into the holes as shown in Figure 62.7, thus setting up a system of eddies. The streamline curvature results in a pressure at the tap "mouth" different from the true fluid pressure. These factors in combination result in a higher pressure at the tap mouth than the true fluid pressure, a positive pressure error. The magnitude of this pressure error is a function of the Reynolds number based on the shear velocity and the tap diameter [5]. Larger tap diameters and high velocities give larger errors [5]. The effect of compressibility on tap errors is not well understood or demonstrated, although correlations for this effect have been suggested [5]. It is possible to reduce tap errors by moving the location of the tap to a nonaccelerating flow location or use pressure taps of smaller diameter. The effect of edge burrs is also noteworthy. All burrs must be removed. There is also an error that results with the angle of attack of the Pitot tube with the flow direction. Figure 62.8

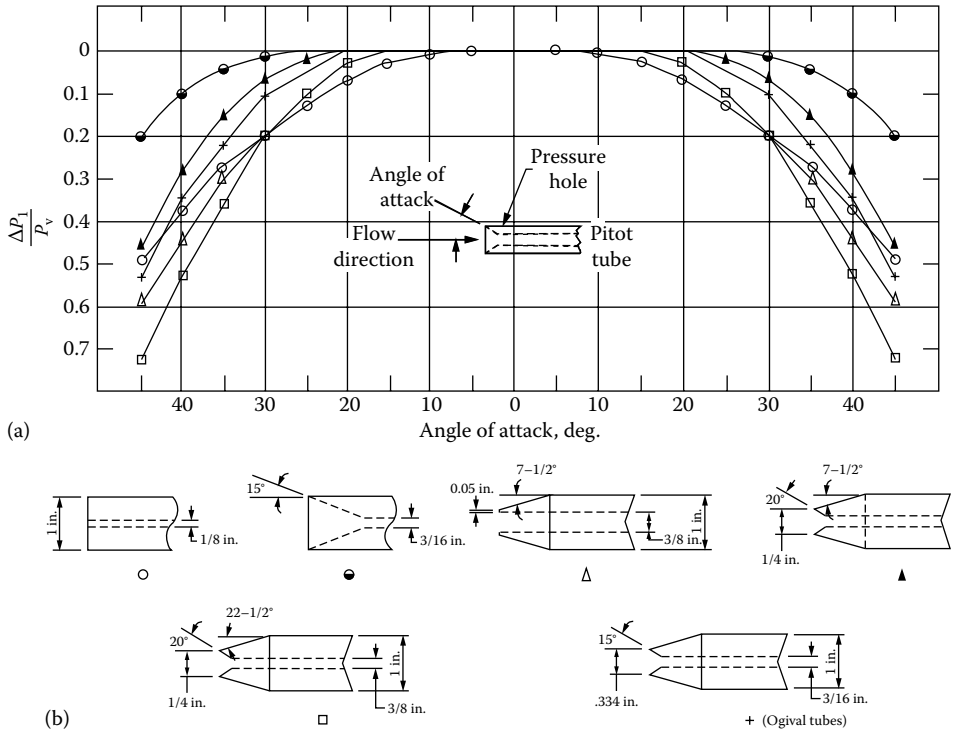


FIGURE 62.8 (a) Variation of total pressure indication with angle of attack and (b) geometry for Pitot tubes. (From ASME/ANSI PTC 19.2-2010, *Instruments and Apparatus, Part 2, Pressure Measurements*, ASME, Washington, DC, 2010. With permission.)

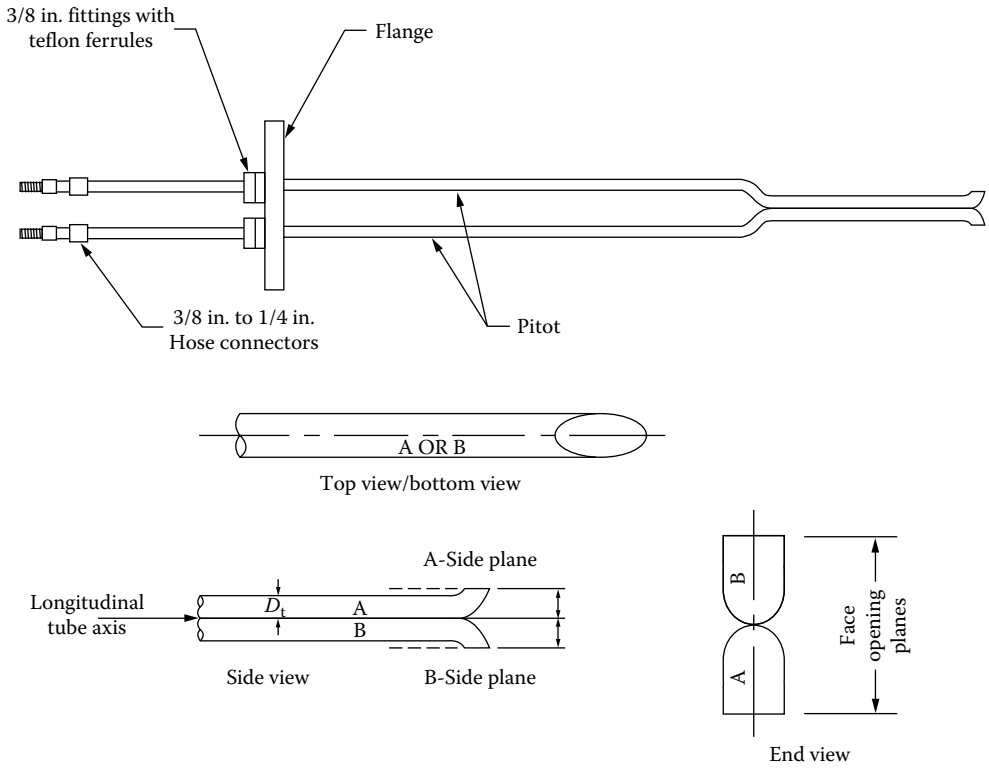


FIGURE 62.9 An S-type Pitot tube for use in gas flow measurement will have specific design parameters. For example, the diameter of the tubing D_t, a gas probe, will be between 0.48 and 0.95 cm. There should be equal distances from the base of each leg of the Pitot tube to its face-opening plane, dimensions d_1, d_2. This distance should be between 1.05 and 1.50 times the external tubing diameter, D_t The face openings of the Pitot tube should be aligned as shown. This configuration of the type S Pitot tube results in a correction coefficient of approximately 0.84. (From EPA, CFR 40 Part 60, Appendix A—Test Methods, September 20, 2011. http://www.epa.gov/ttnemc01/methods/methods2.html.)

shows the variation of total pressure indications as a function of the angle of attack. It can be seen that little error results if the angle of attack is less than ±10°.

A widely used variation of the Pitot-static tube is the type S Pitot tube assembly shown in Figure 62.9. It must be carefully designed and fabricated to ensure that it will properly measure the static pressure. The "static" tube faces backward into the wake behind the probe where the pressure is usually somewhat lower than the undisturbed static pressure. The type S Pitot tube therefore requires the application of a correction factor (usually in the range of 0.84). This correction factor will be valid only over a limited range of velocity measurement. The type S Pitot tube does, however, have the advantage of being compact and relatively inexpensive. A type S Pitot tube can be traversed across a duct or stack to determine the velocity profile and hence total volumetric flow. This is discussed later.

62.2 The Pitot Tube in Flow with Variable Density [3]

When a Pitot-static tube is used to determine the velocity of a constant-density fluid, the stagnation pressure and static pressure need not be separately measured: it is sufficient to measure their difference. A high-velocity gas stream, however, can undergo an appreciable change of density in being brought to rest at the front of the Pitot-static tube; under these circumstances, stagnation and static pressures must

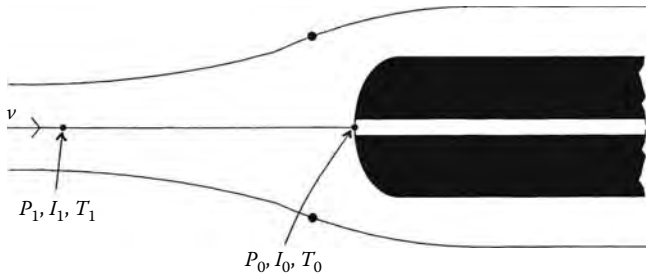


FIGURE 62.10 Pitot static tube in uniform subsonic flow.

be separately measured. Moreover, if the flow is initially supersonic, a shock wave is formed ahead of the tube, and, thus, results for supersonic flow differ essentially from those for subsonic flow. Consider first the Pitot-static tube in uniform subsonic flow, as in Figure 62.10.

The process by which the fluid is brought to rest at the nose of the tube is assumed to be frictionless and adiabatic. From the energy equation for a perfect gas, it can be shown that [3,6].

$$\frac{v^2}{2} = C_p(T_0 - T) = C_p T_0 \left\{1 - \left(\frac{p}{p_0}\right)^{(\gamma-1)/\gamma}\right\} \tag{62.15}$$

where

v is the velocity
C_p is the specific heat at constant pressure
T is the absolute temperature of the gas
T_0 is the absolute temperature at stagnation conditions
p is the total pressure
γ is the ratio of specific heats

For measuring T_0, it is usual to incorporate in the instrument a small thermocouple surrounded by an open-ended jacket. If T_0 and the ratio of static to stagnation pressure are known, the velocity of the stream can then be determined from Equation 62.15.

The influence of compressibility is best illustrated using the Mach number, M. It can be shown that [3]

$$\frac{p_0}{p} = \left(1 + \frac{\gamma - 1}{2} M^2\right)^{\gamma/(\gamma-1)} \tag{62.16}$$

For subsonic flow, $[(\gamma - 1)/2]M^2 < 1$ and so the right side of Equation 62.16 can be expanded by the binomial theorem to give

$$\frac{p_0}{p} = 1 + \frac{\gamma}{2} M^2 + \frac{\gamma}{8} M^4 + \frac{\gamma(2-\gamma)}{48} M^6 + \cdots \tag{62.17}$$

$$p_0 - p = \frac{p\gamma M^2}{2}\left\{1 + \frac{M^2}{4} + \left(\frac{2-\gamma}{24}\right) M^4 + \cdots\right\} \tag{62.18}$$

$$= \frac{1}{2\rho v^2}\left\{1 + \frac{M^2}{4} + \left(\frac{2-\gamma}{24}\right) M^4 + \cdots\right\} \tag{62.19}$$

TABLE 62.1 Variation of "Compressibility Factor" for Air

M	$\frac{p_0 - p}{1/2\rho v^2}$
0.1	1.003
0.2	1.010
0.3	1.023
0.4	1.041
0.5	1.064
0.6	1.093
0.7	1.129
0.8	1.170
0.9	1.219
1.0	1.276

The bracketed quantity is the compressibility factor and represents the effect of compressibility. Table 62.1 indicates the variation of the compressibility factor with M for air with $\gamma = 1.4$.

It is seen that for $M < 0.2$, compressibility affects the pressure difference by less than 1%, and the simple formula for flow at constant density is then sufficiently accurate. For larger values of M, however, the compressibility must be taken into account.

For supersonic flow, Equation 62.16 is not valid because a shock wave forms ahead of the Pitot tube, as shown in Figure 62.11, and, thus, the fluid is not brought to rest isentropically. The nose of the tube must be shaped so that the shock wave is detached, that is, the semiangle must be greater than 45.6° [3].

If the axis of the tube is parallel to the oncoming flow, the wave can be assumed normal to the streamline leading to the stagnation point. The pressure rise across the shock can therefore be given by

$$\frac{p_2}{p_1} = \frac{1+\gamma M_1^2}{1+\gamma M_2^2} \tag{62.20}$$

In the subsonic region downstream of the shock, there is a gradual isentropic pressure rise that can be represented as

$$\frac{p_0}{p_1} = \frac{p_0}{p_2}\frac{p_2}{p_1} = \left(1+\frac{\gamma-1}{2}M_2^2\right)^{\gamma/(\gamma-1)}\frac{1+\gamma M_1^2}{1+\gamma M_2^2} \tag{62.21}$$

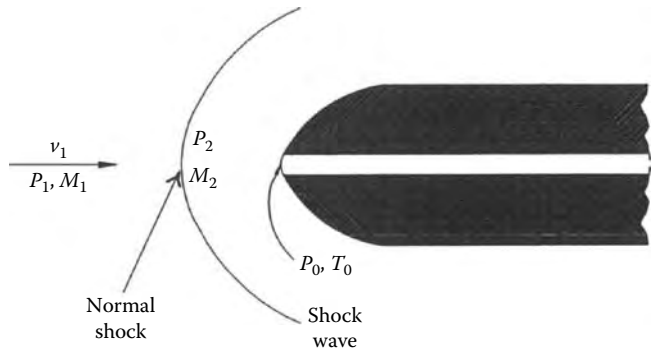


FIGURE 62.11 Pitot tube in supersonic flow switching a shock wave formed ahead of the pitot tube.

Finally, one obtains Rayleigh's formula

$$\frac{p_0}{p_1} = \left\{ \frac{(\gamma+1)^{\gamma+1}}{2\gamma M_1^2 - \gamma + 1} \left(\frac{M_1^2}{2} \right)^{\gamma} \right\}^{1/(\gamma-1)} \tag{62.22}$$

This expression for air reduces to

$$\frac{P_0}{P_1} = \frac{166.9 M_1^7}{(7M_1^2 - 1)^{2.5}} \quad \text{when } \gamma = 1.4 \tag{62.23}$$

Although a conventional Pitot-static tube gives satisfactory results at Mach numbers low enough for no shock waves to form, it is unsuitable in supersonic flow because its "static holes" or "pressure taps," being in the region downstream of the shock, do not then register p_1, nor do they register p_2 since this is found only on the central streamline, immediately behind the normal part of the shock wave. Consequently, p_1 is best determined independently—for example, through an orifice in a boundary wall well upstream of the shock. Where independent measurement of p_1 is not possible, a special Pitot-static tube can be used in which the static holes are much further back (about 10 times the outside diameter of the tube) from the nose. The oblique shock wave on each side of the tube has by then degenerated into a Mach wave across which the pressure rise is very small.

When $M_1 = 1$, the pressure rise across the shock is infinitesimal and, thus, Equations 62.16 and 62.22 both give

$$\frac{p_0}{p_1} = \left\{ \frac{(\gamma+1)}{2} \right\}^{\gamma/(\gamma-1)} = 1.893 \text{ (for air)} \tag{62.24}$$

A small value of p_0/p therefore indicates subsonic flow and a larger value supersonic flow.

Notice that Equation 62.22 enables the upstream Mach number to be calculated from the ratio of stagnation to static pressure. Since the stagnation temperature does not change across a shock wave,

$$C_p T_0 = C_p T_1 + \frac{v_1^2}{2} = C_p \frac{v_1^2}{\gamma R M_1^2} + \frac{v_1^2}{2} \tag{62.25}$$

Thus, v_1 can also be calculated if T_0 is determined.

62.3 Volumetric Flow Measurements

The currently accepted method for measuring volumetric gas flow in ducts and stacks involves the use of Pitot tubes to obtain the velocity at points of equal area of the cross-sectional areas of the stack [7]. For example, Figure 62.12 shows a case where the circular stack of cross-sectional area A has been divided into 12 equal areas. An estimate of the average volumetric flow velocity is determined using the following relationship:

$$\bar{v}_n \approx \frac{\sum v_n A_n}{A} = \frac{A_i \sum v_n}{N A_i} = \frac{1}{N} \sum v_n \tag{62.26}$$

where

A_i is the one segment of the equal-area segments
N is the number of equal-area segments
v_n is the velocity measured at each point of equal-area segments

This relationship shows that one can estimate the average volumetric flow velocity by taking velocity measurements at each point of equal area and then calculate the arithmetic mean of these measurements.

Traverse Point	Distance % of Diameter
1	4.4
2	14.6
3	29.6
4	70.4
5	85.4
6	95.6

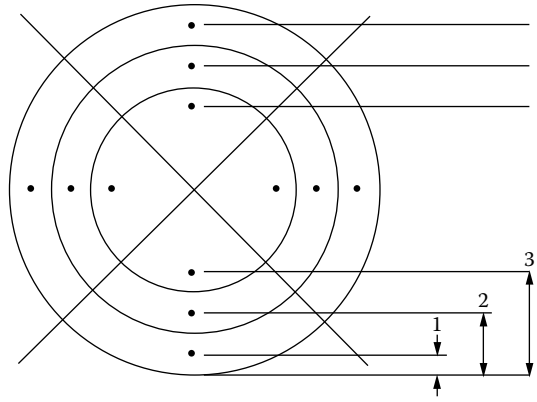


FIGURE 62.12 Example showing circular stack cross section divided into 12 equal areas, with location of traverse points indicated.

It is clearly seen that a different result would be obtained if one were to simply take velocity measurements at equidistant points across the measurement plane and then take the arithmetic mean of these measurements. What would result in this case would be the path-averaged velocity, $\overline{v}_p$, which would be in error.

The sampling site and the number of traverse points designated will affect the quality of the volumetric flow measurement. The acceptability of the sampling procedure is generally determined by the distances from the nearest upstream and downstream disturbances (obstruction or change in direction) to gas flow. The minimum requirements for an acceptable sampling procedure can be found in the literature [7].

An automated system for accomplishing this measurement is shown in Figure 62.13 [8].

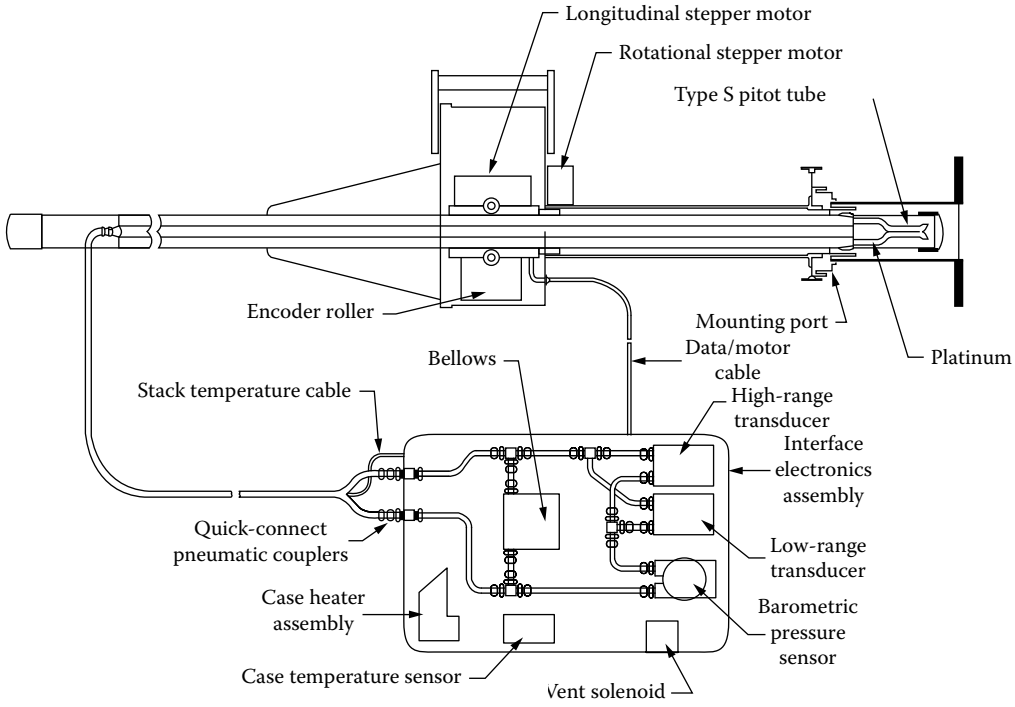


FIGURE 62.13 Automated probe consists of a type S Pitot tube and a platinum RTD mounted onto a type 319 stainless steel probe. (From Elliott, T. C., *Power*, May, 31, 1995. With permission.)

62.4 A Hybrid System

A hybrid system that combines sonic (acoustic) and Pitot tube technology has been developed to measure volumetric flow in large ducts and stacks [9–12]. A block diagram of this system is shown in Figure 62.14. The sensors (Figure 62.15) are mounted on opposite sides of the stack or duct at an angle θ to the flow direction. The acoustic portion of the sensor measures the flight time of the sound waves with and against the gas flow. It can easily be shown [9] that by transmitting and receiving the sound waves in opposite directions, the path average velocity of the gaseous medium can be determined from

$$\bar{v}_p = \frac{d}{2\cos\theta}\left(\frac{\tau_2 - \tau_1}{\tau_1\tau_2}\right) \text{m s}^{-1} \tag{62.27}$$

where

$\bar{v}_p$ is the path average velocity of the gas m s^{-1}
d is the distance between the transceivers (m)
θ is the angle, in degrees, of the path of the transducers with the vertical
τ_1 is the flight time of the sound with the gas flow (s)
τ_2 is the flight time of the sound against gas flow (s)

The result of this part of the total measurement is the area under the velocity curve plotted in Figure 62.16. The Pitot tubes provide differential pressure measurements at two points within the stack. The differential pressure is converted to velocity in a unique manner. The flight times of the

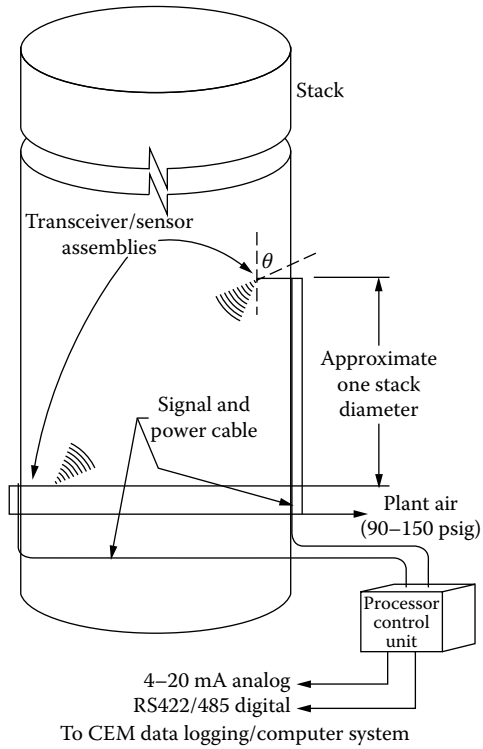


FIGURE 62.14 Block diagram of the hybrid system.

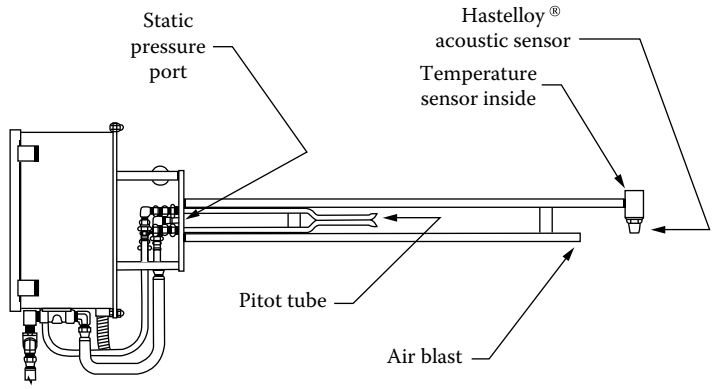


FIGURE 62.15 Acoustic probe contains acoustic, Pitot, and temperature sensors.

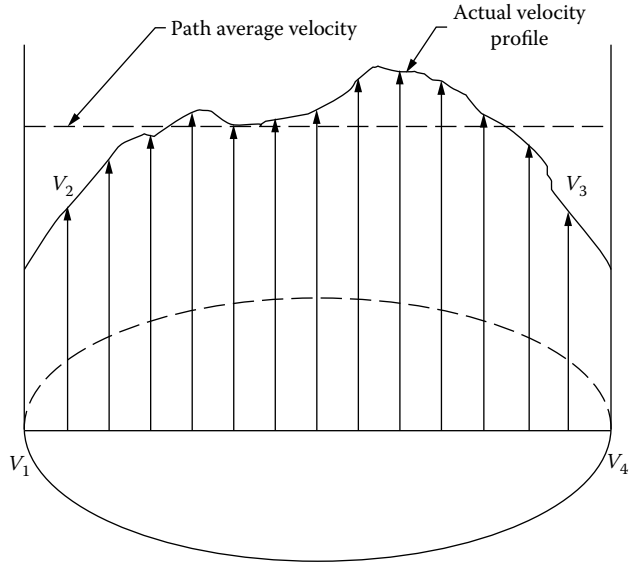


FIGURE 62.16 The velocity profile in a typical large duct or stack can vary greatly, thus changing the total volumetric flow. The hybrid system assumes V_1 and V_4 to be zero, measures V_2 and V_3 using the Pitot tubes, and provides the path average (area under the curve) using the acoustic portions of this sensor. $V_1 = V_4 = 0$ at walls.

acoustic wave, when properly combined with the temperature sensor reading, provide a measurement of the molecular weight of the wet flue gas. This value is then used to obtain the point velocity measurements shown as V_2 and V_3 in Figure 62.16. The actual flow profile curve is then estimated using the values V_1, V_2, V_3, and V_4 and the area under the flow profile curve generated by the acoustic portion of the system. The final part of the measurement involves using the static pressure measurements and the stack temperature measurements to calculate the total standard volumetric flow in scfh (wet).

The vast majority of fossil-fired utility generating units and many industrial sources within the United States have installed and certified continuous emission monitoring systems (CEMS) in accordance with the EPA's 40 CFR Part 75 CEM Rule as required under the Acid Rain Program as well as the NOX Budget Program and Clean Air Rule (CAIR). Under Part 75, sources are required to report emissions data on

a quarterly basis and submit monitoring plan information that identifies the source, power generating units, the emission being monitored, analyzer manufacturer, and sample acquisition methods [13].

Most commercially available flue gas flow monitors operate using one of the following three principles for measuring velocity and volumetric flow: ultrasonic pulse detection, differential pressure, and thermal detection (convective cooling).

The ultrasonic pulse detection system is a form of the hybrid system where corrections to the flow shape are provided from other correlated data such as change in plant loading. It has been reported in Ref. [13] that the ultrasonic flow monitoring technology comprises approximately 67% of the installed base in the United States with differential pressure or Pitot tube based systems having a share of 29%. The remaining 4% of the flow monitoring system are thermal and others.

62.5 Commercial Availability

There is a variety of materials used to construct Pitot tubes. The reasons for this are that Pitot tubes are used to measure a wide range of fluids. For example, to use a type S Pitot tube in a large power plant stack with a wet scrubber where the environment is extremely hostile and corrosive, stainless steel 316 or C276 (Hastelloy®) must be used. This, of course, makes the price of the Pitot tube as varied as its application. Many of the basic type S Pitot tube probes themselves are manufactured by a few small companies who, in turn, supply them on an OEM basis to others.

A typical type S Pitot tube assembly, such as that shown in Figure 62.9, can be purchased from:

EMRC
3730 North Pellegrino Drive
Tucson, AZ 85749
Tel: (520) 749-2167
Fax: (520) 749-3582
www.gasflowmonitors.com

A typical modern Pitot-static assembly, such as that shown in Figure 62.6, can be purchased from:

Dwyer Instruments, Inc.
102 Indiana Highway 212
(P.O. Box 373)
Michigan City, IN 46361
Tel: (219) 879-8000
Fax: (219) 872-9057
www.dwyer-inst.com

More complex, custom-designed and fabricated Pitot-static probes for use on aircraft are available at

Rosemount Aerospace, Inc.
14300 Judicial Road
Burnsville, MN 55306-4898
Tel: (612) 892-4300
Fax: (612) 892-4430
www.goodrich.com/Goodrich/Business/Sensors-and-Integrated Systems

Table 62.2 lists a number of manufactures/vendors that sell Pitot tube and general differential pressure measurement instrumentation.

TABLE 62.2 Sample of Manufacturers/Vendors

Name	Address	Telephone/Fax	Probe Type
EMRC	3730 North Pellegrino Dr. Tucson, AZ 85749	Tel: (520) 749-2167 Fax: (520) 749-3582 www.gasflowmonitors.com	Type S Pitot probe
Dwyer Instruments, Inc.	102 Indiana Hwy. 212 (P.O. Box 373) Michigan City, IN 46361	Tel: (219) 879-8000 Fax: (219) 872-9057 www.dwyer-inst.com	Pitot-static tubes and type S Pitot probe
Rosemount Aerospace, Inc.	14300 Judicial Rd. Burnsville, MN 55306-4898	Tel: (612) 892-4300 Fax: (612) 892-4430 www.goodrich.com	Flow angle sensors, Pitot/Pitot-static tubes, vane angle of attack sensors, temperature sensors, ice detectors, and pressure transducers
Air Monitor Corporation	1050 Hopper Avenue Santa Rosa, CA 95406	Tel: (707) 544-2706 (800) AIRFLOW Fax: (707) 526-9970 www.airmonitor.com	Multipoint, self-averaging
United Sciences, Inc.	201 Commonwealth Drive Warrendale, PA 15086	Tel: 704-778-6934 Fax: 724-778-6959 www.ustirata.com	Auto-PROBE 2000® automated Method 2 Testing
Teledyne Monitor Labs	35 Inverness Drive East Englewood, CO 80113	Tel: 800-422-1499, 303-792-3300 Fax: 303-799-4853 www.teledyne-ml.com	Ultrasonic flow sensor

References

1. W. F. Hughes and J. A. Brighton, *Theory and Problems of Fluid Dynamics*, 3rd edn., New York: McGraw-Hill, 1999.
2. J. B. Evett and C. Liu, *Fluid Mechanics and Hydraulics: 2500 Solved Problems*, New York: McGraw-Hill, 1989.
3. B. S. Massey, *Mechanics of Fluids*, Princeton, NJ: Van Nostrand, 1968.
4. R. P. Benedict, *Fundamentals of Temperature, Pressure and Flow Measurements*, 3rd edn., New York: John Wiley & Sons, 1984.
5. ASME/ANSI PTC 19.2-2010, *Instruments and Apparatus, Part 2, Pressure Measurement*, Washington, DC: ASMW, 2010.
6. S. P. Parker, *Fluid Mechanics Source Book*, New York: McGraw-Hill, 1988.
7. EPA, CFR 40 Part 60, Appendix A—Test methods, Revised September 20, 2011.
8. T. C. Elliott, CEM system: Lynchpin holding CAA compliance together, *Power*, May 1995, 31–40.
9. J. A. Kleppe, Principles and applications of acoustic sensors used for gas temperature and flow measurement, *Proceedings of the SENSOR EXPO*, Boston, MA, May 1995, pp. 337–374.
10. J. A. Kleppe, Method and apppartus for flow rate mesurement, U.S. Patent No. 5,392, CY 5, February 28, 1995.
11. Guidelines for flue gas flow rate monitoring, EPRI TR-104527, Project 1961-13 Final Report, June 1995.
12. A. Mann and J. A. Kleppe, A report on the performance of a hybrid flow monitor used for CEMS and heat rate applications, *Proceedings of the EPRI 1996 Heat Rate Improvement Conference*, Dallas, TX, May 1996, Part 33, pp. 1–13.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

63

Thermal Dispersion Mass Flow Meters

63.1 Introduction **63**-1
63.2 Description **63**-2
63.3 Principle of Operation **63**-5
63.4 Dry-Sensor Microprocessor-Based System **63**-11
63.5 Flow Meter Selection, Installation, and Flow Calibration **63**-12
Manufacturers
References **63**-14

John G. Olin
Sierra Instruments, Inc.

63.1 Introduction

Thermal dispersion mass flow meters measure the mass flow rate of fluids, primarily gases, flowing through a closed conduit. Their first-general description is attributed to L. V. King, who, in 1914 [1], published his famous King's Law, revealing how a heated wire immersed in a fluid flow measures the mass velocity at a point in the flow. He called his instrument a "hot-wire anemometer." Since then, hot-wire and hot-film anemometers have seen wide use in fluid mechanics research and as light-duty mass flow meters and handheld air velocity meters. This class of instruments is described in Chapter 59 [2].

It was not until the 1960s and 1970s that we saw the emergence of industrial-grade thermal dispersion mass flow meters that could solve the wide range of general industry's more ruggedized needs for directly measuring the mass flow rate of air, natural gas, and other gases in pipes and ducts. This chapter describes that class of instruments with emphasis on currently available technology. Future advances in the technology of thermal dispersion mass flow meters are described in Refs. [3,4].

Thermal dispersion mass flow meters measure the heat convected into the boundary layer of the gas flowing over the surface of a heated velocity sensor immersed in the flow. Since it is the molecules of the gas, which bear its mass, that carry away the heat, thermal dispersion mass flow meters directly measure mass flow rate. Capillary tube thermal mass flow meters constitute a second type of thermal mass flow technology, but their principle of operation and their applications are sufficiently different that the American Society of Mechanical Engineers has published separate national standards for each type [5,6]. Capillary tube thermal mass flow meters are not covered in this chapter.

Typical gases monitored by industrial thermal dispersion mass flow meters include air, methane, natural gas, carbon dioxide, nitrogen, oxygen, argon, helium, hydrogen, propane, and stack gases, as well as mixtures of these gases and mixtures of hydrocarbon gases. Common applications are combustion air; preheated air; compressed air; boilers; electric power plants; drying of materials; food and beverage industries; natural gas distribution; aeration and digester gas monitoring in waste water treatment plants; cogeneration with biogas; fuel gas; flare gas; semiconductor manufacturing; heating, ventilation, and air conditioning; and single- and multipoint stack gas monitoring.

63.2 Description

Thermal dispersion mass flow meters directly measure the mass flow rate of single-phase pure gases and gas mixtures of known composition flowing through pipes or other flow conduits. They also have limited application to single-phase liquids of known composition. In most of the following text, we shall assume that the fluid is a gas, without the loss of applicability to liquids. Multivariable versions also provide an output for gas temperature.

Thermal dispersion mass flow meters have two primary configurations: in-line and insertion. Figure 63.1a and b, respectively, shows these two configurations and their major components. Figure 63.1c

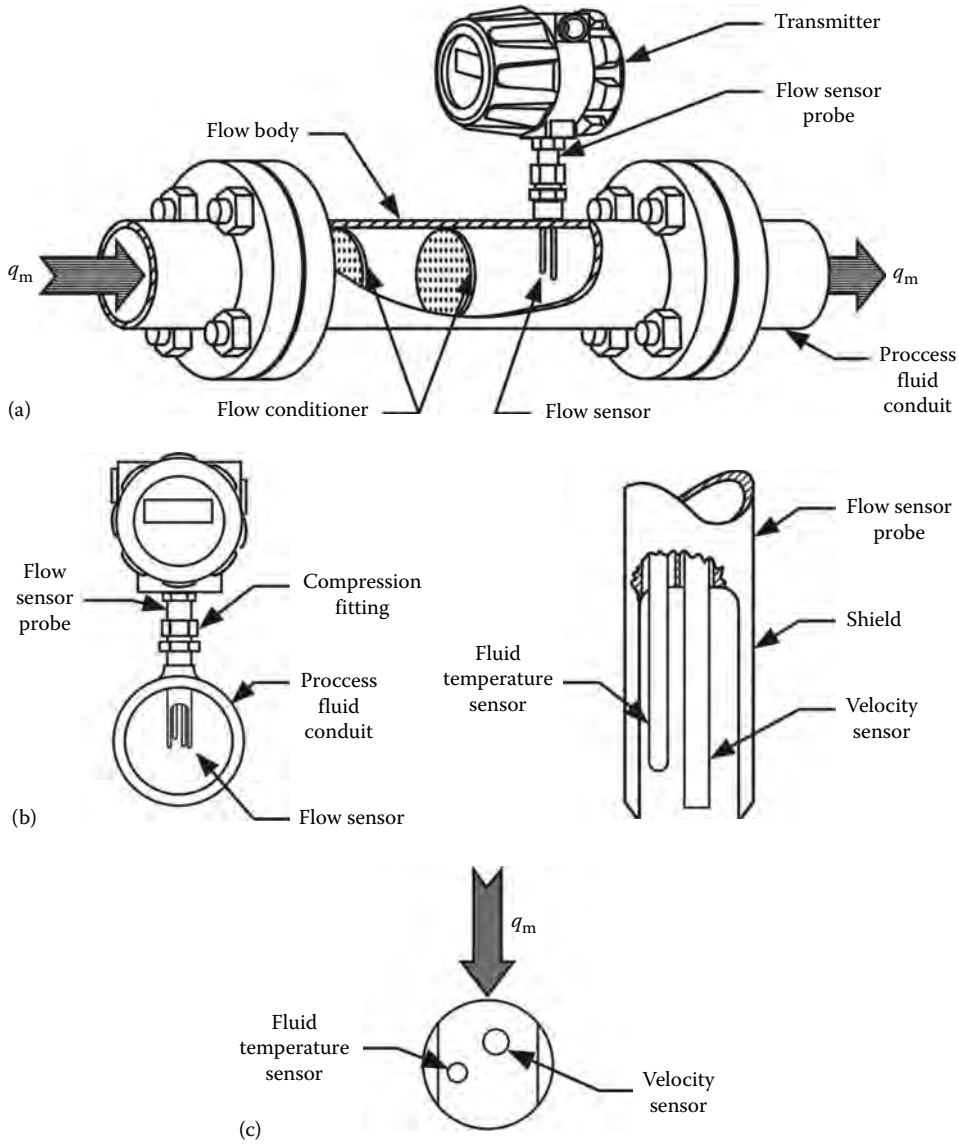


FIGURE 63.1 In-line and insertion meter configurations of thermal dispersion mass flow meters. (a) In-line configuration, (b) insertion configuration, installed, and (c) detail of flow sensor, insertion meter shown. (Adapted from Olin, J.G., Advances in thermal dispersion mass flow meters—Part 1: Principle of operation, Sierra Instruments Technical Report No. SI-JGO-2010-1, 2010. With permission from Sierra Instruments, Inc.)

shows the flow sensor that is common to both configurations, although smaller in-line meters may not have the shield. In-line meters are applied to pipes and ducts with diameters typically ranging from about 10 to 100 mm, but some manufacturers offer sizes up to 300 mm. Process connections include flanges, pipe threads, and compression fittings. A built-in flow conditioner reduces the length of upstream straight pipe required to achieve independence of upstream flow disturbances.

Insertion meters [7] usually are applied to larger pipes, ducts, and other flow conduits having equivalent diameters typically ranging from approximately 75 mm to 5 m. Because insertion meters are more economical than in-line meters, they also have found wide use as flow switches. Compression fittings and flanges are commonly used process connections. Insertion meters measure the mass velocity at a point in the conduit's cross-sectional area, but for applications with smaller conduits, they may be flow calibrated to measure the total mass flow rate through the conduit. Multipoint insertion meters measure the mass velocities at the centroids of equal areas in the cross section of large pipes, ducts, and stacks. The total mass flow rate through the entire conduit is the average mass velocity of the several points multiplied by the total cross-sectional area and the standard mass density of the gas [8].

Figure 63.2a shows a traditional thermal dispersion flow sensor used in in-line and insertion mass flow meters intended for industrial-grade applications. Other kinds of thermal dispersion flow sensors intended for lighter-duty applications are described in Ref. [2]. The flow sensor in Figure 63.2a has a velocity sensor and a separate temperature sensor immersed in the flow stream. For that reason, thermal dispersion mass flow meters are also named "immersible" thermal mass flow meters. The velocity sensor has a single electrically self-heated temperature sensor element located in its tip that both heats the velocity sensor and measures its own average temperature T_1. The gas temperature

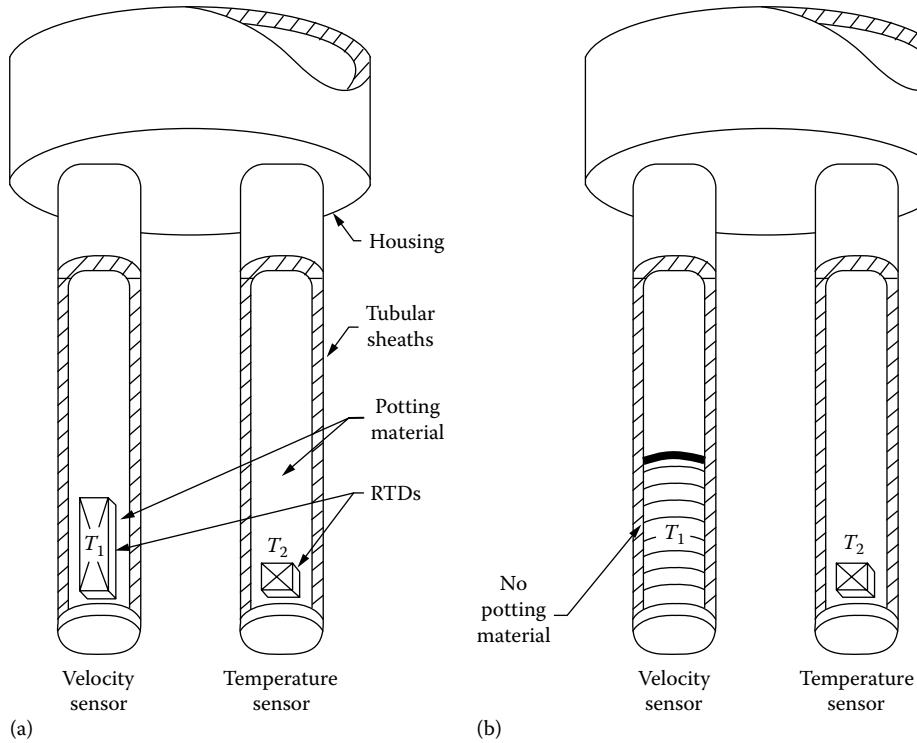


FIGURE 63.2 Two kinds of thermal dispersion mass flow sensors. (a) Flow sensor with potted velocity sensor and (b) flow sensor with dry velocity sensor. (Adapted from Olin, J.G., Advances in thermal dispersion mass flow meters—Part 1: Principle of operation, Sierra Instruments Technical Report No. SI-JGO-2010-1, 2010. With permission from Sierra Instruments, Inc.)

sensor has a single non-self-heated temperature sensor element T_2 located in its tip that measures the gas temperature T. The velocity sensor and the temperature sensor are mounted side by side. Each is enclosed in a rugged, sealed, single-ended, corrosion-resistant metallic tube, usually composed of 316 stainless steel or a nickel alloy. The introduction of this kind of rugged construction in the 1960s and 1970s is responsible for transforming thermal anemometers into industrial-grade instruments. In traditional velocity sensors of the kind shown in Figure 63.2a, the T_1 sensor is potted into the tip of the tubular sheath. Typically, the potting compound, or filler material, is ceramic cement or epoxy. Heat sink grease also has been used for this purpose. In the following text, we shall call this kind of velocity sensor a "potted" one.

For higher-accuracy and higher-stability applications, the temperature sensing elements in the velocity sensor and the temperature sensor in Figure 63.2a and b are either wire-wound or thin-film platinum resistance temperature detectors (RTDs) protected by a thin insulation layer of glass or ceramic. The electrical resistance of RTDs increases as temperature increases, providing the means for transducing their electrical output into temperature. The platinum RTD sensor element in the velocity sensor is called the T_1 element and has a relatively low electrical resistance in the range of about 10–30 Ω. The platinum RTD element in the temperature sensor is called the T_2 element and has a relatively high electrical resistance in the range of 300–1000 Ω. Other types of temperature sensing elements, such as thermistors, thermocouples, and micro-electronic machined devices, have been used for applications with lower-accuracy requirements. In the following text, we shall assume that the T_1 and T_2 elements are platinum RTDs.

In the field, the outside temperature external to the flow sensor can change and be different from its value at the time of flow calibration. For that reason, heat can be conducted in or out of the stems of the velocity sensor and the temperature sensor. Additionally, heat can be conducted from the hot velocity sensor to the cooler temperature sensor via their stems. Both effects are further complicated because they depend on the mass flow rate. These phenomena are collectively called "stem conduction." Stem conduction is a large fraction of the total heat supplied to the velocity sensor and is an unwanted quantity. Left uncorrected, stem conduction constitutes a major source of error in measuring mass flow rate. Flow sensors with long stems have far less stem conduction than those with short, stubby stems.

The use of the potting material between the T_1 element in Figure 63.2a and the internal surface of the sheath has potential long-term stability problems because the filler material can crack or otherwise degrade due to differences in the thermal expansion coefficients of the potting and the sheath materials when exposed to gas temperatures that cycle, change frequently, or are elevated. Any change in the potting material causes a change in the "skin resistance" of the velocity sensor and thereby its stability. Skin resistance and stem conduction are the two major factors that can degrade measurement accuracy if not managed properly. This will be discussed further in the following text.

The construction and assembly of the T_1 element in the flow sensor in Figure 63.2b eliminate the skin resistance problem by (1) avoiding altogether the use of any potting materials between the T_1 element and the internal surface of the sheath and (2) using mating materials that have the same coefficient of thermal expansion. In this design, the T_1 element is cylindrical wire-wound platinum RTD. Potting materials are avoided altogether by means of tightly fitting, as in swaging or press fitting, the wire-wound T_1 element into the sheath. Such velocity sensors are known as "dry" sensors, in contrast with potted velocity sensors fabricated with cements or epoxies that are wet when mixed. In contrast with the velocity sensor, any degradation of potting materials in the temperature sensor changes only its time response, a relatively minor effect.

In operation, the gas temperature sensor in the thermal dispersion mass flow meter measures the gas temperature T. The sensor drive in the transmitter electronics delivers an electrical current I_1 to the velocity sensor such that it is self-heated to an average temperature T_1 that is elevated above the gas temperature. Since it is the molecules of the gas, which bear its mass, that flow over the heated velocity sensor and carry away its heat, thermal dispersion flow meters directly measure the mass flow rate q_m

of the gas or gas mixture. Heat convected from the velocity sensor in this manner depends on the properties of the gas, and therefore the composition of the gas must be known.

In yet another thermal flow sensor construction for in-line meters, the flow sensor is embedded in the wall of the flow body and is not immersed in the flow. This flow sensor consists of a heater element with adjacent upstream and downstream temperature sensing elements. The difference in the two temperatures increases as flow increases, providing the output. This construction is used primarily for low-flow liquid applications.

The transmitter shown in Figure 63.1 is the electronic system that provides the flow sensor drive and many other functions for the flow meter. It accepts the inputs from the two temperature sensing elements as well as the heating current I_1 input and transforms these independent variables into linear analog outputs of the primary dependent variable, mass flow rate q_m, and, in the case of multivariable versions, the gas temperature T. The transmitter is housed in an enclosure that conforms with relevant codes, such as hazardous area codes or area classifications. Digital transmitters with digital displays in engineering units facilitate additional functions. Later, we shall describe microprocessor-based digital transmitters that also provide automatic correction for changes in gas temperature.

Many transport properties of the gas that are involved in convective heat transfer, such as thermal conductivity, viscosity, and Prandtl number, depend on temperature. Likewise, the thermal conductivity of some of the materials in the flow sensor depends on temperature. For this reason, thermal dispersion mass flow meters must correct for changes in gas temperature. In traditional flow meters, this is done by means of an analog Wheatstone bridge at the front end of the flow sensor drive. The velocity sensor and the temperature sensor are located at opposite legs of the bridge. This provides compensation for changes in fluid temperature by adjusting overheat of the velocity sensor. The bridge voltage is a high-level output signal in the order of several volts that provides a high signal-to-noise ratio. The Wheatstone bridge and its temperature-compensation properties are thoroughly described in the literature [2,8–11]. Advanced flow meters with a microprocessor-based flow sensor drive digitally correct for changes in temperature without requiring a Wheatstone bridge [3].

The flow sensor drive in thermal dispersion mass flow meters has two modes of operation: the constant temperature differential mode and the constant current mode. In the constant temperature differential mode of operation, the flow sensor drive maintains at a constant value the difference $\Delta T = T_1 - T$ between the heated velocity sensor T_1 and the gas temperature T. The output signal is the electrical power W supplied to the heated velocity sensor that is required to keep ΔT constant. In the constant current mode of operation, the flow sensor drive maintains at a constant value the current I_1 supplied to the heated velocity sensor. In this case, the output signal is ΔT. Measuring mass flow rate with constant current operation is slower than constant temperature differential operation because the temperature of the entire mass of the velocity sensor must change when velocity changes and also because the masses of the velocity and temperature sensors may be imbalanced. In the following text, we assume constant temperature differential operation and that therefore ΔT is a constant, usually in the range of 20–100 K.

63.3 Principle of Operation

The purpose of this section is to provide a brief description of some of the important principles involved in finding the two output variables—the mass flow rate q_m and the gas temperature T. Figure 63.3 shows the principle of operation of the industrial-grade thermal dispersion flow sensor with the dry velocity sensor shown in Figure 63.2b. The flow sensor is operated in the constant temperature differential mode. The transmitter is the modern microprocessor-based system shown in Figure 63.4. In the following text, we shall call this flow sensor/transmitter combination the "dry-sensor microprocessor-based system." The fluid is a pure gas or gas mixture of known composition. The following treatment specifically applies to the dry velocity sensor, but has general applicability to potted velocity sensors and liquid flows. Refs. [3,9] are good back-up sources for the principle of operation described here, although Ref. [9] specifically applies to hot-wire anemometers.

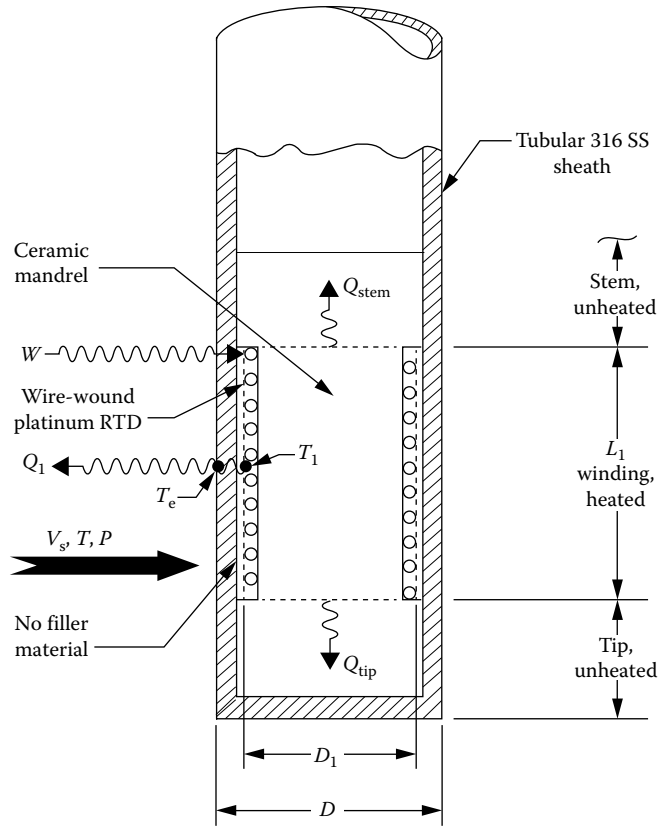


FIGURE 63.3 Principle of operation of the flow sensor with a dry velocity sensor. (Adapted from Olin, J.G., Advances in thermal dispersion mass flow meters—Part 1: Principle of operation, Sierra Instruments Technical Report No. SI-JGO-2010-1, 2010. With permission from Sierra Instruments, Inc.)

The temperatures T_1(K) and T_2(K) are found from the two platinum RTD temperature sensing elements by solving the following equations for temperature using the quadratic formula:

$$R_1 = R_{1,0}\left[1 + A_1\left(T_1 - T_0\right) + A_2\left(T_1 - T_0\right)^2\right] \tag{63.1}$$

$$R_2 = R_{2,0}\left[1 + B_1\left(T_2 - T_0\right) + B_2\left(T_2 - T_0\right)^2\right] \tag{63.2}$$

where

R_1, R_2 are the electrical resistance of temperature sensing elements T_1 and T_2 at temperatures T_1 and T_2, respectively (Ω)

$R_{1,0}$, $R_{2,0}$ are the electrical resistance of temperature sensing elements T_1 and T_2 at temperature T_0 (Ω)

T_0 is the reference temperature, usually 273.15 (K) [0 (°C)]

A_1, B_1; A_2, B_2 are constants provided by the manufacturer (K^{-1}); (K^{-2})

Equations 63.1 and 63.2 specifically apply if temperatures T_1 and T_2 are above 0 °C. Assuming that stem conduction is negligible for the temperature sensor, we can solve Equation 63.2 for T_2. This results in the following relationship for the desired gas temperature output variable:

$$T = T_2 \tag{63.3}$$

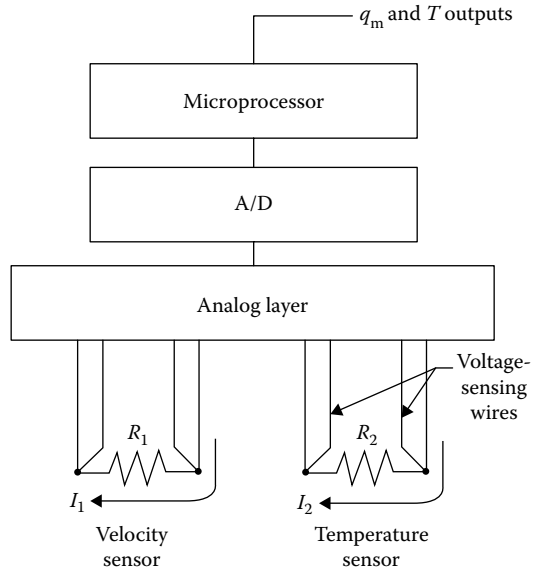


FIGURE 63.4 The dry-sensor microprocessor-based system. The flow meter drives the velocity sensor so that the temperature difference $\Delta T = T_1 - T$ is maintained constant. The system automatically corrects for changes in gas temperature. (Adapted from Olin, J.G., Advances in thermal dispersion mass flow meters—Part 1: Principle of operation, Sierra Instruments Technical Report No. SI-JGO-2010-1, 2010. With permission from Sierra Instruments, Inc.)

The relationship between electrical resistance and temperature for the T_1 sensor sometimes is expressed as

$$R_1 = R_{1,0}\left[1 + \alpha\left(T_1 - T_0\right)\right] \tag{63.4}$$

In Equation 63.4, the term α is the temperature coefficient of resistivity (K^{-1}) and has a value of approximately 0.0039 (K^{-1}) for platinum RTDs. Equation 63.4 is quite accurate for gas temperatures in the range of 0 °C–100 °C.

The dry-sensor microprocessor-based system measures the primary inputs R_1, R_2, and I_1. It then converts them into secondary inputs T_1 and T_2 using Equations 63.1 and 63.2 and computes the electrical power W supplied to the T_1 heated element using the relationship $W = I_1^2 R_1$ (W). The transmitter operates in the constant temperature differential mode and therefore maintains the temperature difference $\Delta T = T_1 - T$ at a constant value. The following text describes some of the important relationships needed to convert these inputs into a linear output of mass flow rate q_m.

In steady state, the first law of thermodynamics states that the energy (or power) flowing into a defined control volume equals the energy flowing out of the control volume. The control volume to which we shall apply this law is the entire heated section of the velocity sensor. As shown in Figure 63.3, this control volume is enclosed in the dotted lines and is defined by length L_1 and diameter D_1. It consists of the cylindrical ceramic mandrel; the R_1 platinum RTD winding around the mandrel that is heated by electrical power W; and a ceramic or high-temperature glass insulating binder for the winding. The first law for this control volume is

$$W = Q_1 + Q_{tip} + Q_{stem} \tag{63.5}$$

In Equation 63.5, Q_{tip} and Q_{stem} are the heat rates (W) conducted from the heated control volume to the tip and stem of the velocity sensor, respectively.

Collectively, they constitute the heat rate lost via the previously mentioned "stem conduction." Q_1 is the heat rate (W) transferred from the heated control volume to the flowing gas by convection and radiation. Q_1 is expressed as

$$Q_1 = h_e(\pi D L_1)\Delta T \tag{63.6}$$

In Equation 63.6, D and L_1 are the outside diameter (m) and the length (m) of the heated section of the cylindrical velocity sensor, respectively. h_e is the effective film coefficient for convection (W/m^2/K), defined as

$$h_e = \frac{h}{1 + h\pi D L_1 R_{skin}} \tag{63.7}$$

where

h is the film coefficient for convection averaged over the surface of the heated section of the velocity sensor (W/m^2/K)

R_{skin} is the skin resistance described in the following text (K/W)

In Equation 63.7, we have ignored radiative heat transfer because it is negligible, except when the gas temperature is very high. The average temperature T_e of the external surface of the velocity sensor over length L_1, shown in Figure 63.3, is less than the average temperature T_1 of the control volume. This is due to the temperature drop that is required to pass the heat rate Q_1 through the intervening layers of material between the control volume and its external surface. These intervening layers are called the "skin." R_{skin}, the average thermal resistance of the skin over length L_1, i.e., the "skin resistance," is defined by

$$Q_1 = \frac{(T_1 - T_e)}{R_{skin}} \tag{63.8}$$

As evidenced by Equation 63.7, at higher flow rates and thus higher values of h, the skin resistance R_{skin} causes a droop in the output signal and consequently a reduction in sensitivity. This is why skin resistance is an unwanted quantity that should be minimized in flow sensor design, should not change over time, and should be taken into account. R_{skin} is the sum of the thermal resistances of the annular layers in the skin of the velocity sensor and is calculated as follows:

$$R_{skin} = \left(\frac{1}{2\pi L_1}\right)\sum_{j=1}^{n}\left[\frac{\ln\left(D_{j+1}/D_j\right)}{k_j}\right] \tag{63.9}$$

where

n is the total number of layers in the skin

$j = 1,\ldots,n$

D_j and D_{j+1} are the inside and outside diameters of layer j, respectively ($D_1 < \cdots < D_{n+1} = D$)

k_j is the thermal conductivity of layer j

The potted velocity sensor shown in Figure 63.2a has three layers in its skin—the thin insulation layer over the platinum RTD, the potting material, and the sheath. The dry velocity sensor in Figure 63.2b also has three layers in its skin—the thin insulation layer over the platinum RTD winding, a thin air gap with an average thickness in the order of 0.005 mm [0.0002 (in.)], and the sheath. The

thin air gap in the dry velocity sensor exists because, although the mating parts are in tight contact via their press fit, slight imperfections in their circularity can cause thin air gaps over a fraction of the mating surface. R_{skin} ranges from 3 to 10 W/K for potted velocity sensors and from 2 to 5 W/K for dry velocity sensors.

h and T are found based on the earlier principles. Given h and T, the next step in finding the mass flow rate q_m is to find the Reynolds number Re using an empirical correlation for convective heat transfer. This correlation has the following two forms, depending upon whether it is solved for the Nusselt number Nu or the Reynolds number Re:

$$\text{Nu} = F_1(\text{Re}, \text{Pr}) \quad \text{or} \quad \text{Re} = F_2(\text{Nu}, \text{Pr}) \tag{63.10}$$

where

Nu = hD/k = Nusselt number, the nondimensional heat-transfer parameter

Re = $\rho VD/\mu$ = Reynolds number, the nondimensional ratio of dynamic forces to viscous forces, based on the outside diameter D of the velocity sensor

Pr = $\mu c_p/k$ = Prandtl number, the nondimensional ratio of the diffusivity of momentum to the diffusivity of heat

k is the thermal conductivity of the gas (W/m/K)

ρ is the mass density of the gas (kg/m^3)

μ is the absolute viscosity of the gas (kg/s/m)

c_p is the coefficient of specific heat of the gas at constant pressure (J/kg/K)

The simplest example of the two forms of the correlation is

$$\text{Nu} = b_1 \text{Pr}^{b_2} \text{Re}^{b_3} \quad \text{or} \quad \text{Re} = \left[\frac{\text{Nu}}{b_1 \text{Pr}^{b_2}} \right]^{1/b_3} \tag{63.11}$$

where b_i is the dimensionless constants, i = 1, 2, or 3.

In Equations 63.10 and 63.11, all properties and dimensions refer to the velocity sensor. In the earlier text, we have assumed that the flow is nonzero, is in the continuum flow regime (i.e., is not in a high vacuum), and is incompressible (i.e., the Mach number is less than approximately 0.3).

The law of conservation of mass (the continuity equation) requires that

$$\rho V = \rho_s V_s \tag{63.12}$$

In Equation 63.12, V is the actual velocity (m/s) measured at the midpoint of the heated section of the velocity sensor, often called the "point velocity." The subscript $()_s$ refers to standard, or base, conditions of standard temperature T_s and absolute pressure P_s. Many applications use standard conditions of T_s = 70 °F = 21.11 °C = 294.26 K and P_s = 1 atm = 101325 Pa = 1.01325 bara. Another class of applications uses standard conditions, often called "normal" conditions, of T_s = 0 °C = 273.15 K and P_s = 1 atm = 1.01325 bara. The constant standard mass density ρ_s (kg/standard m^3) for gases obeying a modified perfect gas law equation of state is calculated as follows:

$$\rho_s = \frac{P_s M}{z_s R T_s} \tag{63.13}$$

where

M is the molecular weight of the pure gas or gas mixture (kg/kg-mole)

R is the universal gas constant (m^3 bara/kg-mole/K)

z_s is the dimensionless compressibility of the gas at standard conditions

In Equation 63.13, P_s and T_s are the standard absolute pressure (bara) and the standard absolute temperature (K) of the gas, respectively.

Based on Equation 63.12, the Reynolds number Re defined in Equation 63.10 also can be expressed as

$$\mathrm{Re} = \frac{\rho_s V_s D}{\mu} \tag{63.14}$$

From Equations 63.10 to 63.14, we arrive at

$$V_s = \left(\frac{\mu}{\rho_s D}\right) F_2(\mathrm{Nu}, \mathrm{Pr}) \tag{63.15}$$

In the earlier text, V_s is called the point "mass velocity" (standard m/s) because when it is multiplied by the constant mass density ρ_s, it yields the mass flow rate per unit area. This is the mathematical reason why thermal dispersion mass flow meters directly measure mass flow rate.

In most insertion meter applications, the desired final output variable is the mass velocity V_s (standard m/s). In this case, V_s is calculated using Equation 63.15. If the desired output is the actual velocity V (m/s), then V is found from $V = (\rho_s/\rho)V_s$. If the desired final output is the total mass flow rate q_m (kg/s) flowing through a pipe or other flow conduit, then the following relationship is used:

$$q_m = \left(F_c A_{pipe}\right) \rho_s V_s \tag{63.16}$$

In Equation 63.16, A_{pipe} is the cross-sectional area (m^2) of the pipe or flow conduit; and the term F_c is defined as $F_c = V_{s,ave}/V_s$, where $V_{s,ave}$ is the average point mass velocity (standard m/s) over the cross-sectional area A_{pipe}. Most insertion meters are inserted such that the midpoint of the heated section of the velocity sensor is at the centerline of the pipe. In this case and with a fully developed velocity profile in the pipe, typical values of F_c are $F_c = 0.5$ for $\mathrm{Re}_{pipe} \leq 2{,}000$, and $F_c = 0.83 \pm 0.03$ for $\mathrm{Re}_{pipe} \geq 40{,}000$, where Re_{pipe} is the Reynolds number based on the *internal* diameter of the pipe [12]. For in-line meters, Equation 63.16 is not used explicitly. Instead, the factor $(F_c A_{pipe})$ becomes a meter factor determined via flow calibration. Tables 63.1 and 63.2 provide a list of the formulas used by the four-temperature microprocessor-based system to calculate the output variables of insertion meters and in-line meters, respectively.

TABLE 63.1 Calculation of Output Variables for Insertion Flow Meters

Variable	Units	Symbol	Formula
Point mass velocity, the primary output	Standard m/s	V_s	V_s from Equation 63.15 or from flow calibration
Actual velocity	m/s	V	$(\rho_s/\rho)V_s$
Mass flow rate	kg/s	q_m	$\rho_s A_{pipe} F_c V_s$
Mass flow rate, expressed as volumetric flow rate referenced to standard conditions $()_s$	Standard m^3/s	$q_{v,s}$	$A_{pipe} F_c V_s$
Volumetric flow rate	m^3/s	q_v	$(\rho_s/\rho) A_{pipe} F_c V_s$
Total mass flow rate for multipoint measurements in N equal areas	kg/s	q_m	$(\rho_s A_{pipe}/N) \sum_{i=1}^{N} V_{s,i}$

TABLE 63.2 Calculation of Output Variables for In-Line Flow Meters

Variable	Units	Symbol	Formula
Mass flow rate, the primary output	kg/s	q_m	q_m from flow calibration
Mass flow rate, expressed as volumetric flow rate referenced to standard conditions $(\ )_s$	Standard m^3/s	$q_{v,s}$	q_m/ρ_s
Volumetric flow rate	m^3/s	q_v	q_m/ρ

63.4 Dry-Sensor Microprocessor-Based System

Figure 63.4 shows a simplified block diagram of the digital microprocessor-based thermal dispersion mass flow meter with the dry flow sensor shown in Figure 63.2b. It is operated in the constant temperature differential mode. As before, we call this configuration the dry-sensor microprocessor-based system. The voltage sensing wires in Figure 63.4 make the measurement of the RTD resistances independent of the length of the flow sensor cable, facilitating remote location of the transmitter. The temperature sensing current I_2 is less than 1 mA to avoid self-heating the T_2 sensor. The analog layer shown in Figure 63.4 includes precision resistors for measuring the currents I_1 and I_2 and may include a bridge circuit.

The dry-sensor microprocessor-based system shown in Figure 63.4 digitally linearizes the q_m output and, optionally, the T output and provides analog outputs for these variables. The system also automatically corrects for changes in gas temperature. The system enables a number of digital functions, such as a multivariable digital readout and user interface; digital RS 232 and RS 485 communications; flow meter diagnostics, validation, calibration adjustment, and reconfiguration; flow totalization; and alarms.

Figure 63.5 shows typical outputs for the dry-sensor microprocessor-based system for air, methane, and argon. These curves are plotted in the conventional manner with the mass velocity V_s shown as

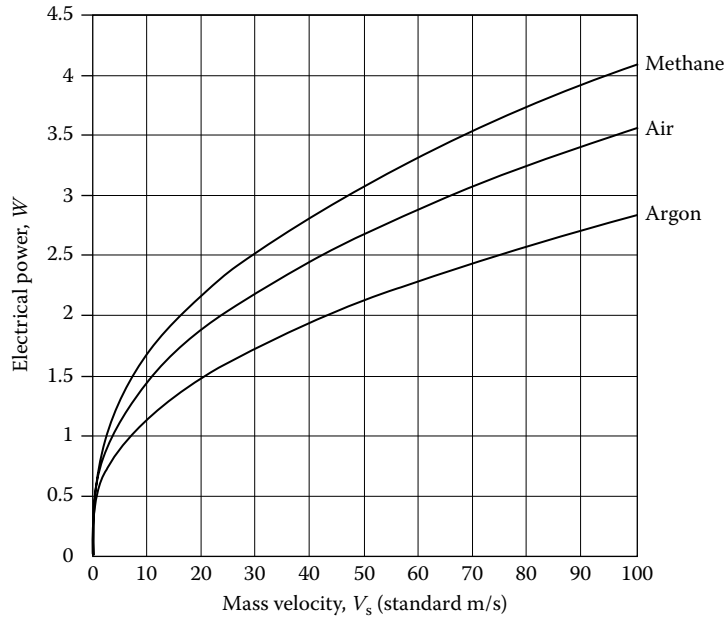


FIGURE 63.5 Typical outputs of the dry-sensor microprocessor-based system for three different gases. $T = 300$ K; $P = 1$ bara; $\Delta T = 50$ K. The "standard" conditions for V_s (standard m/s) are 70 °F and 1 atm. (Adapted from Olin, J.G., Advances in thermal dispersion mass flow meters—Part 1: Principle of operation, Sierra Instruments Technical Report No. SI-JGO-2010-1, 2010. With permission from Sierra Instruments, Inc.)

the independent variable and the electrical power W shown as the dependent variable, whereas in the system they have reversed roles. The three curves reflect the different values of thermal conductivity of the three gases, i.e., $k_{methane} > k_{air} > k_{argon}$.

Figure 63.5 also reveals the nonlinear, logarithmic nature of the output. Log versus Log plots of the three curves will reveal nearly straight lines over approximately 1–150 standard m/s. This logarithmic property is responsible for the exceptional rangeability and low-velocity sensitivity of thermal dispersion mass flow meters. A rangeability as high as 100:1 is common. Even higher rangeabilities are achieved with multirange flow calibration. Detectable minimum point mass velocities as low as approximately 0.1 standard m/s (20 standard ft/min) are reported by some manufacturers. In the early days of analog electronics, it was difficult to linearize the output of thermal dispersion mass flow meters. But now, with microprocessor-based electronics, it is not a problem, and the nonlinear, logarithmic nature of the output bears only advantages.

63.5 Flow Meter Selection, Installation, and Flow Calibration

In all cases, specifications and instructions provided by the manufacturer should be followed in selecting, sizing, and installing in-line and insertion thermal dispersion mass flow meters. Table 63.3 is helpful in the flow meter selection process. The specifications in Table 63.3 are typical of currently available thermal dispersion mass flow meters, but may vary from manufacturer to manufacturer. Accuracy specifications may apply to gas temperatures and gas pressures that lie within bands around their respective values at flow calibration. References [4,5,11] provide additional information regarding flow meter selection, installation, and flow calibration.

The vast majority of applications are gas flow applications because they benefit from the exceptional low-flow sensitivity and wide rangeability of measurement. Thermal dispersion technology is not well suited for liquid flow applications because at the zero-flow condition, a majority of the heat budget is carried away by the liquid via conduction, instead of the desired convection. This is caused

TABLE 63.3 Typical Specifications for Thermal Dispersion Mass Flow Meters

Quantity	Specification
Gases	Most clean gases, including air, methane, Ar, CO_2, He, N_2, O_2, propane, and mixtures of these components
In-line flow body sizes[a]	0.25, 0.5, 1.0, 1.5, 2.0, 4.0 in.; DN6 to DN100
In-line flow meter mass flow rate range for air[b]	0–1.5 kg/s; 0–2,600 scfm
Insertion flow meter mass velocity range for air[c]	0–140 normal m/s; 0–30,000 standard ft/min
Temperature range[d]	−40 °C to 200 °C; −40 °F to 392 °F
Pressure range	0.01–16 bara
Accuracy	±1% of reading plus 0.5% of full scale
Repeatability	±0.2% of full scale
Rangeability	100:1
Time response[e]	3 s (constant power operation); 1.2 s (constant ΔT operation)
Stability	1 year; typical drift 1%–2% per year (potted V-sensor) 10 years; typical drift 0.1% per year (dry V-sensor)

[a] Some manufacturers offer sizes up to 12.0 in. (DN300).
[b] Based on the point mass velocity ranges for insertion flow meters cited later.
[c] "Normal" conditions are 0°C and 1 atm, and "standard" conditions are 70°F and 1 atm.
[d] High-temperature models are available up to approximately 450 °C = 842 °F.
[e] Time response is the time required to reach 63% of the final value (i.e., the 1 sigma value).

by the high thermal conductivity of liquids relative to gases. The result is reduced measurement sensitivity for liquid flows. Additionally, for liquid flows, the temperature differential $\Delta T = T_1 - T$ must not exceed an upper critical limit, or else at higher flow rates, the liquid may flash to the vapor phase and subsequent cavitation may occur, creating unwanted erratic readings. For water flows, this upper critical limit in ΔT is approximately 10 °C–20 °C. The constant differential temperature mode of operation is preferred for liquid flows because ΔT is controlled, whereas in the constant current mode, ΔT varies and may exceed the upper critical limit. Application of thermal dispersion technology to liquid flows has been limited to cases, such as ultra-low flow applications, where it offers advantages over other technologies.

As with most kinds of flow meters, the performance of thermal dispersion mass flow meters can be degraded if the flow meter is installed where flow conditions are different than those for which it was flow calibrated. Improper installation is the single biggest cause of measurement inaccuracy for any kind of flow meter. Components in the piping system upstream and, to a far lesser extent, downstream of the flow meter can create nonuniformities in the flow profile, swirls, and turbulence. All of these phenomena degrade performance. Such flow-disturbing components include single or multiple elbows, expansions, contractions, tees, valves, and pumps. Fortunately, viscous forces in a sufficiently long length of straight pipe upstream and downstream of the flow meter reduce swirl and drive the flow toward a fully developed velocity profile. In-line flow meters with a built-in flow conditioner greatly reduce the need for long runs of straight pipe upstream of the flow meter. For example, the in-line flow meter shown in Figure 63.1a has a built-in flow conditioner upstream of the flow sensor consisting of two separated perforated plates. Flow meters with this built-in flow conditioner require only 1 diameter of straight pipe upstream when located downstream of a single elbow. In contrast, for this same case, an orifice plate flow meter with a 0.7 beta ratio requires 28 diameters of straight pipe upstream.

Because the critical dimensions of the flow sensor of thermal dispersion mass flow meters are so small, manufacturing technology is generally incapable of maintaining sufficiently small tolerances to ensure a high degree of reproducibility from flow sensor to flow sensor. Additionally, the internal diameters of the pipes used in in-line flow bodies have substantial variations. For these reasons, every general-purpose thermal dispersion mass flow meter is flow calibrated by the manufacturer, just like most other kinds of flow meters. Exceptions may include flow switches and low-accuracy flow meters.

The gas flow calibration facilities of manufacturers, flow calibration laboratories, and users should be capable of (1) generating a stable, steady-state, reproducible gas mass flow rate; (2) accommodating the entire mass flow rate range specified; (3) having a flow calibration standard that has an accuracy at least three times better than the flow meter under test; and (4) reproducing the gas composition, temperature, and pressure to be encountered in the actual application. Gas flow calibration facilities are of two types: open loop and closed loop.

Closed-loop facilities are preferred because they allow flow calibration at elevated pressures and temperatures and with gases other than air. A good pressurized closed-loop system has the following major components, listed in flow sequence: (1) a gas charging source, such as a compressed gas tank; (2) a flow source, such as a high pressure axial or centrifugal in-line pump; (3) a flow controller, such as a precision flow control valve; (4) a flow-conditioning section upstream of the flow calibration standard; (5) the in-line flow calibration standard (more than one may be needed to cover the mass flow rate range); (6) a flow-conditioning section upstream of the flow meter under test; and (7) the flow meter under test. Additionally, the facility should have accurate gas temperature and pressure instrumentation at both the flow calibration standard and the flow meter under test, pressure relief and venting components, and a vacuum pump for evacuating the system prior to charging. High-accuracy in-line flow calibration standards include custody-transfer grade multipath ultrasonic flow meters, turbine flow meters, flow nozzles, and positive-displacement flow meters.

63.5.1 Manufacturers

Following is a list of major manufacturers of thermal dispersion mass flow meters for general industrial applications. Details regarding their contact information, products, and current pricing are found in their websites.

- Sierra Instruments, Inc.; Monterey, California
- Endress and Hauser, Inc.; Reinach, Switzerland
- Fluid Components International, LLC; Costa Mesa, California

References

1. King, L. V. 1914. On the convection of heat from small cylinders in a stream of fluid: Determination of the convection constants of small platinum wires with application to hot-wire anemometry. *Phil. Trans. Roy. Soc*. A214: 373–432.
2. Agarwall, J. K. and J. G. Olin, Thermal anemometry, in Webster, J. G. and H. Eren (eds.), *The Measurement, Instrumentation and Sensors Handbook*, 2nd edn., CRC Press, Boca Raton, FL, 2013.
3. Olin, J. G. 2010. Advances in thermal dispersion mass flow meters—Part 1: Principle of operation. Sierra Instruments Technical Report No. SI-JGO-2010-1.
4. Olin, J. G. 2010. Advances in thermal dispersion mass flow meters—Part 2: Installation and flow calibration. Sierra Instruments Technical Report No. SI-JGO-2010-2.
5. ANSI/ASME MFC-21.2-2011. 2011. Measurement of fluid flow by means of thermal dispersion mass flowmeters. Washington, DC: ANSI/ASME.
6. ANSI/ASME MFC-9M, 2008, Measurement of fluid flow by means of capillary tube thermal mass flowmeters and controllers. Washington, DC: ANSI/ASME.
7. ANSI/ASME MFC-15M. 2001, Insertion metering. Washington, DC: ANSI/ASME.
8. Olin, J. G. 1993. A thermal mass flow monitor for continuous emissions monitoring systems (CEMS). *Proceedings of ISA/93 International Conference, Exhibition and Training Program* 93–404: 1635–1653.
9. Bruun, H. H. 1995. *Hot-Wire Anemometry: Principles and Signal Analysis*. Oxford, U.K.: Oxford University Press.
10. Olin, J. G. 1999. Industrial thermal mass flowmeters—Part 1: Principles of operation. *Measurement. Control* 193(February): 83–90.
11. Olin, J. G. 1999. Industrial thermal mass flowmeters—Part 2: Applications. *Measurement. Control* 194(April): 1–6.
12. Miller, R. W. 1996. *Flow Measurement Engineering Handbook*, 3rd edn. New York: McGraw-Hill.

64

Laser Anemometry

64.1 Principle of Operation **64**-3
Dual-Beam Approach • Frequency Shifting • Signal Strength • Measuring Multiple Components of Velocity
64.2 Signal Processing **64**-9
Nature of the Signal • Digital Signal Processing
64.3 Seeding and Other Aspects **64**-11
64.4 Data Analysis **64**-12
LDV Systems
64.5 Extension to Particle Sizing **64**-13
64.6 Phase Doppler System: Principle **64**-13
Fringe Model Explanation
64.7 Conclusion **64**-16
Acknowledgments **64**-16
References **64**-16
Further Information **64**-18

Rajan K. Menon
TSI Inc.

Laser anemometry, or *laser velocimetry,* refers to any technique that uses lasers to measure velocity. The most common approach uses the Doppler shift principle to measure the velocity of a flowing fluid at a point and is referred to as Laser Doppler Velocimetry (LDV) or Laser Doppler Anemometry (LDA). This technique (also known as dual beam, differential Doppler, or fringe mode technique), incorporating intersecting (focused) laser beams, is also used to measure the motion of surfaces, for example, in industrial applications [1]. In some special flow situations, another approach using two *nonintersecting,* focused laser beams known as *Laser two focus* (also known as Laser transit velocimetry or L2F) technique is used to measure flow velocity at a point [2]. Laser illumination by light sheets is used to make global flow measurements in a plane and is referred to as *Particle Image Velocimetry* (PIV) [3]. The strength of PIV (including particle tracking velocimetry) lies in its ability to capture instantaneous velocity fields, extract turbulence structures within the flow, observe transient phenomena, and examine unsteady flows. The development of this technique to obtain both spatial and temporal information about flow fields is making this a powerful diagnostic tool in fluid mechanics research [4–6]. More recently, using pulsed-laser volume illumination (in place of light sheet illumination), combined with multi-camera capture of particle images in a volumetric region in the flow, has provided the instantaneous velocity vector field in the three-dimensional domain [7]. Other approaches to measure global flow velocities come under the category of molecular tagging velocimetry [8] or Doppler global velocimetry [9].

The noninvasive nature of the LDV technique and its ability to make accurate velocity measurements with high spatial and temporal resolution, even in highly turbulent flows, have led to the widespread use of LDV for flow measurement. Flow velocities ranging from micrometers per second to hypersonic speeds have been measured using LDV systems. Measurements of flows in rotating machinery, including in the interblade region of rotors, flows in combustors and engines, flows at high temperatures, and flows in small spaces and in other hostile environments have been performed using the LDV technique [10–17].

The versatility and the widespread use of the LDV approach to measure flows accurately have resulted in referring to this technique as *laser velocimetry* or *laser anemometry*. Many details of the technique, including some of the early developments of the hardware, are provided in the book by Durst [18]. A bibliography of the landmark papers in LDV has been compiled by Adrian [19].

For the case of spherical scatterers, the technique has also been extended to measure the size of these particles. In this case, the scattered light signal collected by a suitably placed receiver system is converted into an electrical signal (by a photodetector) and processed to obtain the diameter of the particle, using the phase Doppler technique [20,21].

The first reported fluid flow measurements using LDV principles were by Yeh and Cummins [22]. Although in this case an optical arrangement referred to as the reference beam system was used to measure the Doppler shift, in almost all measurement applications, what is referred to as the dual beam or differential Doppler arrangement is used. This arrangement, also referred to as the "fringe" mode of operation, uses two intersecting laser beams (Figure 64.1) to measure one velocity component.

The advantages of the LDV technique in measuring flows include (1) a small measuring region (i.e., point measurement), (2) high measurement accuracy, (3) the ability to measure any desired velocity component, (4) accurate measurement of high-turbulence intensities, including flow reversals, (5) a large dynamic range, (6) no required velocity calibration, (7) no probe in the flow (does not disturb the flow; measures in hostile environments), and (8) good frequency response.

LDV technique relies on the light scattered by scattering centers in the fluid to measure flow velocity. These scattering centers will also be referred to as *particles,* with the understanding that bubbles or anything else that has a refractive index different from that of the fluid could be the source of scattered light. The particles, whose velocities are measured, must be small enough (generally in the micrometer range) to follow the flow variations and large enough to provide signal strength adequate for the signal processor to give velocity measurements. It should be noted that the signal exists only when a "detectable" particle is in the measuring volume and, hence, is discontinuous. This, along with other properties of the signal, places special requirements on the signal processing and the subsequent data analysis systems. The scattered light signal is processed to obtain the Doppler shift frequency and from that the velocity of

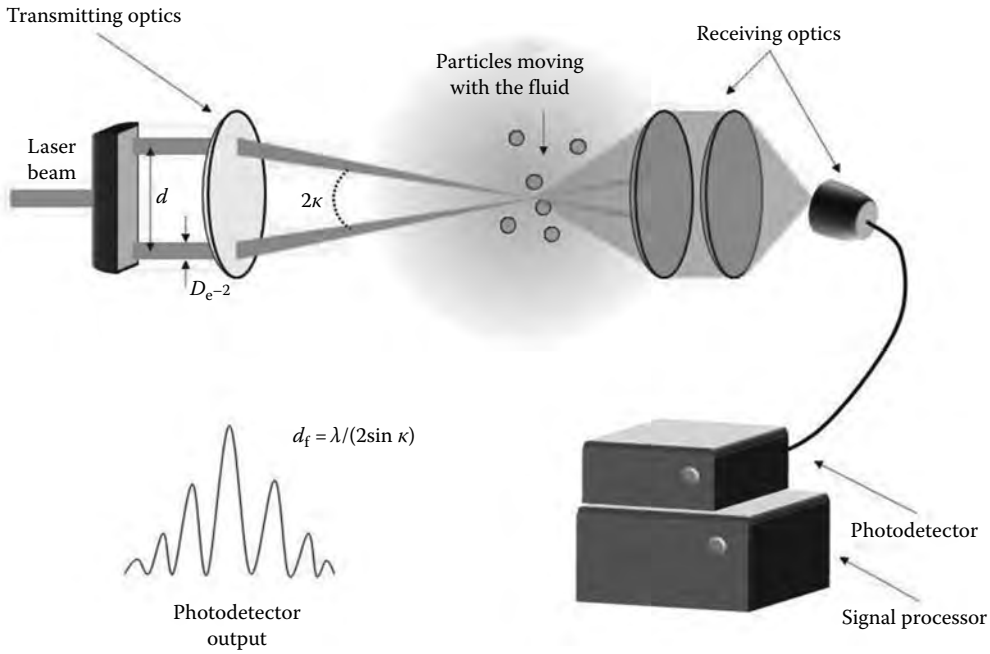


FIGURE 64.1 Schematic of a dual-beam system.

the particle. Hence, the rate at which the velocity measurements are made depends on the rate of particle arrival. It is desirable to have a high particle concentration to obtain a nearly continuous update of velocity. In carefully controlled experiments, the LDV system can provide very high accuracy (0.1% or better in mean velocity) measurements. Thermal anemometer systems are generally able to measure lower turbulence levels compared to that by an LDV system [23]. While the direct measurement of the Doppler-shifted frequency from a single laser beam caused by a moving particle is possible, most LDV systems employ the heterodyne principle to obtain and process only the Doppler shift (difference) frequency.

64.1 Principle of Operation

64.1.1 Dual-Beam Approach

The dual-beam approach is the most common optical arrangement used for LDV systems for flow measurement applications. The schematic (Figure 64.1) shows the basic components of a complete LDV system to measure one component of velocity. The transmitting optics includes an optical element to split the original laser beam into two parallel beams and a lens system to focus and cross the two beams. The intersection region of the two beams becomes the measuring region. The receiving optics (shown to be set up in the forward direction) collects a portion of the light scattered by the particles, in the fluid stream, passing through the beam-crossing region (measuring volume) and direct this light to a photodetector, which converts the scattered light intensity to an analog electrical signal. The frequency of this signal is proportional to the velocity of the particle. A signal processor extracts the frequency information from the photodetector output and provides this as a digital number corresponding to the instantaneous velocity of the particle. The data processing system obtains the detailed flow properties from these instantaneous velocity measurements. The idealized photodetector signal, for a particle passing through the center of the measuring volume, is shown in the lower left side of Figure 64.1. Actual signals will have noise superimposed on them, and the signal shape will vary, depending on the particle trajectory through the measuring volume [21,24].

64.1.1.1 Fringe Model Description

While there are several ways to describe the features of a dual-beam system, the description based on a fringe model is, perhaps, the simplest. For simplicity, the diameter and the intensity of both the beams are assumed to be the same. After the beams pass through the transmitting lens, the diameter of each beam continuously decreases to a minimum value (beam waist) at the focal point of the lens and then increases again. Thus, the beam waists cross where the two laser beams intersect (at the focal point of the lens), and the wavefronts in the beams interfere with each other, creating a fringe pattern [21]. In this pattern, assuming equal-intensity beams and other needed qualities of the beams, the light intensity varies from zero (dark fringe) to a maximum (bright fringe), and the fringes are equally spaced. The particles in the flow passing through the intersection region (measuring region) scatter light in all directions. An optical system, including a receiving lens (to collimate the scattered light collected) and a focusing lens, is used to collect the scattered light and focus it on to the detector. The aperture in front of the receiver is used to block out stray light and reflections and collect only the light scattered from the measuring region.

As a particle in the flow, with velocity u, moves across the fringes, the intensity pattern of the light scattered by the particle resembles that shown in the lower left of Figure 64.1. The velocity component, u_y (perpendicular to the optical axis and in the plane of the incident beams) can be obtained from the ratio of the distance between fringes (or fringe spacing, d_f), and the time t (=1/f_D) for the particle to cross one pair of fringes, where f_D is the frequency of the signal. The amplitude variation of the signal reflects the Gaussian intensity distribution across the laser beam. Collection (receiving) optics for the dual-beam system can be placed at any angle, and the resulting signal from the receiving system will still give the same frequency. However, signal quality and intensity will vary greatly with the collection optics angle.

64.1.1.2 Doppler Shift Explanation

The description of the dual-beam system using the Doppler shift principle is as follows. At the receiver, the frequencies of the Doppler-shifted light scattered by a particle from beam one and beam two are given by

$$\nu_{D1} = \nu_{01} + \frac{\vec{u}}{\lambda}\left(\hat{r} - \hat{S}_1\right); \quad \nu_{D2} = \nu_{02} + \frac{\vec{u}}{\lambda}\left(\hat{r} - \hat{S}_2\right) \tag{64.1}$$

where

ν_{01} and ν_{02} are the frequencies of laser beam 1 and laser beam 2, respectively
$\hat{r}$ is the unit vector directed from the measuring volume to the receiving optics
$\hat{S}_1$ and $\hat{S}_2$ are the unit vectors in the direction of incident beam 1 and incident beam 2, respectively
$\vec{u}$ is the velocity vector of the particle (scattering center)
λ is the wavelength of light

The frequency of the net (heterodyne) signal output from the photodetector system is given by the difference between ν_{D1} and ν_{D2}:

$$f_D = f_s + \frac{\vec{u}}{\lambda}\left(\hat{S}_2 - \hat{S}_1\right) \tag{64.2}$$

where $f_s = \nu_{01} - \nu_{02}$ is the difference in frequency between the two incident beams. This difference frequency is often intentionally imposed (see Section 64.1.2) to permit unambiguous measurement of flow direction and high-turbulence intensities. Assuming $f_s = 0$, the frequency detected by the photodetector is

$$f_D = \frac{\dot{u}}{\lambda}\left(\hat{S}_2 - \hat{S}_1\right) = 2u_y \sin k \tag{64.3}$$

Hence,

$$u_y = \frac{f_D \lambda}{2\sin k} = f_D d_f \tag{64.4}$$

This is the equation for u_y and shows that the signal frequency f_D is directly proportional to the velocity component u_y. The heterodyning of the scattered light from the two laser beams at the photodetector actually gives both the sum and difference frequency. However, the sum frequency is too high to be detected, and so only the difference frequency ($\nu_{D1} - \nu_{D2}$) is output from the photodetector as an electrical signal. The frequency f_D is often referred to as the Doppler frequency of the output signal, and the output signal is referred to as the Doppler signal.

It can be seen from Equation 64.4 that the Doppler frequency is independent of the receiver location ($\hat{r}$). Hence, the receiver system location can be chosen based on considerations such as signal strength, ease of alignment, and clear access to the measuring region. The expressions for the other optical configurations can be reduced similarly [24], giving the identical equation for the Doppler shift frequency f_D. It should be noted that the fringe description does not involve a "Doppler shift" and is, in fact, not always appropriate. The fringe model is convenient and gives the correct expression for the frequency. However, it can be misleading when studying the details of the Doppler signal (e.g., signal-to-noise ratio [SNR]) and other important parameters, for example, modulation depth or visibility ($\overline{V}$) of the signal [21,24].

The time taken by the particle to cross the measuring volume is referred to as *transit time, residence time,* or *total burst time,* τ_B, and corresponds to the duration of the scattered light signal. The number of cycles (N) in the signal (same as the number of fringes the particle crosses) is given by the product of the transit time (τ_B) and the frequency, f_D, of the signal.

It should also be noted that the fringe spacing (d_f) depends only on the wavelength of the laser light (λ) and the angle (2κ) between the two beams. It can be shown that the effect of the fluid refractive index on these two terms tends to cancel out, and hence, the value of fringe spacing is independent of the fluid medium [25]. The values of λ and κ are known for any dual-beam system, and hence, an actual velocity calibration is not needed. In some cases, an actual velocity calibration using the rim of a precisely controlled rotating wheel has been performed to overcome the errors in measuring accurately the angle between the beams.

The intensity distribution in a laser beam operating in the TEM_{00} mode is Gaussian. Using wave theory and assuming diffraction-limited optics, the effective diameter of the laser beam and the size of the measurement region can be defined. The conventional approach to the definition of laser beam diameter and measuring volume dimensions is based on the locations where the light intensity is $1/e^2$ of the maximum intensity (at the center of the beam). This definition of the dimensions is analogous to that of the boundary layer thickness. The dimensions d_m and l_m of the ellipsoidal measuring volume (Figure 64.2) are based on the $1/e^2$ criterion and are given by

$$d_m = \frac{4f\lambda}{\pi D_{e^{-2}}}; \quad l_m = \frac{d_m}{\tan\kappa}; \quad N_{FR} = \frac{d_m}{d_f} \tag{64.5}$$

N_{FR} is the maximum number of fringes in the ellipsoidal measuring region. Note that as the value of $D_{e^{-2}}$ increases, the measuring volume becomes smaller. In flow measurement applications, this relationship is exploited to arrive at the desired size of the measuring volume.

Consider the following sample optical arrangement of wavelength, $\lambda = 514.5$ nm (green line of argon-ion laser), $D_{e^{-2}} = 1.1$ mm, $d = 35$ mm, and $f = 250$ mm. Then, $\kappa = 4°$, and the measuring volume parameters are $d_m = 149$ μm, and $l_m = 2.13$ mm. The fringe spacing, d_f, is 3.67 μm and the maximum number (N_{FR}) of fringes (same as the number of cycles in a signal burst generated by a particle going through the center of the measuring volume in the y-direction) in the measuring volume is 40. Consider a particle

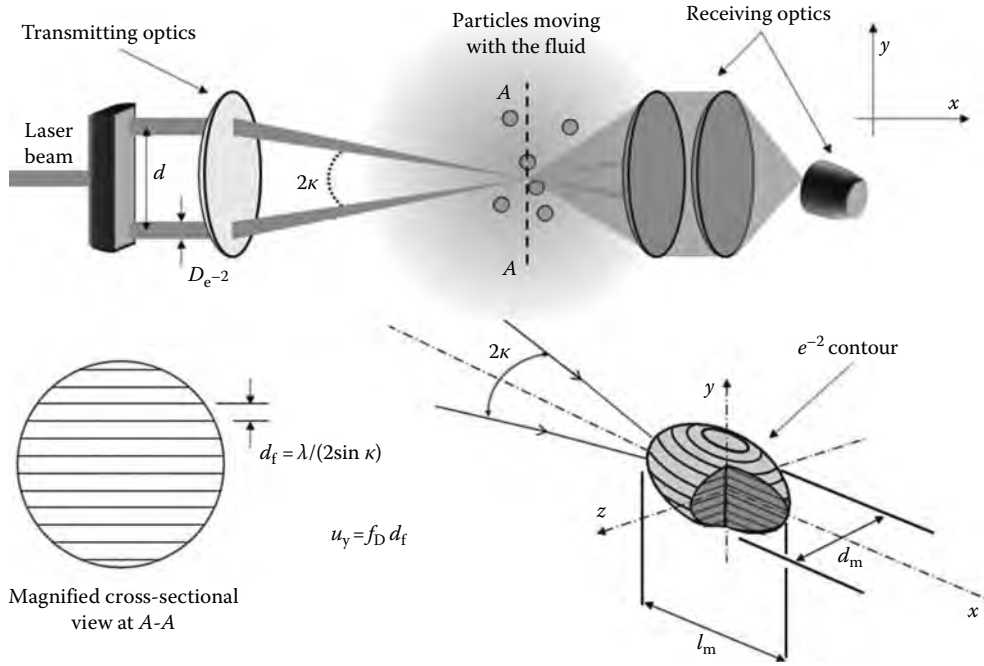


FIGURE 64.2 Details of the beam crossing.

passing through the center of the measuring region with a velocity (normal to the fringes) of 15 m s^{-1}. This would generate a signal with a frequency of 4.087 MHz. The transit time of the particle (same as the duration of the signal) would be approximately 9.93 μs!

64.1.2 Frequency Shifting

The presence of high-turbulence intensity and recirculating or oscillatory flow regions is common in most flow measuring situations. In the fringe model and the Doppler shift (with $f_s = 0$) descriptions of the dual-beam system, the Doppler signal does not indicate the influence of the sign (positive or negative) of the velocity. Further, a particle passing through the measuring volume parallel to the fringes would not cross any fringes and, hence, would not generate a signal having the cyclic pattern, resulting in the inability to measure the zero normal (to the fringes) component of velocity. In addition, signal processing hardware used to extract the frequency information often requires the signals to have a minimum number of cycles. This, as well as the ability to measure flow reversals, is achieved by a method of frequency offsetting referred to as *frequency shifting*. Frequency shifting is also used to measure small velocity components perpendicular to the dominant flow direction and to increase the effective velocity measuring range of the signal processors [25].

By introducing a phase or frequency offset (f_s) to one of the two beams in a dual-beam system, the directional ambiguity can be resolved. From the fringe model standpoint, this situation corresponds to a moving (instead of a stationary) fringe system. A stationary particle in the measuring volume will provide a continuous signal at the photodetector output whose frequency is equal to the difference in frequency, f_s, between the two incident beams. In other words, as shown in Figure 64.3b, the linear curve

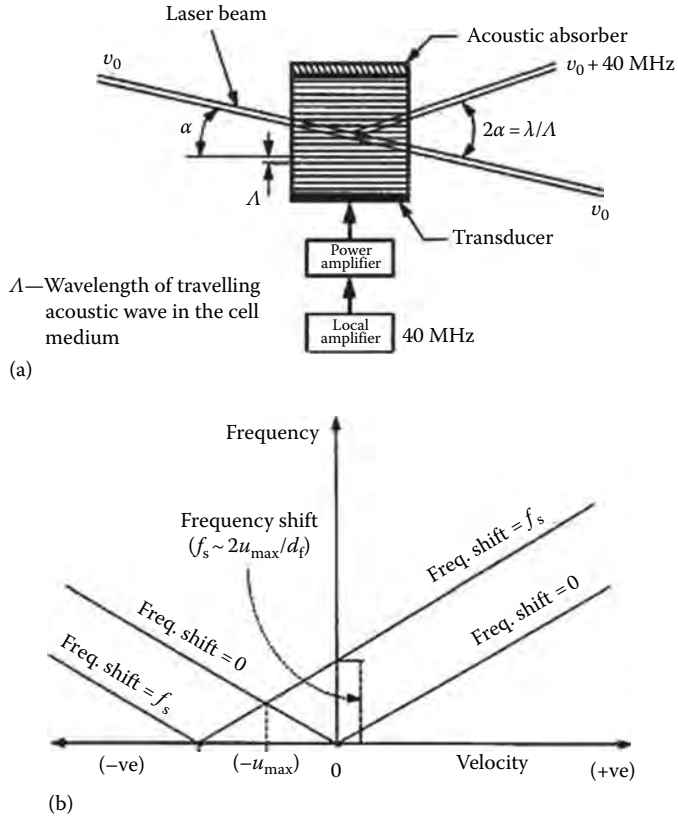


FIGURE 64.3 (a) Bragg cell arrangement; (b) velocity vs. frequency.

between velocity and frequency is offset along the positive frequency direction by an amount equal to the frequency shift, f_s. Motion of a particle in a direction opposite to fringe movement would provide an increase in signal frequency, while particle motion in the direction of fringe motion would provide a decrease in frequency. To create a signal with an adequate number of cycles even while measuring negative velocities (e.g., flow reversals, recirculating flows), a convenient "rule-of-thumb" approach for frequency shifting is often used. The approach is to select the frequency shift ($f_s \sim 2u_{max}/d_f$) to be approximately twice the frequency corresponding to the magnitude of the maximum negative velocity (u_{max}) expected in the flow. This provides approximately equal probability of measurement for all particle trajectories through the measuring volume [21].

Frequency shifting is most commonly achieved by sending the laser beam through a Bragg cell (Figure 64.3a), driven by an external oscillator. Typically, the propagation of the 40 MHz acoustic wave (created by a 40 MHz drive frequency) inside the cell affects the beam passing through the cell to yield a frequency shift of 40 MHz for that beam. By properly adjusting the angle the cell makes with the incoming beam and blocking off the unwanted beams, up to about 80% of input light intensity is recovered in the shifted beam. The Bragg cell approach will provide a 40 MHz frequency shift in the photodetector output signal. To improve the measurement resolution of the signal processor, the resulting photodetector signal is often "downmixed" to have a more appropriate frequency shift (based on the rule-of-thumb shift value) for the flow velocities being measured. Frequency shifting using two Bragg cells (one for each beam of a dual-beam system) operating at different frequencies is attractive to systems where the bandwidth of the photodetector is limited. However, the need to readjust the beam crossing with a change in frequency shift has not made this approach (double Bragg cell technique) attractive for applications where frequency shift needs to be varied [25].

More recently, Bragg cells have been used in a multifunctional mode to split the incoming laser beam into two equal-intensity beams, with one of them having the 40 MHz frequency shift. This is accomplished by adjusting the Bragg cell angle differently. In addition to Bragg cells, rotating diffraction gratings and other mechanical approaches have been used for frequency shifting. However, limits on rotational speed and other mechanical aspects of these systems make them limited in frequency range [18]. Other frequency-shifting techniques also have been suggested for use with LDV [21]. Because so many flow measurement applications involve recirculating regions and high-turbulence intensities, frequency shifting is almost always a part of an LDV system used for flow measurement.

64.1.3 Signal Strength

Understanding the influence of various parameters of an LDV system on the SNR of the photodetector signal provides methods or approaches to enhance signal quality and hence improve the performance of the measuring system. The basic equation for the ratio of signal power to noise power (SNR) of the photodetector signal can be written as [24]

$$\mathrm{SNR} = A_1 \frac{\eta_q P_0}{\Delta f}\left[\frac{D_a}{r_a}\frac{D_{e^{-2}}}{f}\right]^2 d_p^2 \bar{G}\bar{V}^2 \tag{64.6}$$

Equation 64.6 shows that higher laser power (P_0) provides better signal quality. The quantum efficiency of the photodetector, η_q, depends on the type of photodetector used and is generally fixed. The SNR is inversely proportional to the bandwidth, Δf, of the Doppler signal. The term in brackets relates to the optical parameters of the system: the "f-number" of the receiving optics, D_a/r_a, and parameters associated with the transmitting optics, $D_{e^{-2}}/f$. The square dependence of SNR on these parameters makes them the prime choice for improving signal quality and, hence, measurement accuracy. The focal length of the transmitting (f) and receiving (r_a) lenses is generally decided by the size of the flow facility. Using the smallest possible values for these would

increase the signal quality. The first ratio (D_a is the diameter of the receiving lens) determines the amount of the scattered light that is collected, and the second ratio determines the diameter of (and hence the light intensity in) the measuring volume. The last three terms are the diameter, d_p, of the scattering center and the two terms (scattering gain $\overline{G}$, visibility $\overline{V}$) relating to properties of the scattered light. These need to be evaluated using the Mie scattering equations [24] or the generalized Lorenz–Mie theory [21].

64.1.4 Measuring Multiple Components of Velocity

A pair of intersecting laser beams is needed to measure (Figure 64.1) one component of velocity. This concept is extended to measure two components of velocity (perpendicular to the optical axis) by having two pairs of beams that have an overlapping intersection region. In this case, the plane of each pair of beams is set to be orthogonal to that of the other. The most common approach to measure two components of velocity (such a system is referred to as a two-component system) is to use a laser source that can generate multiwavelength beams so that the wavelength of one pair of beams is different from that of the other pair. The Doppler signals corresponding to the two components of velocity are separated by wavelength [21,24].

Historically, LDV systems were assembled by putting together a variety of optical modules. These modules included beam splitters, color separators, polarization rotators, and scattered light collection systems. The size of such a modular system depended on the number of velocity components to be measured.

The use of optical fibers along with multifunctional optical elements has made the systems more compact, flexible, and easier to make measurements. The laser, optics to generate the necessary number of beams (typically, one pair per component of velocity to be measured), photodetectors, and electronics can be isolated from the measurement location. The fibers carrying the laser beams thus generated are arranged in the probe to achieve the desired beam geometry for measuring the velocity components. Hence, flow field mapping is achieved by moving only the fiber-optic probes, while keeping the rest of the system stationary. To achieve maximum power transmission efficiency and beam quality, special single-mode, polarization-preserving optical fibers along with precision couplers are used. In most cases, these fiber probes also have a receiving system and a separate fiber (multimode) to collect (in back scatter) the scattered light and carry that back to the photodetector system [21].

A schematic arrangement of a fiber probe system to measure one component of velocity is shown in Figure 64.4a. In flow measurement applications, LDV systems using these types of fiber-optic probes have largely replaced the earlier modular systems.

The best way to make three-component velocity measurements is to use an arrangement using two probes—a two-component probe and a single-component probe. In this case, the obvious orientation of the optical axis of the system to measure the third component of velocity (u_x) would be perpendicular to that of the two-component system. Unfortunately, access and/or traversing difficulties, often, make this arrangement impractical or less attractive. In most practical situations, the angle between the two probes is selected to be less than 90° [11,16,21].

Another three-component LDV system configuration uses a single fiber-optic probe by effectively combining the laser beams associated with the two-probe arrangement. This system incorporating a single-lens coaxial probe, with five laser beams, is used to make measurements in regions with limited access or map the velocity vector fields in liquid flows in channels and tunnels [25,26].

More recently, the availability of solid-state lasers with reasonable power levels and good beam quality has resulted in configuring compact LDV systems. These compact systems (Figure 64.4b) have the lasers integrated with the required optics to measure one or two components of velocity, simultaneously. Such an arrangement using a diode laser to measure two components of velocity, simultaneously, is shown in Figure 64.5.

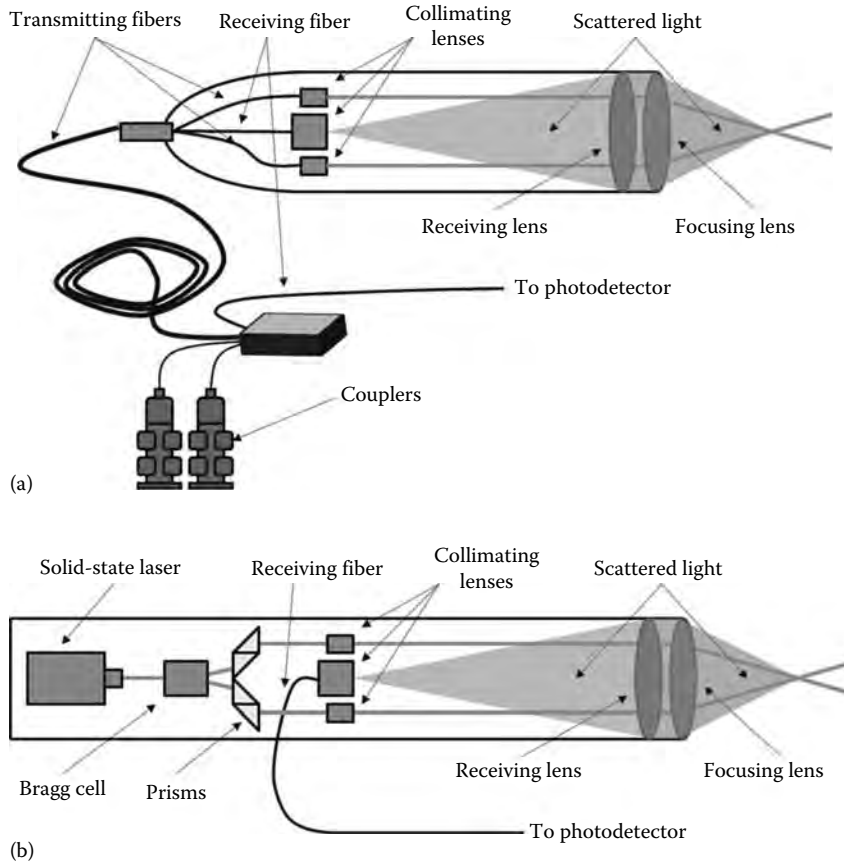


FIGURE 64.4 (a) Schematic arrangement of a fiber-optic system; (b) schematic arrangement of a diode laser-based system.



FIGURE 64.5 Two-component LDV system with diode laser and integrated optics.

64.2 Signal Processing

64.2.1 Nature of the Signal

Every time a particle passes through the measuring region, the scattered light signal level (Figure 64.6) suddenly increases ("burst"). The characteristics of the burst signal are (1) amplitude in the burst not constant, (2) lasts for only a short duration, (3) amplitude varies from burst to burst, (4) presence of noise, (5) high frequency, and (6) random arrival.

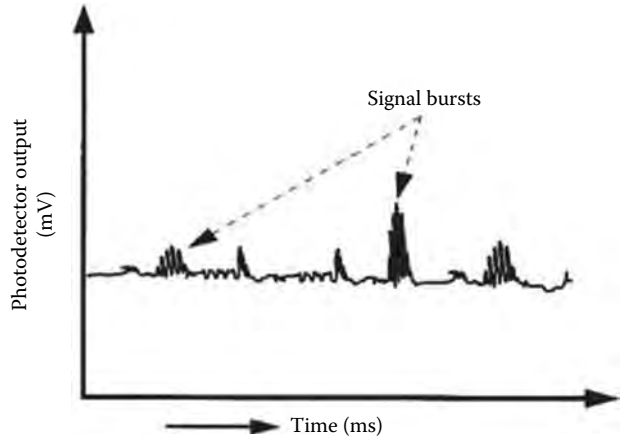


FIGURE 64.6 Time history of the photodetector signal.

The primary task of the LDV signal processor is to extract the frequency information from the burst signal generated by a particle passing through the measuring volume and provide an analog or digital output proportional to the frequency of the signal. The unique nature of the signals demands the use of a special signal processing system to extract the velocity information.

A variety of techniques have been used for processing Doppler signals. Signal processors have been based on spectrum analysis, frequency tracking, photon correlation, frequency counting, Fourier transform, and autocorrelation principles. The evolution of the signal processing techniques shows the improvement in their ability to handle more difficult measuring situations (generally implies noisier signals), give more accurate measurements, and have higher processing speed.

The traditional instrument to measure signal frequency is a spectrum analyzer. The need to measure individual particle velocities, and to obtain the time history and other properties of the flow, has eliminated the use of "standard" spectrum analyzers [24]. The "tracker" can be thought of as a fixed-bandwidth filter that "tracked" the Doppler frequency as the fluid velocity changed. This technique of "tracking flow" worked quite well at modest velocities and where the concentration of scattering centers was high enough to provide an essentially continuous signal. However, too frequently, these conditions could not be met in the flows of most interest.

When the scattered light level is very low, the photodetector output reveals the presence of the individual photon pulses. By correlating the actual photon pulses from a wide bandwidth photodetector, the photon correlator was designed to work in situations where the attainable signal intensity was very low (low SNR). However, as normally used, it could not provide the velocity of individual particles but only the averaged quantities, such as mean velocity and turbulence intensity.

The "counter" type processor was developed next and basically measured the time for a certain number (typically, eight) of cycles of the Doppler signal. Although it measured the velocity of individual particles, its performance depended on the careful setting of amplifier gain and, especially, threshold levels to discriminate burst signals from background noise. Counters were the processors of choice for many years, and excellent measurements were obtained. However, the reliance on user skill, difficulty in handling low SNR signals, possibility of getting erroneous measurements, inclination to ignore signals from small particles, and desire to make measurements close to surfaces and in complex flows led to the need for a better signal processor.

64.2.2 Digital Signal Processing

The latest development in signal processing is in the area of digital signal processors. Developments in high-speed digital signal processing permit the use of these techniques to extract the frequency, from individual Doppler bursts, fast enough to actually follow the flow when the seeding concentration is adequate,

in a wide range of measurement situations. By digitizing the incoming signal and using the Fourier transform or autocorrelation algorithms, these new digital processors can work with lower SNR signals (than counters), while generally avoiding erroneous data outputs [21,27,28]. While instruments using these techniques are certainly not new, standard instruments were not designed to make rapid individual measurements on the noisy, short-duration burst signals with varying amplitudes that are typical of Doppler bursts.

Because the flow velocity and hence the signal frequency vary from one burst to the next, the sampling rate needs to be varied accordingly. And because the signal frequency is not known *a priori,* the ability to optimally sample the signal has been one of the most important challenges in digital signal processing. In one of the digital signal processors, the question of deciding the sample rate is addressed by a burst detector that uses SNR to identify the presence of a signal [28]. In addition, the burst detector provides the duration and an approximate estimate of the frequency of each of the burst signals. This frequency estimate is used to select the output of the sampler (from the many samplers) that had sampled the burst signal at close to the optimum rate. Besides optimizing the sample rate for each burst, the burst detector information is also used to focus on and process the middle portion of the burst where the SNR is maximum. These optimization schemes, followed by digital signal processing, provide an accurate digital output that is proportional to the signal frequency, and hence the fluid velocity and detailed flow statistics [29].

64.3 Seeding and Other Aspects

The performance of an LDV system can be significantly improved by optimizing the source of the signal, the scattering particle. The first reaction of many experimentalists is to rely on the particles naturally present in the flow. There are a few situations (e.g., LDV systems operating in forward scatter to measure water or liquid flow velocities) where the particles naturally present in the flow are sufficient in number and size to provide good signal quality and, hence, good measurements. In most flow measurement situations, particles are added to the flow (generally referred to as seeding the flow) to obtain an adequate number of suitable scatterers. Use of a proper particle can result in orders-of-magnitude increase in signal quality (SNR), and, hence, can have greater impact on signal quality than the modification of any other component in the LDV system. Ideally, the seed particles should be naturally buoyant in the fluid, provide adequate scattered light intensity, have large enough number-concentration, and have uniform properties from particle to particle. While this ideal is difficult to achieve, adequate particle sources and distribution systems have been developed [21,24,30–32].

LDV measurements of internal flows such as in channels, pipes, and combustion chambers result in the laser beams (as well as the scattered light) going through transparent walls or "windows." In many cases, the window is flat and, hence, the effect of light refraction can be a simple displacement of the measuring region. In the case of internal flows with curved walls, each beam can refract by different amounts, and the location of the measuring region needs to be carefully estimated [33]. For internal flows in models with complex geometries, the beam access needs to be carefully selected so that the beams do cross inside. Further, to make measurements close to the wall in an internal flow, the refraction effect of the wall material on the beam path needs to be minimized. One of the approaches is to use a liquid [21,34] that has the same refractive index as that of the wall material.

To make velocity measurements near a boundary or a wall, light reflected from these background surfaces need to be reduced. Fluorescent seed particles [21,30] would generate, as they pass through the measuring region, a signal that is different in wavelength from that of the laser light and, hence, the reflected light. Thus, the impact of reflected light can be minimized by the use of receiver systems with proper optical filters (band pass) that collect the fluoresced wavelength light and filter out other wavelengths. In some cases, surfaces were painted with a fluorescent paint to reduce the influence of reflected light [35].

Properties of the scattered light signal have also been used in discriminate phases (in two-phase flows) to obtain velocity distribution of different phases or distinguish large and small particles. Fluorescent seed particles were used in two-phase flows to discriminate the phases [45], whereas signal amplitude [46] was used to separate gas and particle phases and get the corresponding velocity distributions.

64.4 Data Analysis

The flow velocity is "sampled" by the particle passing through the measuring volume, and the velocity measurement is obtained only when the Doppler signal, created by the particle, is processed and output as a data point by the signal processor. While averaging the measurements to get, for example, mean velocity would seem reasonable, this method may give the wrong answer. This arises from the fact that the number of particles going through the measuring region per unit time is higher at high velocities than at low velocities. In effect, there is a correlation between the measured quantity (velocity) and the sampling process (particle arrival). Hence, a simple average of the data points will bias the mean value (and other statistical parameters) toward the high-velocity end and is referred to as *velocity bias* [36]. The magnitude of the bias error depends on the magnitude of the velocity variations about the mean. If the variations in velocity are sufficiently small, the error might not be significant.

If the actual data rate is so high that the output data are essentially able to characterize the flow (time history), then the output can be sampled at uniform time increments. This is similar to the procedure normally used for sampling a continuous analog signal using an ADC. This will give the proper value for both the mean and the variance when the data rate is sufficiently high compared to the rates based on the Taylor microscale for the temporal variation of velocity. This is referred to as a high data density situation [24].

In many measurement situations, the data rate is not high enough (low data density) to actually characterize the flow. Here, sampling the output of the signal processor at uniform time increments will not work because the probability of getting an updated velocity (new data point) is higher at high velocity than at low velocity. The solution to the velocity bias problem is to weight the individual measurements with a factor inversely proportional to the probability of making the measurement:

$$\bar{U} = \frac{\Sigma u_j \tau_{B_j}}{\Sigma \tau_{B_j}} \tag{64.7}$$

where

u_j is the velocity of particle j

τ_{Bj} is the transit time for particle j

Similar procedures can be used to obtain unbiased estimators for variance and other statistical properties of the flow [24,17]. Modern signal processors provide the residence time and the time between data points along with the velocity data. Comparison of some of the different approaches to do bias correction and other types of biases associated with LDV has been summarized [17,21,37].

A variety of techniques to obtain spectral information of the flow velocity from the random data output of the signal processors have been tried. The goal of all these techniques has been to get accurate and unbiased spectral information to as high a frequency as possible. Direct spectral estimation of the digital output of the processors exhibits the spectrum estimates at high frequency to be less reliable. The "slotting" technique of estimating the autocorrelation of the (random) velocity data followed by Fourier transform continues to be attractive from a computational standpoint. To obtain reliable spectrum estimates at high frequencies, a variety of methods aimed at interpolation of measured velocity values have been attempted. These are generally referred to as *data* or *signal reconstruction* techniques [17,21].

64.4.1 LDV Systems

Tables 64.1 and 64.2 list examples of typical commercial LDV systems. The research systems are generally configured to suit the application requirements, including the need to do flow field mapping. This results in wide variation of possible light sources, signal processors, data analysis packages, and seeding systems. Hence, the system price can vary significantly, and the numbers shown in Table 64.2 are only indicative.

TABLE 64.1 Manufacturers of Laser Doppler Velocimeter Systems

Systems for research and industrial applications

TSI Inc., 500 Cardigan Road, Shoreview, MN 55126, U.S.A., Tel.: 800-874-2811, Fax: 651-490-3824, http://www.tsi.com/Fluid-Mechanic-Systems/

Dantec Dynamics A/S, Tonsbakken 16–18, P.O. Box 121, DK-2740 Skovlunde, Denmark, Tel: +45 44 57 80 00, Fax: +45 44 57 80 01, http://www.dantecdynamics.com/

Miniature LDV systems

Measurement Science Enterprise, Inc., 123 West Bellevue Drive, Suite 1, Pasadena, California 91105-2549, USA, Tel.: 626-577-0566, Fax: 626-577-0565, http://measurementsci.com/index.html

TABLE 64.2 Typical Commercial Laser Doppler Velocimeter Systems

One-velocity-component LDV system	$60,000–90,000
Two-velocity-component LDV system	$100,000–130,000
Three-velocity-component LDV system	$150,000–190,000

Note: The price of the systems varies significantly depending upon the type of lasers, signal processors, and other components of the system.

64.5 Extension to Particle Sizing

In LDV, the frequency of the scattered light signal provides the velocity of the scatterer. Processing the scattered light to get information about the scatterer other than velocity has always been a topic of great interest in flow and particle diagnostics. One of the most promising developments is the extension of the LDV technique to measure the surface curvature and, hence, the diameter of a spherical scatterer [20]. This approach (limited to spherical particles) uses the phase information of the scattered light signal to extract the size information. To obtain a unique and, preferably, monotonic relation between phase of the signal and the size of the particle, the orientation and the geometry (aperture) of the scattered light collection system need to be carefully chosen. In the following, unless otherwise mentioned, the particles are assumed to be spherical.

The light scattered by a particle, generally, contains contributions from the different scattering mechanisms—reflection, refraction, diffraction, and internal reflection(s). It can be shown that, by selecting the position of the scattered light collection set-up, contributions from one scattering mechanism can be made dominant over the others. The aim in phase Doppler measurements is to have the orientation of the receiver system such that the scattered light collected is from one dominant scattering mechanism.

The popularity of the technique is evidenced by its widespread use for measuring particle diameter and velocity in a large number of applications, especially in the field of liquid sprays [21,38,43]. The technique has also been used in diagnosing flow fields associated with combustion, cavitation, manufacturing processes, and other two-phase flows.

64.6 Phase Doppler System: Principle

The phase Doppler approach, outlined as an extension to an LDV system, was first proposed by Durst and Zare [39] to measure velocity and size of spherical particles. The first practical phase Doppler systems using a single receiver were proposed by Bachalo and Houser [20].

A schematic arrangement of a phase Doppler system is shown in Figure 64.7a. This shows a receiver system arrangement that collects, separates, and focuses the scattered light onto multiple photodetectors. In general, the receiving system aperture is divided into three parts, and the scattered light collected through these is focused into three separate photodetectors. For simplicity, in Figure 64.7a, the output of two detectors is shown. The different spatial locations of the detectors (receiving apertures)

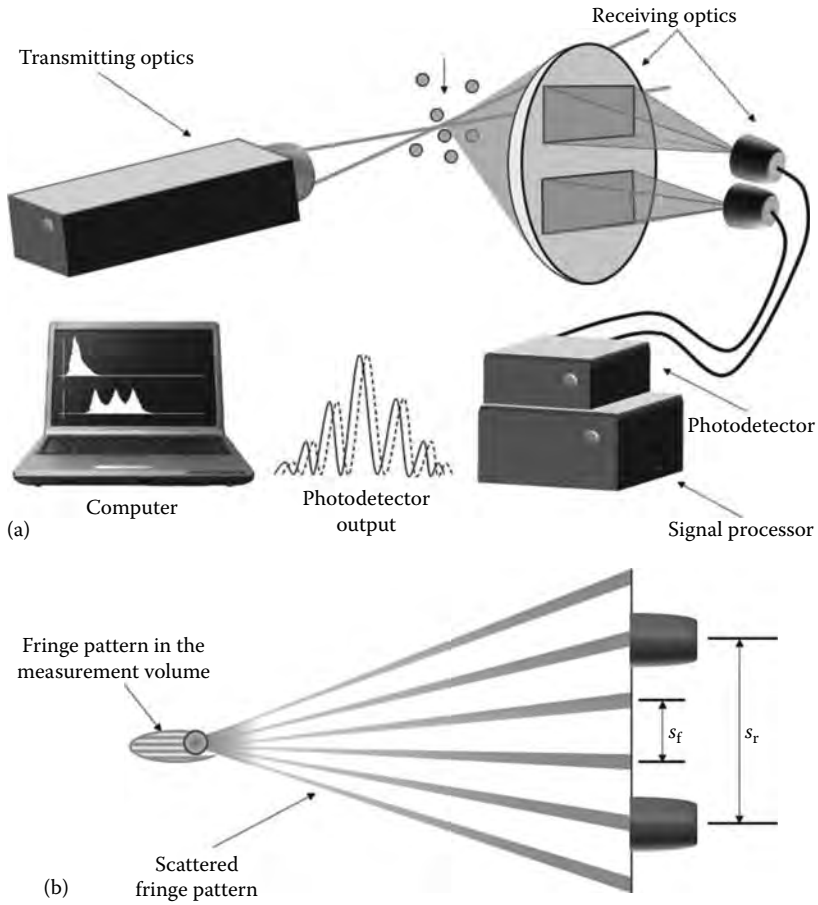


FIGURE 64.7 (a) Phase Doppler system: schematic; (b) phase Doppler system: schematic of the scattered fringes (fringe model).

result in the signals received by each detector having a slightly different phase. In general, the difference in phase between the signals from the detectors is used to obtain the particle diameter, whereas the signal frequency provides the velocity of the particle.

64.6.1 Fringe Model Explanation

The fringe model provides an easy and straightforward approach to arrive at the expressions for Doppler frequency and phase shift created by a particle going through the measuring volume. When the particle moves through fringes in the measuring volume, it "scatters a fringe pattern" (Figure 64.7b). The phase shift in the signals can be arrived at by examining the scattered fringe pattern. If the particle acts like a spherical mirror (dominant reflection) or a spherical lens (dominant refraction), it projects fringes from the measuring volume into space all around as diverging bands of bright and dark light, known as *scattered fringes*. The scattered fringe pattern is collected through the apertures by the receiver system, as shown in Figure 64.7b. The spacing between the scattered fringes at the plane of the receiver is s_f, whereas the spacing between a pair of apertures is denoted by s_r.

The receiver system in Figure 64.7b shows two apertures. The distance between (separation) the centroids of the two receiving apertures is s_r. Scattered fringes move across the apertures in the receiver, as the particle moves through the measuring volume. The scattered fringe pattern thus collected through

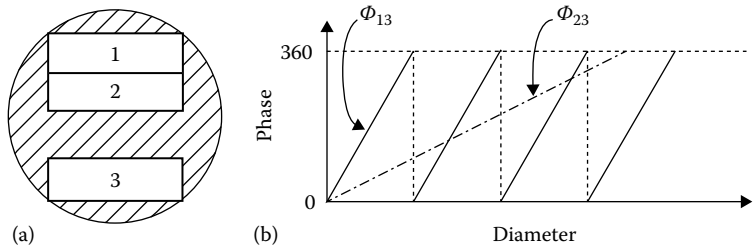


FIGURE 64.8 (a) Three-detector configuration; (b) phase–diameter relationship.

the apertures in the receiver is converted, using photodetectors, into electrical signals. These would be temporally fluctuating signals. The two photodetector output signals are shifted in phase by s_r/s_f times 360° [25]. Large particles create a scattered fringe pattern with a smaller fringe spacing (compared to that for small particles), that is, particle diameter is inversely proportional to s_f, while s_f is inversely proportional to phase difference. Thus, the fringe model shows the particle diameter to be directly proportional to the phase difference. It can also be seen that the sensitivity (degrees of phase difference per micrometer) of the phase Doppler system can be increased by increasing the separation (s_r) between the detectors.

The phase Doppler system shown earlier measures the phase difference between two detectors in the receiver system to obtain particle diameter. This brings in the limitation that the maximum value of phase that could be measured is 2π. A three-detector (using three apertures) arrangement in the receiver system is used to overcome this 2π ambiguity. Figure 64.8 shows the three-detector (aperture) arrangement. Scattered light collected through apertures 1, 2, and 3 is focused into detectors 1, 2, and 3. Φ_{13} is the phase difference between the detectors 1 and 3 and provides the higher phase sensitivity because of their greater separation compared to detectors 2 and 3. As Φ_{13} exceeds 2π, the value of Φ_{23} is below 360° and is used to keep track of Φ_{13}. It should be noted that the simplified approach in terms of geometrical scattering provides a linear relationship between the phase difference and diameter of the particle.

It has been pointed out that significant errors in measured size can occur due to trajectory-dependent scattering [21]. These errors could be minimized by choosing the appropriate optical configuration of the phase Doppler system [21,40]. Intensity-validation technique is a simple processing approach to minimize the errors [41]. Other methods such Dual-mode phase Doppler [21] have also been used to address this problem.

To explore the fundamental physical limits on the applicability of the phase Doppler technique, a rigorous model based on the electromagnetic theory of light has been developed. Computational results based on Mie scattering and comparison with and limitations of the geometric scattering approach have also been outlined [42]. By integrating a phase Doppler velocimeter system with a *rainbow refractometer* system, the velocity, size, and the refractive index of a droplet could be determined [21].

The values of particle velocity and diameter are obtained by processing the photodetector output signals. In general, the signal processing system for velocity measurements is expanded to measure the phase difference between two photodetector signals. The digital signal processing approaches described earlier have been complimented by the addition of accurate phase measurement techniques to get velocity and diameter information, simultaneously, for each particle [38].

Phase Doppler technique has been used to diagnose detailed behavior and characteristics of various different sprays [21,38,43]. This technique has also been used to examine flow and droplet motion in complex flows such as the interaction of an aerated-liquid jet exiting into a cross-stream supersonic flow [44].

Although the phase Doppler technique is limited to the domain of spherical particles, there has always been an interest in extending the technique to size nonspherical particles. In the past, symmetry checks and other similar techniques have been used to ensure the sphericity of particles. Sizing irregular particles is a more complex problem because the local radius of curvature is not a meaningful concept in these cases. Past efforts include going from a single particle approach to using stochastic models for the signals from irregular-shaped and inhomogeneous particles [21].

64.7 Conclusion

LDV has become the preferred technique for measuring flow velocity in a wide range of applications. The ability to measure noninvasively the velocity, without calibration, of any transparent flowing fluid has made it attractive for measuring almost any type of flow. Velocity measurement of moving surfaces by LDV is used to monitor and control industrial processes. Use of laser diodes, fiber optics, and advances in signal processing and data analysis is reducing both the cost and complexity of measuring systems. This noncontact measurement system's (as opposed to mechanical systems) ability to give reliable, repeatable, and accurate measurements has made it attractive for calibrating other devices [47].

The extension of LDV to the phase-Doppler technique provides an attractive, noncontact method for measuring size and velocity of spherical particles. Recent developments in the phase Doppler technique have generated a method to size submicrometer particles as well. These ideas have been extended to examine irregular particles also.

Acknowledgments

The input and comments from Dr. L. M. Fingerson have been valuable in the preparation of this chapter.

References

1. D. Niccum, A new tool for fiber spinning process control and diagnostics, *Int. Fiber J.*, 10(1), 48–57, 1995.
2. Meyers, J. F., Lee, J. W., and Schwartz, R. J., Characterization of measurement error sources in Doppler global velocimetry. *Measure. Sci. Technol.*, 12(4), 357–368, 2001.
3. R. J. Adrian, Particle imaging techniques for experimental fluid mechanics, *Annu. Rev. Fluid Mech.*, 23, 261–304, 1991.
4. I. Grant, *Selected Papers in Particle Image Velocimetry,* SPIE Milestone Series, MS 99, Bellingham, WA: SPIE Optical Engineering Press, 1994.
5. R. J. Adrian, *Bibliography of Particle Velocimetry Using Imaging Methods: 1917–1995,* TAM Report, University of Illinois Urbana-Champaign, Produced and distributed in cooperation with TSI Inc., March 1996, 2009 version is available at http://25-years-piv.dlr.de/wiki/Public/Article/Page/step1.php?sid=ae9f351a9ee34ca47e84f5ed8b10580451434adfc1bc8 (accessed on March 15, 2013).
6. R. J. Adrian and J. Westerweel, *Particle Image Velocimetry*, Cambridge, U.K.: Cambridge University Press, 2010.
7. E. Wolf, C. J. Kähler, D. R. Troolin, C. Kykal, and W. Lai, Time-resolved volumetric particle tracking velocimetry of large-scale vortex structures from the reattachment region of a laminar separation bubble to the wake, *Exp. Fluids*, Online first, September 16, 1–12, 2010, http://www.springerlink.com/content/577833210v421t46/
8. D. G. Bohl and M. M. Koochesfahani, MTV measurements of the vortical field in the wake of an airfoil oscillating at high reduced frequency, *J. Fluid Mech.*, 620, 63–88, 2009.
9. J. F. Meyers, J. W. Lee, and A. A. Cavone, Boundary layer measurements in a supersonic wind tunnel using Doppler global velocimetry, *Proceedings of 15th International Symposium on Applications of Laser Techniques to Fluid Mechanics,* Lisbon, Portugal, July 05–08, 2010.
10. A. Hamed, W. Tabakoff, and R. Wenglarz, Erosion and deposition in turbomachinery, *J. Propulsion Power*, 22(2), 350–360, March–April 2006.
11. P. B. Martin, G. Pugliese, and J. G. Leishman, High resolution trailing vortex measurements in the wake of a hovering rotor, *J. Am. Helicopter Soc.*, 49(1), 39–52, 2004.
12. J. Colby, S. Menon, and J. Jagoda, Flow field measurements in a counter-swirling spray combustor, *AIAA Paper AIAA-2005-4143, 41st AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit*, Tucson, AZ, July 10–13, 2005.

13. K. Jaffri, H. G. Hascher, M. Novak, K. Lee, H. Schock, M. Bonne, and P. Keller, Tumble and swirl quantification within a four-valve SI engine cylinder based on 3D LDV measurements, *SAE*, Paper No. 970792, February 1997.
14. F. Payri, J. M. Desantes, and J. V. Pastor, LDV measurements of the flow inside the combustion chamber of a 4-valve D.I. diesel engine with axisymmetric piston-bowls, *Exp. Fluids*, 22(2), 118–128, 1996.
15. J. Furukawa, Y. Noguchi, and T. Hirano, Investigation of flame generated turbulence in a large-scale and low-intensity turbulent premixed flame with a 3-element electrostatic probe and a 2-D LDV, *Combust. Sci. Technol*., 154(1), 163–178, 2000.
16. R. Semaan and J. W. Naughton, Three component LDA measurements in the near and far field of swirling jets, *AIAA Paper AIAA 2008-761, 46th AIAA Aerospace Sciences Meeting and Exhibit*, Reno, NV, January 7–10, 2008.
17. Z. Zhang, *LDA Application Methods: Laser Doppler Anemometry for Fluid Dynamics*, Berlin, Germany: Springer, 2010.
18. F. Durst, A. Melling, and J. H. Whitelaw, *Principles and Practice of Laser Doppler Anemometry,* 2nd edn., New York: Academic Press, 1981.
19. R. J. Adrian (ed.), *Selected Papers on Laser Doppler Velocimetry,* SPIE Milestone Series, MS 78, Bellingham, WA: SPIE Optical Engineering Press, 1993.
20. W. D. Bachalo and M. J. Houser, Phase Doppler spray analyzer for simultaneous measurements of drop size and velocity distributions, *Opt. Eng*., 23, 583–590, 1984.
21. H. E. Albrecht, M. Borys, N. Damaschke, and C. Tropea, *Laser Doppler and Phase Doppler Measurement Techniques,* Berlin, Germany: Springer, 2003.
22. Y. Yeh and H. Z. Cummins, Localized fluid flow measurements with an He-Ne laser spectrometer, *Appl. Phys. Lett*., 4, 176–178, 1964.
23. L. M. Fingerson and P. Freymuth, Thermal anemometers, in *Fluid Mechanics Measurements,* 2nd edn., R. J. Goldstein (ed.), Washington, DC: Taylor & Francis Publishing, pp. 115–170, 1996.
24. R. J. Adrian, Laser velocimetry, in *Fluid Mechanics Measurements,* 2nd edn., R. J. Goldstein (ed.), Washington, DC: Taylor & Francis Publishing, pp. 175–300, 1996.
25. L. M. Fingerson, R. J. Adrian, R. K. Menon, S. L. Kaufman, and A. Naqwi, Data analysis, in *Laser Doppler Velocimetry and Particle Image Velocimetry, TSI Short Course Text*, St. Paul, MN: TSI Inc., 1993.
26. P. A. Chevrin, H. L. Petrie, and S. Deutsch, Accuracy of a three-component laser Doppler velocimeter system using a single lens approach, *J. Fluids Eng*., 115, 142–147, 1993.
27. K. M. Ibrahim, G. D. Werthimer, and W. D. Bachalo, Signal processing considerations for laser Doppler and phase Doppler applications, *Proceedings of 5th International Symposium on Applications Laser Techniques Fluid Mechanics,* Lisbon, Portugal, 1990.
28. L. Jenson, LDV digital signal processor based on Autocorrelation, *Proceedings of 6th International Symposium on Applications Laser Techniques Fluid Mechanics,* Lisbon, Portugal, 1992.
29. M. P. Schultz and K. A. Flack, The rough-wall turbulent boundary layer from the hydraulically smooth to the fully rough regime, *J. Fluid Mech*., 580, 381–405, 2007.
30. R. K. Menon and W. T. Lai, Key considerations in the selection of seed particles for LDV measurements, *Proceedings of the 4th International Conference on Laser Anemometry,* Cleveland, OH, August 1991.
31. M. F. Reeder, J. W. Crafton, J. Estevadeordal, J. DeLapp, C. McNiel, D. Peltier, and T. Reynolds, Clean seeding for flow visualization and velocimetry measurements, *Exp. Fluids*, 48(5), 889–900, 2010.
32. Y. Ikeda, M. Nishigaki, M. Ippommatsu, and T. Nakajima, Optimum seeding particles for successful laser Doppler velocimeter measurements, *Part. Part. Syst. Charact*., 11, 127–132, 1994.
33. M. L. Lowe, and P. H. Kutt, Refraction through cylindrical tubes, *Exp. Fluids*, 13, 315–320, 1992.
34. R. Budwig, Refractive index matching methods for liquid flow investigations, *Exp. Fluids,* 17, 350–355, 1994.
35. Q. Tian, Near wall behavior of vortical flow around the tip of an axial pump rotor blade, PhD thesis, Aerospace Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA, 2006, http://scholar.lib.vt.edu/theses/available/etd-12122006-112751/unrestricted/TianDissertation.pdf

36. D. K. McLaughlin and W. G. Tiederman, Biasing correction for individual realization of laser anemometer measurements in turbulent flows, *Phys. Fluids*, 16, 2082–2088, 1973.
37. R. D. Gould and K. W. Loseke, A comparison of four velocity bias correction techniques in laser Doppler velocimetry, *J. Fluids Eng*., 115, 508–514, 1993.
38. J. D. Schwarzkopf, J. S. Shakal, and C. Bonuccelli, A new method for minimizing volumetric flux errors associated with PDPA measurements in the dilute region of full cone pressure swirl atomizers, *ICLASS 2006*, Kyoto, Japan, Paper ID ICLASS06-163, August 2006.
39. F. Durst and M. Zare, Laser Doppler measurements in two-phase flows, *The Accuracy of Flow Measurements by Laser Doppler Methods,* Skovulunde, Denmark: Dantec Measurement Technology, pp. 480–489, 1976.
40. A. Naqwi, Innovative phase Doppler systems and their applications, *Part. Part. Syst. Charact*., 11, 7–21, 1994.
41. S. V. Sankar, D. A. Robart, and W. D. Bachalo, An adaptive intensity validation technique for minimizing trajectory dependent scattering errors in phase Doppler interferometry, *Proceedings of the 4th International Congress Optical Particle Sizing,* Nuremberg, Germany, March 1995.
42. A. Naqwi and F. Durst, Light scattering applied to LDA and PDA measurements. 2. Computational results and their discussion, *Part. Part. Syst. Charact*., 9, 66–80, 1992.
43. J. Benajes, R. Payri, S. Molina, and V. Soare, Restore desktop view investigation of the influence of injection rate shaping on the spray characteristics in a diesel common rail system equipped with a piston amplifier, *J. Fluids Eng*., 127(6), 1102–1111, 2005.
44. K.-C. Lin, P. J. Kennedy, and T. A. Jackson, Structures of water jets in a Mach 1.94 supersonic crossflow *42nd AIAA Aerospace Sciences Meeting and Exhibit*, AIAA Paper, pp. 12096–12115, 2004.
45. G. Rottenkolber, R. Meier, O. Schafer, S. Wachter, and K. Dullenkopf, Phase discrimination inside a spray: LDV measurements using fluorescent seeding particles (FLDV*)*, *Laser Techniques for Fluid Mechanics*, R. J. Adrian, D. F. G. Durao, F. Durst, M. V. Heitor, M. Maeda, and J. H. Whitelaw (eds.), Berlin, Germany: Springer-Verlag, pp. 511–527, 2002.
46. A. Mychkowsky, D. Rangarajan, and S. Ceccio, LDV Techniques to simultaneously measure gas and particulate phase velocities in a jet plume in a 2D bubbling fluidized bed, *7th International Conference on Multiphase Flows, ICMF 2010*, May 30–June 4, 2010.
47. K. Shirai, L. Buttner, J. Czarske, and C. Kykal, A new calibration method and device for certified flow measurements with laser velocimetry, *Proceedings of the ASME 2012 International Mechanical Engineering Congress & Exposition, IMECE2012-93100*, November 9–15, 2012, Houston, TX.

Further Information

Buchave, P., W. K. George, and J. L. Lumley, The measurement of turbulence with the laser Doppler anemometer, *Annu. Rev. Fluid Mech*., 11, 443–504, 1979.

Drain, L. E., *The Laser Doppler Technique,* New York: John Wiley & Sons, 1980.

Greated, C. A. and T. S. Durrani, *Laser Systems and Flow Measurement*, New York: Plenum, 1977.

Naqwi, A. and F. Durst, Light scattering applied to LDA and PDA measurements. 1. Theory and numerical treatments, *Part. Part. Syst. Characteriz*., 8, 245–258, 1991.

Proceedings, *Proceedings of the International Symposium* (1 to 16) *on Applications of Laser Techniques to Fluid Mechanics*, Lisbon, Portugal, 1982–2012, http://ltces.dem.ist.utl.pt/lxlaser/

Proceedings, *International Conference* (1 to 8) *on Laser Anemometry—Advances and Applications*, 1985–1999.

C. Tropea, Cameron, A. L. Yarin, J. Foss (eds.), *Handbook of Experimental Fluid Mechanics*, Berlin, Germany: Springer, 2007.

VII

Thermal and Temperature Measurement

65 **Temperature Measurements, Scales, and Calibrations** *Franco Pavese* **65**-1
Definitions of Temperature • Temperature Scales • Thermodynamic Temperature Scale • Empirical Temperature Scales • International Temperature Scales • Differences between the Realized Thermodynamic Temperature Scale and the T_{90} • Calibration of Thermometers • References

66 **Thermal Conductivity Measurement** *William A. Wakeham and Marc J. Assael*..... **66**-1
Introduction • Fundamental Equations • Measurement Techniques • Instrumentation • Appraisal • References

67 **Heat Flux** *Thomas E. Diller*..... **67**-1
Heat Transfer Fundamentals • Heat Flux Measurement • Sensors Based on Spatial Temperature Gradient • Sensors Based on Temperature Change with Time • Measurements Based on Surface Heating • Calibration and Problems to Avoid • Summary • References

68 **Resistive Thermometers** *Jim Burns*..... **68**-1
Introduction • How an RTD Works? • RTD Construction • Classifications of RTDs • Applications • Calibration and Maintenance • Conclusion • Defining Terms • References

69 **Thermistor Thermometers** *Rod White and Meyer Sapoff*..... **69**-1
Principles of Operation • Calibration Equations • Instrumentation • Limitations in performance • Choice of Thermistor • Thermistors in Other Applications • References

70 **Thermocouple Thermometry** *R.P. Reed* **70**-1
Introduction • Terminology • Simplest Thermocouple Thermometer • Thermoelectricity • Basic Thermocouple Circuits • Functional Model of Thermoelectric Circuits • Applications of the Functional Model • Thermocouple Characteristics • Thermocouple Hardware • Quality Assurance by Users • Thermocouple Instrumentation Systems • Thermocouple Thermometry Practice • Information Sources for Thermoelectric Theory and Thermometry • Caveat • Summary • References

71 **Semiconductor Junction Thermometers** *Randy Frank*..... **71**-1
Transistor as a Temperature Sensor • Thermal Properties of Semiconductors: Defining Equations • Integrated Temperature Sensors • Common Applications

of Semiconductor Sensing • Temperature Sensing in Power ICs for Fault Protection and Diagnostics • Reliability Implications of Temperature to Electronic Components • Semiconductor Temperature Sensor Packaging • Defining Terms • Vendors and Suppliers • References

72 **Noncontact Thermometers** *Jacob Fraden* **72**-1
Thermal Radiation (Physical Laws) • Emissivity • Blackbody • Detectors for Thermal Radiation • Pyrometers • IR Thermometry • Special Applications • References

73 **Pyroelectric Detectors** *Jacob Fraden* **73**-1
Pyroelectric Effect • Pyroelectric Materials • Manufacturing Process • Pyroelectric Sensors • Applications • References • Further Readings

74 **Liquid-in-Glass Thermometers** *Rod White and J. V. Nicholas* **74**-1
General Description • Choice of Thermometers • Operation • Errors in Liquid-in-Glass Thermometry • References • Partial List of Vendors and Suppliers

75 **Manometric Thermometers** *Franco Pavese* **75**-1
Vapor Pressure • Gas Thermometry • References

76 **Temperature Indicators** *Jan Stasiek, Tolestyn Madaj, and Jaroslaw Mikielewicz* **76**-1
Melting and Shape/Size Changing Temperature Indicators • Color-Change Temperature Indicators • References • Further Readings • Partial List of Vendors and Suppliers

77 **Fiber-Optic Thermometers** *Brian Culshaw* **77**-1
Introduction • Fiber-Optic Temperature Sensors • Fiber-Optic Point Temperature Measurement Systems • Distributed and Quasi-Distributed Optical Fiber Temperature Measurement Systems • Application for Optical Fiber Temperature Probes • Further Information • Further Readings • Partial List of Vendors and Suppliers

78 **Thermal Imaging** *Herbert M. Runciman* **78**-1
Essential Components • Thermal Imaging Wave Bands • Emission from Source • Atmospheric Transmission • Detectors • Electronics • Optics and Scanning • Temperature References • Imager Performance • Available Imagers • Performance Trade-Offs • Future Trends in Thermal Imaging • References • Further Information

79 **Calorimetry Measurement** *Sander van Herwaarden and Elina Iervolino* **79**-1
Introduction • Calorimeters: Criteria for Their Classification • Scanning Calorimetry Sweeps the Experiment Temperature • Calibration of Calorimeters • Typical Applications of Calorimeters • Thermal Analysis of Materials and Their Behavior with Temperature • Choosing the Proper Calorimeter for an Application • Can the Instrument of Choice Measure the Signals Desired? • Commercially Available Calorimeters • References • Further Information

65

Temperature Measurements, Scales, and Calibrations

65.1 Definitions of Temperature **65**-1
65.2 Temperature Scales **65**-3
65.3 Thermodynamic Temperature Scale **65**-4
65.4 Empirical Temperature Scales **65**-6
Semi-Empirical Scales • Strictly Empirical Scales
65.5 International Temperature Scales **65**-9
"*Mise en pratique*" for the Definition of the kelvin
65.6 Differences between the Realized Thermodynamic Temperature Scale and the T_{90} **65**-10
65.7 Calibration of Thermometers **65**-14
References **65**-16

Franco Pavese
National Institute of Meterology Research

65.1 Definitions of Temperature

The modern definition of temperature according to classical thermodynamics stems from Carnot's well-known *closed reversible cycle*, which must be formed by alternating isotherms and adiabats. A minimum of *two* temperatures is necessary to construct the cycle.

The Carnot cycle is represented in Figure 65.1, with the p–V coordinates originally used by Carnot. Assume that there are two heat reservoirs able to provide any required amount of heat isothermally, at temperatures θ_1 and $\theta_2 > \theta_1$. The working substance absorbs an amount of heat Q_2 from the hot reservoir in expanding isothermally from A to B; from B to C, it does not exchange heat, but performs mechanical work W_2 until it cools down to θ_1; the working substance then undergoes an isothermal reduction in volume from C to D, releasing an amount of heat Q_1 to the cold reservoir; finally, the working substance is returned to its original state A and temperature θ_2 by reversible compression utilizing mechanical work W_1 without any exchange of heat. If the two isotherms differ by an infinitesimal amount $\mathrm{d}\theta$, and the engine produces a total net work of $\mathrm{d}W$ over the full cycle, the amount of heat Q that is transferred from the higher to the lower reservoir may be shown to be such that

$$\frac{đ\,W}{Q} = \mu(\theta)\mathrm{d}\theta \tag{65.1}$$

where the differential $\mathrm{d}W$ is not exact, since the total work performed along a path depends on the path, and $\mu(\theta)$ is a function of temperature θ *only*—the so-called Carnot function. A Carnot engine does *not* depend on the substance used in performing the work. For a cycle performed between two

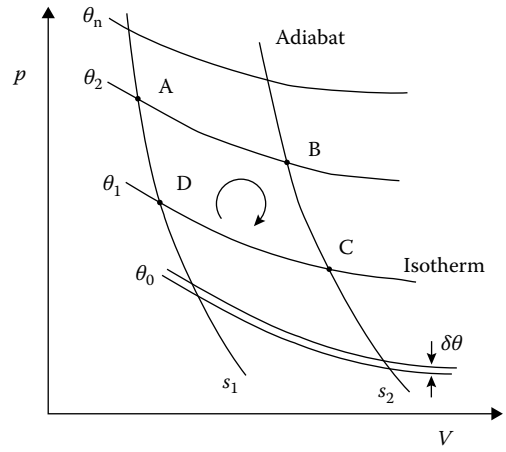


FIGURE 65.1 The Carnot cycle (ABCD), in the p–V representation.

temperatures θ_1 and θ_2 with $\theta_2 > \theta_1$, Equation 65.1 defines the efficiency η equal to W/Q_2 or to $(Q_2 - Q_1)/Q_2$. Then there is a choice of defining the (absolute) temperature in terms of the mechanical work or the thermal effect of heat.

Thomson (Lord Kelvin) made the first recognized attempt to define absolute temperature both ways, based on the theory initiated by Carnot that $\mu(\theta)$ in Equation 65.1 is "universal," as required by an *absolute* temperature. The first of his definitions [1] states that a unit *difference* of absolute temperature precisely equals the maximum *work* that can be obtained from *a unit of heat let down in a cyclic process,* irrespective of the temperature values involved, or

$$W = (\Theta_2^* - \Theta_1^*) \cdot Q \tag{65.2}$$

which is equivalent to setting $\mu(\Theta^*) = 1$ in Equation 65.1. Using this definition, it would be $\theta = \Theta^*$. Thomson assumed a "reference" temperature $\Theta_0^* = 0$. In general, the definition holds for $\mu(\Theta^*) = const$. Therefore, temperature intervals are proportioned to increments of efficiency. For an arbitrary "reference" temperature Θ_0^* and any Θ_n^*, one can write $Q_n/Q_0 = \mu(\Theta^*)^n$, or $\Theta^* = \Theta_0^* + \log(Q/Q_0)/\mu(\Theta^*)$. When $\Theta^* \to \pm\infty$, $\eta \to 1$. In his second definition [2], upon the suggestion of Joule, who found from experiments on compression of air that $\mu(\theta)$ was rather proportional to $1/\theta$, he preferred to change the definition to one that can be derived from Carnot's finite cycle of Figure 65.1:

$$\frac{Q_2}{Q_1} = \frac{\Theta_2}{\Theta_1} \tag{65.3}$$

The use of Θ instead of Θ^* marks the fact that this definition of temperature is different from the one in Equation 65.2. If chosen as the temperature definition, $\theta = \Theta$ would hold. Notice that definition in Equation 65.3 does not place any dimensional constraint on temperature, since only temperature *ratios* are defined, whereas Equation 65.2 does, since it is using temperature *differences*. The relationship between the two definitions is given by $\theta^* = J \log\theta + const$, where J is a constant defined as the *mechanical equivalent of a unit of heat*. Equation 65.3 is the modern definition of thermodynamic temperature, $\Theta = T$, where temperature is assumed proportional to the heat exchanged and where $\Theta \geq 0$. Notice that, instead, Equation 65.2 is equivalent to assuming that temperature increases by the same *proportion* as the corresponding increase in the gas volume, that is, $\Theta^* \propto \ln(V/V_0)$ and the earlier restriction for Θ does not hold. Figure 65.2 pictures the intrinsic logarithmic nature of temperature.

Other definitions of temperature (axiomatic, from statistical thermodynamics, kinetic theory, quantum theory, etc.) can be found in [3].

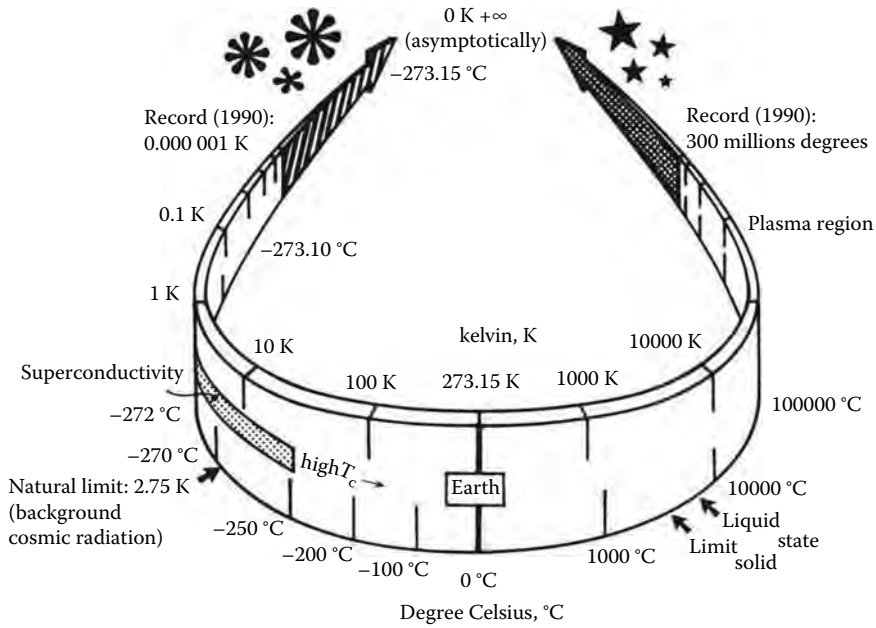


FIGURE 65.2 The logarithmic natural scale of temperatures.

65.2 Temperature Scales

Temperature cannot be measured in the same way as other fundamental quantities, for example, length. The unit size of the degree having been defined, it cannot subsequently be labeled "unit interval" and used to measure temperature in the same way as the meter in length measurements. That is, an additive procedure cannot be used for temperature, by which its value is determined from the number of the "unit intervals" contained in it. Temperature values, instead, can only be determined by comparing two temperatures, one of which is the reference, and by observing whether they are equal—or which one is higher—by observing, in accordance with the zeroth law of thermodynamics, whether there is—or not—a heat flow, and by noting its direction. To assign a numerical value to each temperature, one has to first "order" the measured temperatures, that is, to establish a *scale* in which the heat flows always in the same direction, and then *assign* a sign to the flow. Accordingly, it can be said that $\{\ldots T_1 > T_2 > T_3 \ldots\}$: temperature is a simply ordered manifold. By this procedure, one cannot yet assign a numerical value to different temperatures, but only give them a serial number that will be altered by the arbitrary addition of any new measurement. Nor can one say yet that the value of any T_2 is closer to that of T_1 than to that of T_3 (in fact, one cannot assign a distance between two temperatures), even if he or she should decide to take the interval, say $\{T_1, T_2\}$, as the unit interval, because neither intervals nor ratios can be compared by means of heat flow measurements.

As mentioned before, the definitions of "absolute" temperature are given only as ratios or differences of values. Therefore, normalization constants are required to make the relationship of temperature with another physical quantity determined. These definitions may be said to provide a "blank tape," which is necessary for temperature measurement, but leaves free the choice as regards both the *size* of the graduation in unit intervals and the *position* of the "zero" value on the tape. These are the two degrees of freedom characteristic of any *linear* scale. In the case, for instance, of the unit of length, the *meter*, it is appropriate to define it in such a way as to maintain the additive property in the classical isotropic space. In this respect, in the absolute temperature definition of Equation 65.3, the space of heat—that is, of energy—is used, whereas in his first definition (Equation 65.2), Lord Kelvin used the space of mechanical work equivalent to heat. The transformation between the two definitions is not linear, but logarithmic, as a reflection of the fact that the first defines a temperature ratio and the second a temperature difference.

The definition in the space of heat has been preferred because it closely approximates the numerical values obtained with gaseous substances—already experimentally available at Kelvin's times—and, more specifically, because it matches *exactly* the physical properties of a model state of the matter called "the ideal gas," more or less approximated by real gaseous low-density substances. This limit model has unique characteristics assuming that the "gas" constituents are particles not subjected to any external force field (not even to gravity) and not interacting with each other, but only subjected to perfectly elastic—nondissipative—collisions; the model is independent of the size of the atoms and of the scale. In the classical ideal-gas theory, particles are in fact not characterized by any parameter, namely, mass, but they are identical, so indistinguishable from each other. Starting from the modern generalization of the model of *quantum* ideal gas, they satisfy the equation of state

$$pV = gT \tag{65.4}$$

where g is a constant [4].

In the case of the classical ideal gas, g is defined by Rn, (n being the amount of substance and R is the gas constant) or by kN (N being the number of particles (entities) and k the Boltzmann constant). Equation 65.4 *alone* defines the ideal gas model for absolute temperature. Actually, the plural form "ideal *gases*" is often found. For an empirical temperature θ, on the contrary, the definition of the model requires, in addition to Equation 65.4, the condition $(\delta U/\delta V)\theta = 0$, which is no longer implied. The reason is that other conditions are assumed to be added to the basic model. For example, a common (but *unnecessary*) condition is the definition of a relationship $c(T)$ for the specific heat (e.g., $c(T) = const$). Or, of a "structure," for example, molecular, with an additional kinetic energy associated to each noninteracting particle. Therefore, we may have different *gases*, depending on whether they are monoatomic or polyatomic, have a spin, and so on. In such cases, they can be considered as the limit condition for vanishing intermolecular and external interactions of real gas species. The ideal gas model was fully developed in the second half of the nineteenth century on the basis of the kinetic theory of gases. To summarize, it is because temperature is defined in the heat (energy) space that the absolute lower end of the temperature scale is a zero (only asymptotically approached) and not an (negative) infinity.

65.3 Thermodynamic Temperature Scale

First, let us remark again that only an empirical temperature θ can be measured. The aim of this section is to show how a usable relationship can be established between a scale based on θ and the "thermodynamic" temperature T. Second, it must be pointed out again that none of the definitions of temperature given in the preceding section defines a unique scale. Each only puts constraints on the scale form. An example is given in Equation 65.2, where the values of two constants can be chosen arbitrarily by setting the zero point in the scale (Kelvin first definition) and the size of the degree. As a matter of curiosity, with θ = 273.16 K and with the old value 373.16 K for the boiling point of water, one would have $\theta^*/[\text{unit}] = 738 \cdot \log T/\text{K} - 1798$. On this scale, the value of the tungsten melting point (3673 K) is 833 [units], smaller in absolute value than that of the normal boiling point of ^{4}He (4.2 K), which would be –1336 [units].

Kelvin's second definition Equation 65.3 was used as the basis for the first internationally accepted temperature scale at the first meeting of the Conférence Générale des Poids et Mesures (CGPM) held in 1887 [5]; it was the "normal hydrogen scale" using a hydrogen gas thermometer developed by Chappuis at the BIPM. Temperature $\theta := t$ was defined with the use of *two* fixed points T_1 and T_0, and was assumed to be proportional to the pressure increase (at constant volume), $(p - p_0)/p_0 = \alpha t$. Pressure has a defined value p_0 at the first fixed point T_0, the ice point, which was assumed to be 0 °C (this symbol, and the term "degree Celsius" later adopted in 1948, replaced the *previous*, "degree centigrade"). The second fixed point T_1, necessary to define α, was the normal boiling point of water, defined as 100 °C: $\alpha \approx 0.003661\ ^\circ\text{C}^{-1}$.

The size of the unit was, therefore, so chosen that the interval between the normal melting point of water and its normal boiling point be *exactly* 100 °C. In this way, the *size* of the degree Celsius is exactly that of the historical degree centigrade defined by Celsius in 1742. The temperature is defined, as a consequence, in terms of the "Celsius scale": $T_{emp} = 1/\alpha + t$, with $1/\alpha = 273.15$ °C *exactly*. The coefficient of cubic thermal expansion of the gas α is defined as $\lim_{p_0 \to 0} \alpha_v = \lim_{p_0 \to 0} \alpha_p = \alpha$, where α_v and α_p are obtained by measuring the volumes or the pressures, respectively, of a sample of gas enclosed in a bulb thermostated at 0 °C and then at 100 °C: values of $1/\alpha$ ranging from 273.25 °C to 273.08 °C were actually measured.

Subsequently, however, international preference moved to the use of only *one* fixed point for the definition of temperature. The main objection [6] to the use of a two-point definition was that the definition of the absolute temperature is affected in this case by the experimental uncertainty in the determination of a (0.01–0.02° at the time), while in the Celsius temperature, an empirical scale, it is exact by definition. Celsius temperature, an empirical scale, is exact by definition. The reverse seemed more appropriate. In 1954, the Comité International des Poids et Mesures endorsed the one-point definition, which was included by the next CGPM in 1960 [5] in the revision of the 1948 scale. Temperature T_{48} was defined proportional to gas pressure, at a constant volume, $(p/p_0)_V = T$. The size of the temperature unit was then set by defining the numerical value of a *single* thermodynamic state, which is actually the only definition necessary for a scale that assumes as well $T = 0$ K at zero thermodynamic temperature.

This single state was selected to be the triple point of water: it is today reproducible to within ±50 μK: it was given the value, *exact by definition*, 273.16 K. The term "kelvin" of this symbol was adopted in 1954; subsequently, in 1968, the symbol of the unit was changed to K, *no longer* °K. At that time, this value was consistent, within the uncertainty of its determination, with the Celsius scale (in which the triple point of ice is 0.01 °C) since the temperature value of the normal boiling point of water was still 100 °C. After subsequent measurements performed in 1976–1988, the "0 °C–100 °C" historical interval lost its exactness, when the interval was found shorter by (26 ± 2) mK. However, the CGPM in 1989 [5] decided to maintain the 1954 definition of the Celsius temperature, that is, to maintain the *size* of the degree Celsius equal to that of the kelvin. Therefore, the temperature value of water boiling at 101325 Pa is 99.974 °C, a big change for the more accurate thermodynamic tables. If it had been decided, instead, to maintain *also* the difference between ice point and normal boiling point of water equal to 100 °C (a decision having no physical basis), the temperature of the triple point of water ought to have been redefined as: 273.16 × (100:99.975) = 273.23 K.

The present definition for the unit of the fundamental physical quantity known as thermodynamic temperature, symbol *T*, is the kelvin, symbol K, defined as the fraction 1/273.16 of the thermodynamic temperature of the triple point of water. The CIPM in 2005 [5] clarified the definition of the triple point of water by specifying the isotopic composition of the water to be that of Vienna Standard Mean Ocean Water (V-SMOW). This was the situation until a process started in the scientific community and later within the Conférence Internationale des Poids et Mesures (CGPM), concerning the redefinition of the kelvin by using the Boltzmann constant. In its Resolution 1 of October 2011 [5], the CGPM "takes note of the intention of the International Committee for Weights and Measures to propose a revision of the SI as follows: the International System of Units, the SI, will be the system of units in which:

- The ground state hyperfine splitting frequency of the caesium 133 atom $\Delta\nu(^{133}Cs)_{hfs}$ is exactly 9 192 631 770 Hz.
- The speed of light in vacuum *c* is exactly 299 792 458 m/s.
- The Planck constant *h* is exactly $6.626\,06 \times 10^{-34}$ J/s.
- The elementary charge *e* is exactly $1.602\,17 \times 10^{-19}$ C.
- The Boltzmann constant *k* is exactly $1.380\,6488 \times 10^{-23}$ J/K.
- The Avogadro constant N_A is exactly $6.022\,14 \times 10^{23}$ reciprocal mole.
- The luminous efficacy K_{cd} of monochromatic radiation of frequency 540×10^{12} Hz is exactly 683 lumen per watt.

where

1. The hertz, joule, coulomb, lumen, and watt, with unit symbols Hz, J, C, lm, and W, respectively, are related to the units second, meter, kilogram, ampere, kelvin, mole, and candela, with unit symbols s, m, kg, A, K, mol, and cd, respectively, according to $Hz = s^{-1}$, $J = m^2\ kg\ s^{-2}$, $C = s\ A$, $lm = cd\ m^2\ m^{-2} = cd\ sr$, and $W = m^2\ kg\ s^{-3}$.
2. The symbol X in this Draft Resolution represents one or more additional digits to be added to the numerical values of h, e, k, and N_A, using values based on the most recent CODATA adjustment, from which it follows that the SI will continue to have the present set of seven base units, in particular the kelvin will continue to be the unit of thermodynamic temperature, but its magnitude will be set by fixing the numerical value of the Boltzmann constant to be equal to exactly $1.380\ 6 \times 10^{-23}$ when it is expressed in the SI unit $m^2\ kg\ s^{-2}\ K^{-1}$, which is equal to $J\ K^{-1}$".

Actually, the CGPM in October 2011 did not yet take any final decision about the earlier provided matter, so no change in the kelvin definition is in force after that meeting at least until the next CGPM meeting would possibly confirm the previous intention. Table 65.1 summarizes chronologically the developments of thermometry. In principle, any thermodynamic law can be used to obtain a thermodynamic scale.

65.4 Empirical Temperature Scales

The determination of thermodynamic temperature from first principles is a difficult experiment. Every measurement is, in fact, only an approximation, because of imperfections in the model used for describing the basic thermodynamic law, of insufficient control of the secondary experimental parameters that are included in the model (the so-called corrections), or of experimental random errors. However, in nature, a temperature-independent phenomenon is rather the exception than the rule. Any temperature-dependent physical quantity can, in principle, be used to measure temperature. An important element to consider for the evaluation of a candidate empirical scale, apart from the trivial necessity for the physical quantity to behave monotonically with temperature (e.g., water density is not), is its being or not a linear transformation of absolute temperature, allowing the experimentalist to assign numerical values sequentially to the thermodynamic states. The relationship between the physical quantity and the thermodynamic temperature must be carefully specified; this specification is often difficult to find in the literature, especially over long periods of time. It is well known that scale-conversion calculations are cumbersome and errors frequently occur. Most of the possible empirical scales are not linear in thermodynamic temperature. All empirical scales require *calibration*.

65.4.1 Semi-Empirical Scales

The only element that makes these scales different from a primary thermodynamic scale is the increased number of conventions stipulated for their definition. The defining equation may be the exact representation of a thermodynamic law, but the value of some of the parameters either cannot be calculated with sufficient accuracy, or accurate calculations are not estimated convenient. For instance, Curie's law for paramagnetic substances is $\chi = \chi_0 + C/(T + A + B/T)$. This equation, which also is nonlinear in T, can be exploited to determine temperature by measuring the magnetic susceptibility χ, but the three constants must be experimentally determined by measuring the susceptibility values (χ_1, …χ_3) at three reference temperatures ($T_{r,1}$, …, $T_{r,3}$), whose values are *defined*. The procedure can then be considered a *calibration* of the experimental setup. The reference points are at least two. Since they are always selected from well-defined thermodynamic states, it is assumed that, when the state is exactly reproduced, the thermodynamic temperature is too, independently of the experimental setup or of the experimentalist. A scale so defined assumes no approximations in the defining thermodynamic equation and no simplifications in the experiment, as would be the case of a primary thermodynamic measurement. Consequently, the

TABLE 65.1 Short History of Thermometry

1600	Thermoscope invented by Galileo Galilei	The thermoscope has no graduated scale; therefore, temperature evaluation is only qualitative
Mid-1600	Liquid-in-glass graduated thermometers are widely spread	Famous are the alcohol thermometers of the Accademia del Cimento in Florence Their graduation marks are made of colored glass bits attached to the stem
1694	Melting ice and boiling water adopted by G. Renaldini as fixed points	Fixed points must be established by reference to natural phenomena, whose temperature is assumed to be intrinsically determined
1724	Modern temperature scale proposed by G. Fahrenheit, defined by a thermometer, a law, and three fixed points	Fahrenheit's thermometer was a mercury-in-glass one. Thermal expansion vs. temperature was assumed linear. Three fixed points were defined: 0 °F: temperature of a mixture of water, ice, and ammonium chloride 32 °F: temperature of melting ice 96 °F: temperature of human body
1742	The temperature interval between melting ice and boiling water subdivided into 100 equal parts by A. Celsius	The degree Celsius was the one eventually adopted
1821	Thermoelectric effect discovered by T.J. Seebeck	The Seebeck effect is the basis for the thermometers designated as thermocouples
1854	The modern absolute temperature definition, suggested by W. Thomson, is based on Carnot cycle	A scale whose definition does not depend on a specific substance is called "absolute"
1871	First electrical resistance thermometer built by C.W. Siemens	The electrical resistance thermometer is still today the best practical thermometer
1879	Blackbody radiation law proposed by J. Stefan; forms the basis for the radiation thermometers, with Planck's law	The radiation law is strongly nonlinear in temperature, since it depends on T^4
1967	The XIII Conférence Générale des Poids et Mesures adopts the kelvin scale with a single fixed point: 273.16 K	The unit of temperature is designated "kelvin"; its symbol is K
1990	The XVIII Conférence Générale des Poids et Mesures adopts the present version of the International Temperature Scale, the ITS-90	The ITS-90 provides the best to-date practical approximation of the thermodynamic scale and offers a reproducibility that is better than the thermodynamic scale
2006	Definition of a "*mise en pratique*" of the kelvin	ITS-90, PLTS-2000, and thermodynamic scale transformed in realizations of the "*mise en pratique*" of the kelvin. A technical annex contains changes in scales definitions
2011	CGPM takes note of the change of the kelvin definition in the frame of "the possible future revision of the International System of Units, the SI"	If the present CCU proposal will stand, the kelvin will be defined by means of the Boltzmann constant set to a stipulated value. This will not have any practical effect on the *mise en pratique* in force since 2006

scale definition need *not* specify any characteristics of the experimental equipment or any measurement procedure. Measurements have to be performed according to state-of-the-art techniques required by the accuracy to be achieved. With respect to primary measurements, the measurement procedure is simplified in that calibration at the fixed points avoids measuring several secondary parameters and performing cumbersome and uncertain correction calculations. The two (minimum) defining points are called *fixed points* of the scale. None need to be the defining point of the kelvin scale. The scale definition assigns them *conventional* temperature values. Though these values are assumed to be *exact* thermodynamic temperature values, they are actually the best up-to-date approximations. The empirical temperature θ defined by these scales is considered to be satisfactory enough as an approximation of the thermodynamic temperature T.

65.4.2 Strictly Empirical Scales

Quite frequently, the thermodynamic relationship between the measured physical quantity and thermodynamic temperature is not known to sufficient accuracy, or this relationship is not convenient for direct use; it might be as well that measurements are executed with instruments that cannot or need not be fully characterized. A temperature scale stemming from situations such as these is substantially different from semi-empirical scales because the defining equation becomes a purely computational model to obtain an approximation of the thermodynamic temperature T. The model must prove to fit T over the whole chosen range between the fixed points within the chosen uncertainty, on the basis of *experimental* $T(y)$ data (actually only an estimate $\theta(y)$ is available, as data are affected by experimental errors). The number p of parameters necessary to fit the model equation to the required accuracy determines the number of the fixed points necessary to *calibrate* the thermometer. The optimum spacing between these fixed points on the temperature scale is determined by the characteristics of the model, that is, by the mathematical function used, but at the same time, it is obviously constrained by the availability of suitable thermodynamic states in nature.

Such a scale requires an *interpolating instrument*. Since the relationship between the chosen physical property (e.g., electrical resistance) and the thermodynamic temperature is not calculated from first principles but is derived from experimental data, the use of that temperature scale is limited to a specific device, or at least to a family of devices. This point will be made clearer by the example of the platinum resistance thermometer. In the first place, the experimental data pertain exclusively to a *specific* substance, platinum, as they cannot be extended, by means of a theory, to be valid for other metals. Other restrictions must be considered as well, for instance, the bulk material must be in well-defined chemical (oxidized…) and physical (strain-free…) states, and the results are valid only for a given range of chemical purity. When some of the characteristics of the interpolating instrument can be specifically quantified, the range of the accepted values for these parameters must be defined as well in the scale, to enable one following this prescription to reproduce the results, that is, the scale. If this is not possible, with, for example, diode thermometers or thermistors, results, *and that scale*, will be valid only for the specific production-lot of devices from which results were obtained. The thermometer used in a scale of these types is known as an "interpolating instrument," because it is not itself required to reproduce a thermodynamic property, but only precisely the smooth selected function that is used to *interpolate* temperature values between the values assigned to the fixed points. To summarize, there are three constitutive elements associated with an empirical scale:

1. An interpolating instrument
2. A mathematical definition
3. A set of *fixed points*

The need of an *interpolating instrument* has already been illustrated. The *mathematical definition* must conveniently be the simplest function (or set of functions) representing, within the stated accuracy, the relationship between the measured quantity and the thermodynamic temperature; it has a number of free parameters, a characteristic of an *empirical* scale, whose numerical values must be obtained from "calibration" of the interpolation instrument at a number of reference points. The number of these *defining fixed points*, called "fixed points," must be equal to the number of free parameters, and their position must be such as to permit the best compromise between mathematical requirements and what is available in nature. The temperature value of the reference point is *defined*, that is, *fixed* by the scale definition. A scale of this type, which is the most common, is affected by a specific shortcoming, which must be kept in mind, called *nonuniqueness*, a *nonstatistical* term expressing the concept that in each particular unit implementing the *interpolating instrument* (i.e., the thermometer), the relationship between the measured quantity and temperature is slightly different. Therefore, the mathematical interpolating procedure defined by the scale *cannot* adequately represent the measurements performed with each individual unit of the thermometer, but can only approximate the physical behavior of these units. In other words, if the scale definition is applied to two specific units of the thermometer placed in an isothermal enclosure, there will be a measurable difference between the temperature values supplied

by their calibration tables and the temperatures actually measured. This is true at every temperature *except* at the fixed points, at which the units have originally been calibrated. Calibration at the fixed points means, by definition, to associate the measured numerical value of the property (e.g., of electrical resistance) to each temperature value of the calibration table (expressed in the scale, e.g., T_{90}). The nonuniqueness effect cannot be reduced, since it is a nonstatistical systematic error intrinsic to thermometer fabrication. Consequently, it must neither be mixed up with errors in the scale realization, nor be ascribed to the limited ability of the mathematical definition to match the physical behavior of each individual thermometer; it is due, instead, to technical limits of the interpolating instrument itself. Still in connection with these scales, another problem, concerning instead their mathematical definition, is the *smoothness* of the selected mathematical model with respect to the physical relationship between the measured quantity and thermodynamic temperature. In order to model this relationship accurately, high-order polynomials are often required, which tend to oscillate between the constraints. A striking example is the difference between the International Temperature Scale of 1990 (ITS-90) and the thermodynamic scale, whose smoothness will be illustrated in Section 65.6.

65.5 International Temperature Scales

The measurement of temperature is a necessity in most laboratories. As a consequence of the accuracy requirements, quite early, an ITS was agreed upon a strictly empirical scale. At its first meeting after the Convention du Mètre, the CGPM (actually a diplomatic body) in 1887 endorsed the "Normal Hydrogen Scale." In 1927, the CGPM adopted the first ITS. Since then, two other such international scales were agreed: in 1948 and 1968. A historical review of this process can be found in many texts, for example [3,7–11]. The present version of the temperature scale (ITS-90) was adopted in 1990 and it is still in charge, but in 2006, a basic change in meaning has been adopted (see next section). The meaning of an "official" scale is not to be considered a bureaucratic one, but scientific and technical. As in all types of measurements, to improve accuracy, not only random errors must be reduced, but the largest possible number of causes of systematic deviations as well has to be eliminated. An international body, within which the recognized specialists in the temperature measurements field meet and discuss their experiences and those of their colleagues, makes it possible to obtain the best approaches for reducing the error limits. The first task of such a body is to collate the best estimates of the thermodynamic temperature and then to assess methods for its measurement to the state-of-the-art accuracy. In carrying out this work, it is often realized that the direct measurement of the thermodynamic temperature is affected by an uncertainty larger than the reproducibility level obtainable in empirical temperature measurements. Consequently, so far, it has proved advantageous to define an empirical scale that can be reproduced more accurately and is internationally agreed upon (the ITS). Another equally important role of this body is to assess the *traceability of lower-level approximations* of the thermodynamic scale to the ITS and to fill a gap of information between the ITS definition and the international codes developed for industry (e.g., by OIML, IEC) or prepared by other international bodies (e.g., IUPAC, IUPAP). At the same time, the body reviews critically the scientific and technical information necessary for code updating. The change in the philosophy of the meaning of the temperature scales that occurred in 2006 has an influence also on these matters, as illustrated in the next section.

65.5.1 *"Mise en Pratique"* for the Definition of the kelvin

In 2005, the CCT approved a document, officially endorsed in 2006, containing a basic change in the meaning of temperature scale "*Mise en pratique* for the definition of the kelvin" [11]. This document provides the information needed to perform a practical measurement of temperature in accord with the SI. This document states: "for temperatures other than the triple point of water, direct measurements of thermodynamic temperature require a primary thermometer based on a well-understood physical system whose temperature may be derived from measurements of other quantities. In practice, primary thermometry is difficult and time consuming and not a practical means of disseminating the kelvin.

As an alternative, the International Temperature Scale provides an internationally accepted recipe for realizing temperature in a practical way." At present, "the ITS-90 and the PLTS-2000 [between 1 K and 0.9 mK] define temperatures T_{90} and T_{2000} that are good approximations to thermodynamic temperature." Concluding, "Recent developments in thermometry have motivated the creation of a broader, more flexible document that incorporates the temperature scales in current use: the *mise en pratique* for the definition of the kelvin. As envisioned by the CCT, the *mise en pratique* will serve as a reference for

- The text of the ITS-90 and PLTS-2000
- A Technical Annex of material deemed essential to realization of the ITS-90 or PLTS-2000, but not included in the scale definitions themselves
- Descriptions of primary thermometers for direct measurement of thermodynamic temperature
- Assessment of the uncertainty of the ITS-90, PLTS-2000, and measurements made by primary thermometry"

The rational of the *mise en pratique* is that, since it was promulgated as the realization of the kelvin, the status of the official temperature scales—and of the thermodynamic scale—is now that of a specific realization of the *mise en pratique*. This basic change has several important implications that should carefully be kept in mind [12,13]. ITS-90 and PLTS-2000 have not changed their regulatory status, but the new method approved by the CIPM has relaxed some of the previous constraints placed on temperature standards:

1. Adjustment to their definitions can now be implemented by simply updating the Technical Annex. In fact, the specifications for the triple point of water and the e-H_2 fixed points of the ITS-90 have recently been refined to account for isotopic effect [11].
2. Addition "in due course" of other methods to the *mise en pratique* is also foreseen, at present with the explicit mention to thermodynamic temperature.

This means that the new method intrinsically does not prevent the acceptance of a multiplicity of methods for the *mise en pratique*, even though they may apply to the same temperature range and have different levels of uncertainty. This simply means the addition to the *mise en pratique* of scales that had previously only the status of "approximations" to the ITS-90, without formal status, showing levels of accuracy—and precision—lower than that of the ITS-90. A summary of the official scale definitions is reported in Table 65.2.

65.6 Differences between the Realized Thermodynamic Temperature Scale and the T_{90}

For the most accurate thermo-physical and chemical–physical data, the differences $T - T_{90}$ are significant. For this reason, the Technical Annex to the *mise en pratique* of the kelvin now includes an official relationship between thermodynamic temperature T and T_{90}.

Interpolating functions were generated by CCT for representing this relationship [14]. The method used was to subdivide the range of the *critically selected 1046 experimental temperature values* available until 2010 in the temperature range from 4.2 K (normal boiling point of ^{4}He) to 1358 K (copper fixed point of the T_{90}) into 42 narrow subranges, and for each of them to obtain a *consensus* mean value and attribute it to the mean temperature value of the subrange; for this purpose, in the ranges 2.0–35 and 77–273.16 K, a weighted mean was used—basically this choice corresponds to keep intact the short-range local behavior of the experimental data. Then, the set of 42 data was fitted using:

- Between 8.0 and 273.16 K, a logarithmic polynomial of degree 7
- Between 273.16 and 1358 K, a rational polynomial of degree 4
- Between 2.0 and 8.0 K, the difference set to zero, exactly
- Between 1.0 and 2.0 K, $T - T_{2000} = T_{2006} - T_{2000}$, where T_{2000} is the PLTS-2000 scale
- Between 0.65 and 1.0 K, $T - T_{2006}$, where T_{2006} is from [15]

TABLE 65.2 International Temperature Standards from 1927 to 2011 and the Future

	ITS-27	IPTS-48 (from 1960)		IPTS-68		ITS-90	PLTS-2000	"*Mise en Pratique*"	Future (Anticipated)
	1927	ITS 1948	1954[a]	1968	1976	1990	2000	2005	≥2014
Unit	Degree centigrade, °C	Degree Celsius, °C	Degree kelvin, °K; degree Celsius, °C	kelvin, K; degree Celsius, °C	kelvin, K; degree Celsius, °C	kelvin, K; degree Celsius, °C	kelvin, K; degree Celsius ,°C	kelvin, K; degree Celsius, °C	kelvin, K; degree Celsius, °C
Reference T	0.000°C and 100.000°C	0° and 100°	273.16 K	273.16 K	273.16 K	273.16 K	273.16 K	273.16 K	None
Range	0.000°C up	90.2 K up	90.2 K up	13.8 K up	13.8 K up	0.65 K up	0.9 mK–1 K	ITS-90, PLTS-2000, thermodynamic scale realizations; Technical Annex with changes of the above included (see text)	0.9 mK up
Interpolating instrument (cryogenic range, K)	None	SPRT	SPRT	SPRT; T_{58} (^{4}He vp) T_{62} (^{3}He vp) recommended	SPRT; T_{58} (^{4}He vp) T_{62} (^{3}He vp) recommended	0.65–3.2 ^{3}He vp 1.25–5 ^{3}He vp 3.0–24.6 ICVGT 13.8 up SPRT	0.9 mK–1 K melting curve ^{3}He		Possible re-definition of the kelvin in terms of Boltzmann constant k Scale major revisions of: ITS-xx, Other implementations of the "*mise en pratique*," Technical Annex
Definition fixed points (cryogenic range)	None	nbp O_2	nbp O_2	tp H_2, vp H_2, nbp H_2, nbp Ne, tp O_2, nbp O_2	tp H_2, vp H_2, nbp H_2, nbp Ne, tp O_2, nbp O_2	ICVGT: 3–5 K ^{3}He or ^{3}He vp, tp H_2, tp Ne; SPRT: tp H_2, vp H_2, nbp H_2, tp Ne, tp O_2, tp Ar, tp Hg	None		
Mathematical definition (cryogenic range)	None	Single interpolation polynomial	Single interpolation polynomial	Reference function plus correction polynomial for sub-ranges	Same, but new reference function	vp: polynomials. ICVGT: quadratic polyn. SPRT: New ref. function and corr. polynomials; multiple definitions in some sub-ranges	Polynomial		
Thermodynamic scale	Excluded from defin.[b]	Excluded from defin.	Excluded from defin.	Excluded from defin.	Excluded from defin.	Excluded from defin.	Excluded from defin.	Included "in due time"	Included

[a] In 1964, a provisional definition between 13 K and 90 K was adopted, CCT-64.

[b] Before 1927, the thermodynamic hydrogen "normal scale" (CVGT) was the standard; until 2005, the international scales were only empirical and an approximation of the thermodynamic T. SPRT, standard platinum resistance thermometer; vp, vapor pressure scale; nbp, normal boiling point; tp, triple point.

TABLE 65.3 $T - T_{90}$ Functions

Range (K)	Interpolating Function	Parameters	
0.65–1.0 K	$T - T_{90} \equiv T - T_{2000}$ $(T - T_{90})/\text{mK} = \sum_{i=0\ldots3} d_i\,(T_{90}/\text{K})^i$	$d_0 = -14.0651$ $d_1 = 40.9970$ $d_2 = -44.1079$	$d_3 = 16.5315$
1.0–2.0 K	$T - T_{90} \equiv T_{2006} - T_{2000}$ $(T - T_{90})/\text{mK} = \sum_{i=0\ldots5} a_i\,(T_{90}/\text{K})^i$	$a_0 = 8.7999$ $a_1 = -54.8216$ $a_2 = 101.4590$	$a_3 = -83.5816$ $a_4 = 32.2307$ $a_5 = -4{,}7513$
2.0–8.0 K	$T - T_{90}$: = 0	—	
8.0–273.16 K	$(T - T_{90})/\text{mK} = \sum_{i=0\ldots7} b_i \cdot (\log_{10}(T_{90}/273.16\ \text{K}))^{i+1}$	$b_0 = 4.42457 \cdot 10^1$ $b_1 = -1.76311 \cdot 10^2$ $b_2 = -1.53985 \cdot 10^3$ $b_3 = -3.63685 \cdot 10^3$	$b_4 = -4.19898 \cdot 10^3$ $b_5 = -2.61319 \cdot 10^3$ $b_6 = -8.41922 \cdot 10^2$ $b_7 = -1.10322 \cdot 10^2$
273.16–1358 K	$(T - T_{90})/\text{mK} = (T_{90}/\text{K}) \cdot \sum_{i=0\ldots4} c_i\,(273.16\ \text{K}/T_{90})^{2i}$	$c_0 = 0.0497$ $c_1 = -0.3032$ $c_2 = 1.0254$	$c_3 = -1.2895$ $c_4 = 0.5176$

Source: Report of Working Group 4, Doc. CCT/08-13rev, 2008, Comité Consultatif de Thermométrie (CCT), Bureau International des Poids et Mesures, Sèvres, France, http://www.bipm.org; Fischer, J. et al., *Int. J. Thermophys.*, 32, 12, 2011.

The functions are reported in Table 65.3. The basic feature of the results is the following:

- The functions fit the 42 data points well within their uncertainty.
- The existing discontinuity of the first derivative $\mathrm{d}(T - T_{90})/\mathrm{d}T$ at 273.16 K arising from the ITS-90 *reference functions is maintained*, whose value, however, depends on each thermometer used, due to ITS-90 nonuniqueness.
- The function set to zero between 4.2 and 8 K produces *another discontinuity* of the first derivative at 8 K.

Since second-order discontinuities in interpolating functions are quite inconvenient outside strict metrological applications, namely, for the correction of thermophysical or physical–chemical data, a different calculation has been published in [16] for correcting the data of these types between 2.3 K (lambda point of ^{4}He) and 1238 K (close to copper fixed point of the T_{90}), having the following characteristics:

- Directly interpolates the whole set of 1046 selected experimental data, without the intermediate step of the 42 local-averages or further assumptions—like weighting
- Covers the full range with a *single function*, continuous up to the first derivative over the full range: the four-knots function is a *cubic spline*, for optimal smoothness (see Table 65.4)
- By providing the function also in the form of its explicit *polynomial representation* between pairs of knots (Cox–de Boor recursion formula), that is, for each of the four subintervals, allows the user to compute the $T - T_{90}$ function only in the temperature subinterval(s) of one's interest (see Table 65.4)
- Removes the discontinuity of the first derivative at 273.16 K and does *not* retain the difference $(T - T_{90})$: = 0 at 273.16 K, though the resulting value is <+0.0005 K—the spline function does not force the function to any specific value at any knot
- Also removes the second-order discontinuity of the official function at 8.0 K

The deviations of the function in Table 65.4 with respect to the official set of functions in Table 65.3 are minor for many users and are limited within one standard deviation of the data, as shown in Figures 65.3 and 65.4 [16] for the full range (with detail below 50 K in the insert), and above 273.16 K, respectively. In Figure 65.4, curve (d), closest to (a), indicates the variability within CCT elaborations referring to a previous one using a different interpolating function—see Doc. CCT/08-13rev, (a) in [14].

TABLE 65.4 $T - T_{90}$ Functions

Four Subintervals Spline		
Spline Function[a] (2.3–1238 K)	Cubic Polynomials[b]	
	Subinterval (T_{knots})	Coefficients a_{ij}
−0.003073048	2.3–90 K a_0	−0.000593965
0.002521906	a_1	0.000084027
−0.007092070	a_2	−0.000003974
−0.007160607	a_3	0.000000052
0.011774420	90–300 K a_0	−0.002642255
0.015114007	a_1	−0.000063857
0.000330789	a_2	0.000000602
0.060097369	a_3	0.000000004
0.042668756	300–450 K a_0	0.003884825
	a_1	0.000157792
	a_2	−0.000001213
	a_3	−0.000000002
	450–1238 K a_0	0.012749953
	a_1	−0.000047281
	a_2	0.000000582
	a_3	−0.000000001

Source: Pavese, F. et al., *J. Chem. Thermodynam.*, 43, 75, 2011; Pavese, F., *J. Chem. Thermodynam.*, 44, 179, 2012.

[a] Spline coefficients b_j.

[b] Coefficients a_{ij} of the corresponding cubic polynomial (Cox–de Boor recursion formula: $p_i = a_{0i} + h(a_{1i} + h/2\,(a_{2i}\,h + 1/3\,a_{3i}\,h^2)$, $h = x - \tau_i$, where τ_i is the x value of the lower knot of the ith subinterval) for each of the four subintervals.

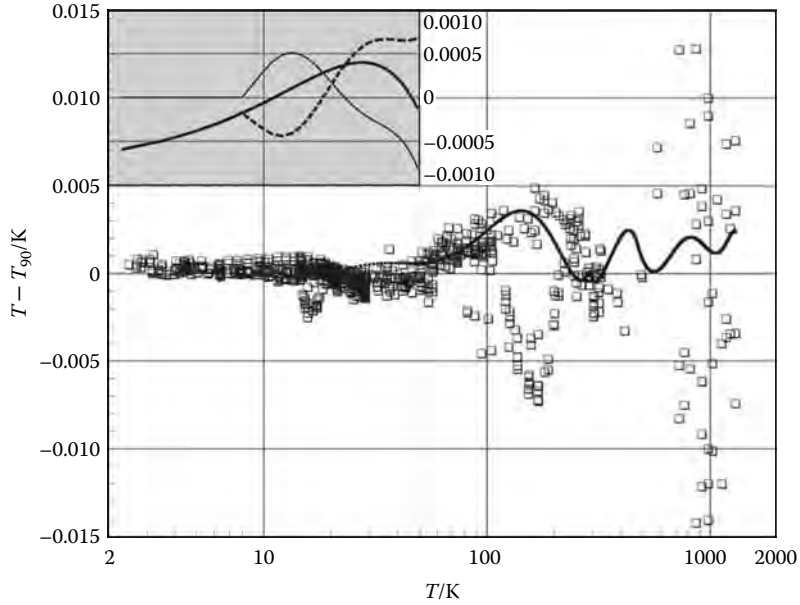


FIGURE 65.3 $T - T_{90}$ as computed by different functions. (From Pavese, F. et al., *J. Chem. Thermodynam.*, 43, 75, 2011; Pavese, F., *J. Chem. Thermodynam.*, 44, 179, 2012.) Full range: using the spline in Table 65.4.

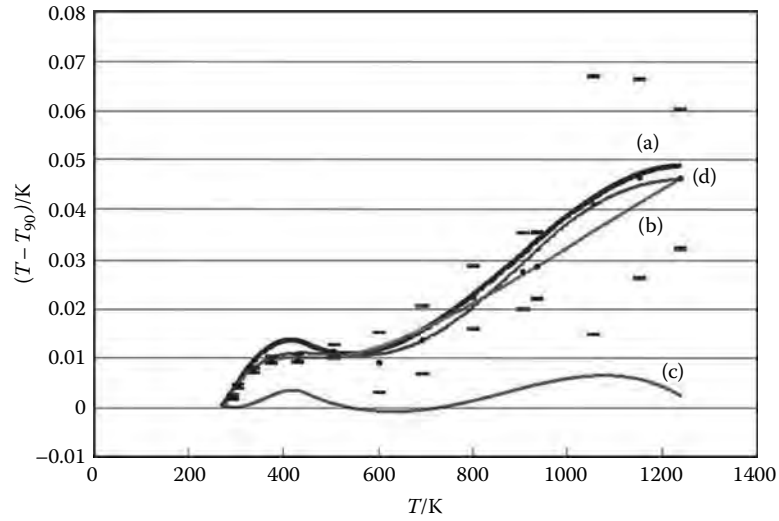


FIGURE 65.4 $T - T_{90}$ as computed by different functions. (Adapted from Pavese, F. et al., *J. Chem. Thermodynam.*, 43, 75, 2011; Pavese, F., *J. Chem. Thermodynam.*, 44, 179, 2012.) Range above 273.16 K: (a) using the spline in Table 65.4; (b) official function in Table 65.3 [14]; (c) difference (a)–(b). The short segments are the uncertainty limits of the standard uncertainty of each of the 42 local averages shown as thick dots. For curve (d), see text.

In the Figure 65.3, note that in this case there is not a first-derivative discontinuity at 8.0 K, nor there is at 273.16 K. The thick line is the difference between the spline function and the official functions in Table 65.3 (σ = 0.9 mK). The open squares are the residuals of the spline fit, indicating the dispersion of the original 1046 experimental data. *In the insert*, for $T < 50$ K the spline function (thick line), the official functions in Table 65.3 (thin line) and the difference between the two (broken line) are reported.

65.7 Calibration of Thermometers

"Laboratory scales" more or less based on thermodynamic determinations have always proliferated. However, having demonstrated how delicate a matter the temperature definition is, it seems now appropriate to draw attention on the great care that accurate temperature measurements demand. This care should start from the recognition of the need of using *calibrated* thermometers.

This is a general requirement when using any sensor or any transducer for performing quantitative measurements: a non-calibrated device is a gadget, not a measuring instrument.

To allow measurements whose uncertainty can be evaluated, a calibration is mandatory, and the calibration certificate must still be valid at the time the measurements are performed. This is by far less than trivial for thermometers, since all of them are likely to shift their characteristics over time after they are subjected, inherently to their function, to thermal cycling, from room to higher temperatures or to lower temperatures. The higher is the quality of the thermometers in this respect, the more sensitive they are, in general, to mishandling, especially mechanical shocks or stresses. The consequence is that the validity of a calibration within the declared uncertainty is invariably limited in time. The time span of validity depends on the conditions of use and cannot be easily predicted. In conclusion, thermometers should not only carry an initial calibration, but also be recalibrated at convenient intervals of time.

Concerning calibration in itself, it is a highly specialized job. In addition, except in cases when the measured values need only to be compared with the previous ones for an internal check of the "trend" of a process, like it may happen in industrial applications, the calibration must be "traceable" to higher-rank standards pertaining to a metrological chain—national or international. On this subject matter, see [10] and [17].

Each country has a National Metrology Institute heading a National Calibration System (whose name changes from country to country) of laboratories formally accredited to perform calibrations traceable to the national standards, each within a stated uncertainty. The system makes references to Metrology Regional Organizations [18], eventually headed by the Bureau International des Poids et Mesures (BIPM). Official monographs dedicated to methods and equipment for calibrating thermometers on the current official scales, namely, the ITS-90 (see Table 65.2), can be found in [19], most being currently under revision.

TABLE 65.5 Types of Contact Thermometers and Their Typical Precision

Type		Typical Temperature Range (K)	Typical Precision (mK)
1. Thermocouples[a]			
1.1	Type T	77 to >30	10[b]
1.2	Au + 0.07 at %Fe-KP	<4 to 100	10[b]
1.3	Au + 2.10 at %Co-KP	<4 to 300	10[b]
2. Resistance			
2.1	Platinum		
	Standard	14 to >300	Better than 0.5
	Industrial	20 to >300	50[c]
2.2	Rhodium–iron		
	Standard	0.5 to >30	Better than 0.3
	Industrial	<4 to >300	10
2.3	Platinum–cobalt		
	Industrial	2 to >300	10
2.4	CLTS	4–300	100
2.5	Germanium	1–100	DT/T = 0.00024
2.6	Carbon	0.5–100	200
2.7	Carbon–glass	0.5–100	DT/T = 0.002[d]
2.8	Thermistor	77 to >300	100[e]
2.9	Cernox®	0.1–325	25[f]
2.10	Ge on GaAs	0.2–300	10
3. Capacitance		2–300	100[g]
4. Diode			
4.1	Gallium arsenide	4 to >300	50
4.2	Silicon	4–300	50
5. Gas			
5.1	Vapor pressure	0.5–100	1–100
5.2	Gas	2 to >300	1–100
6. Mechanical (bimetal)			1000

[a] Other types available, suitable only for industrial use: type E and K down to 77 K; type S and R down to 220 K. KP = chromel P.

[b] Best precision for laboratory use in the differential mode and with offset for parasitic e.m.f.'s (up to several microvolts) corrected. Typical precision in the absolute mode (reference junction at refrigerant temperature): 100 mK.

[c] Can improve to 10 mK for selected units above 50 K.

[d] The relative precision in mK/K is preferred for this stable thermometer because of its R log(T) characteristics. Each unit does not cover the whole temperature range. The same applies for the carbon type.

[e] Can improve to a few millikelvin in narrow ranges above 200 K.

[f] Up to 77 K, 153 mK at 300 K.

[g] Not usable between 40 and 70 K, because of a slope inversion.

TABLE 65.6 Precision Requirements for the Measuring Instrument for Some Contact Thermometers

Sensor	Thermal Parameter (Max.)	Sensitivity	Max. Accuracy (Absolute and Relative) Required for			
			1 K	0.1 K	0.01 K	0.001 K
Thermocouple	emf/mV	($\mu V\ K^{-1}$)	(μV and (% or ppm))			
Type T	6	10–40	10 (0.6)	1 (0.06)	0.1 (60 ppm)	—
Au + 0.07 at %Fe-KP	5	12–22	12 (0.4)	1 (0.04)	0.1 (40 ppm)	—
Au + 2.1 at %Co-KP	14	5–70	5 (0.5)	0.5 (0.05)	0.05 (50 ppm)	—
Resistance	R_0/Ω	(ΩK^{-1})	($m\Omega$ and (% or ppm))			
Platinum	25	0.004–0.1	4 (0.4)	0.4 (0.04)	0.04 (40 ppm)	4 $\mu\Omega$ (4 ppm)
Platinum (industr.)	100	0.08–0.8	80 (8)	8 (0.8)	—	—
Rhodium-05–0.5% iron	100	0.15–0.6	150 (0.4)	15 (0.04)	1.5 (40 ppm)	0.15 (4 ppm)
CLTS	290	0.23	230 (0.1)	23 (0.01)	2.3 (10 ppm)	—
	$R_{4.2\,K}$ (Ω)					
Germanium	1000	0.1–1000	100 (10)	10 (1)	1 (0.1)	0.1 (0.01)
Carbon–glass	1000	0.1–1000	100 (10)	10 (1)	1 (0.1)	—
Diodes	$V_{4.2\,K}$/V	($mV\ K^{-1}$)	(mV and (% or ppm))			
Gallium arsenide	1.5	0.6–3	0.6 (0.05)	0.06 (50 ppm)	6 μV (5 ppm)	—
Silicon	2	2–60	2 (0.2)	0.2 (0.02)	0.02 (20 ppm)	—
Capacitance	$C_{4.2\,K}$/pF	($pF\ K^{-1}$)	(pF and %)			
	15000	(0)–300	150 (1)	15 (0.1)	1.5 (0.01)	—

The direct use of the thermodynamic scale for the calibration of thermometers is not generally advisable. Its correct realization is an extremely specialized job [3], and no official documents for the best practice to follow in its realizations are available so far from the specialized bodies in thermal metrology.

In Table 65.5, the typical precision that can be expected from a number of types of thermometers (generally suitable for the temperature range below room temperature, but some also suitable above; radiation thermometers not included) are reported: obviously any calibration can be expected to exceed these values in accuracy. In addition, when used, they place different requirements to the precision of the measuring equipment, some quite stringent, as shown in Table 65.6. Free consultancy on thermal metrology and specifically on thermometer use and calibration can be obtained by requesting from the nonprofit organization http://www.evitherm.org.

References

1. W. Thompson (1848), On an absolute thermometric scale founded on Carnot's theory of the motive power of heat, and calculated from Regnault's observations, *Phil. Magazine*, 33, 313–317.
2. W. Thomson (1851), On the dynamical theory of heat, with numerical results deduced from Mr Joule's equivalent of a thermal unit, and M. Regnault's observations on steam". *Trans. Royal Soc. Ed.* 20 (part II): 261–268, 289–298.
3. F. Pavese and G. F. Molinar (1992), Modern gas-based temperature and pressure measurements, *International Monograph Series on Cryogenic Engineering*, New York: Plenum Publishing, p. 514, and references therein. Augmented II Edition, 2012, Springer, New York, p. 660, also available as e-Book.
4. P. T. Landsberg (1961), *Thermodynamics*, New York: Interscience Publishers.
5. CIPM (1887), *Procès Verbaux 85*, Bureau International des Poids et Mesures, Sèvres, France; CGPM (1889), *Comptes Rendus des Séances de la I Conférence Générale*, Bureau International des Poids et Mesures, Sèvres, France, 35; CGPM (1960), *Comptes Rendus des Séances de la XI Conférence Générale*, Bureau International des Poids et Mesures, Sèvres, France, 124–133; CGPM (1989), *Comptes Rendus des Séances de la XVIII Conférence Générale*, Bureau International des Poids et Mesures,

Sèvres, France; CIPM (2005), Recommendation 2 (CI-2005): *Clarification of the Definition of the Kelvin, Unit of Thermodynamic Temperature*, International Committee for Weights and Measures, Bureau International des Poids et Mesures, Sèvres, France; CGPM (2011), On the possible future revision of the International System of Units, the SI, *Comptes Rendus des Séances de la XXIV Conférence Générale*, Resolution 1, Bureau International des Poids et Mesures, Sèvres, France, http://www.bipm.org/en/convention/cgpm/resolutions.html (accessed on March 14, 2013).

6. J. De Boer (1965), Temperature as a basic physical quantity, *Metrologia*, 1, 158–169.
7. Supplementary Information for the ITS–90, Monograph, Bureau International des Poids et Mesures, Sèvres, France, http://www.bipm.org/en/publications/mep_kelvin/its-90_supplementary.html (accessed on March 16, 2013).
8. J. F. Schooley (1986), *Thermometry*, Boca Raton, FL: CRC Press Inc.
9. T. J. Quinn (1983), *Temperature*, London, U.K.: Academic Press.
10. J. V. Nicholas and D. R. White (1994), *Traceable Temperatures*, Brisbane, Queensland, Australia: Wiley & Sons
11. *Mise en pratique* for the definition of the Kelvin & Technical annex for the International Temperature Scale of 1990 (ITS-90), 2006, Bureau International des Poids et Mesures, Sèvres, France, http://www.bipm.org/utils/en/pdf/MeP_K.pdf (accessed on March 16, 2013).
12. F. Pavese (2007), Possible implications of the principle of the 'mise en pratique' in its application to the kelvin, *Int. J. Thermophys*., 28, 1766–1774.
13. D. C. Ripple, R. Davis, B. Fellmuth, J. Fischer, G. Machin, T. J. Quinn, P. P. M. Steur, O. Tamura, and D. R. White (2010), The roles of the mise en pratique for the definition of the kelvin, *Int. J. Thermophys*., 31, 1795–1808.
14. (a) Report of Working Group 4, Doc. CCT/08-13rev, 2008, Comité Consultatif de Thermométrie (CCT), Bureau International des Poids et Mesures, Sèvres, France, http://www.bipm.org/en/committees/cc/cct/ (accessed on March 16, 2013). (b) J. Fischer, M. de Podesta, K. D. Hill, M. Moldover, L. Pitre, R. L. Rusby, P. P. M. Steur, O. Tamura, R. White, and L. Wolber (2011), Present estimates of the differences between thermodynamic temperatures and the ITS-90, *Int. J. Thermophys*., 32, 12–25.
15. J. Engert, B. Fellmuth, and K. Jousten (2007), A new ^{3}He vapour–pressure based temperature scale from 0.65 K to 3.2 K consistent with the PLTS-2000, *Metrologia*, 44, 40–52.
16. (a) F. Pavese, P. Ciarlini, and P. P. M. Steur (2011), Thermodynamic temperature differences from the ITS-90 for the correction of thermodynamic property data, *J. Chem. Thermodynam*., 43, 75–79; (b) F. Pavese (2012), Note on comment on 'Thermodynamic temperature differences from the ITS-90 for the correction of thermodynamic property data' by F. Pavese, P. Ciarlini, and P. P. M. Steur by J. Fischer, Chairman of WG4 of CCT, *J. Chem. Thermodynam*., 44, 179–180.
17. F. Pavese (2012), Mathematical and statistical tools in metrological measurements, Chapter E6-35-78 on EOLSS, Unesco Encyclopedia of Life Support Systems, http://www.eolss.net (accessed on March 14, 2013).
18. Bureau International des Poids et Mesures, Sèvres, Regional metrology organizations (RMOs), Practical info. http://www.bipm.org/en/practical_info/useful_links/rmo.html (accessed on March 14, 2013).
19. Bureau International des Poids et Mesures, Sèvres, Mise en pratique of the definition of the kelvin (2011). http://www.bipm.org/en/publications/mep_kelvin/ ; Bureau International des Poids et Mesures, Sèvres, Techniques for Approximating the ITS-90 (R.E. Bedford, G. Bonnier, H. Maas and F. Pavese), 1990. http://www.bipm.org/en/publications/mep_kelvin/its-90_techniques.html (accessed on March 14, 2013).



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

66

Thermal Conductivity Measurement

William A. Wakeham
Imperial College, London

Marc J. Assael
Aristotle University of Thessaloniki

66.1 Introduction **66**-1
66.2 Fundamental Equations **66**-2
66.3 Measurement Techniques **66**-3
Steady-State Methods • Steady-State Methods = Parallel-Plate Instrument • Measurement on Fluids • Measurements on Solids • Transient Methods • Transient Hot-Wire Technique • Hot-Disk Instrument • Transient Hot Bridge • Heated-Needle Probe • Laser-Flash Instrument
66.4 Instrumentation **66**-11
66.5 Appraisal **66**-11
References **66**-11

66.1 Introduction

There are three mechanisms whereby energy can be transported from one region of space to another under the influence of a temperature difference. One is by transmission in the form of electromagnetic waves (radiation); the second is the process of convection, in which a bulk or local motion of the material effects the transport; and the final process is that of thermal conduction when energy is transported through a medium. In most practical situations, energy transport is accomplished by all three processes to some extent, but the relative importance of each contribution varies markedly. For example, within an evacuated region, radiation is the sole mechanism of transport, whereas in an opaque solid, conduction is the only mechanism possible.

These processes of heat transfer are often very important in a wide variety of scientific and industrial applications. In the cooling of cast or crystalline materials (e.g., metals, semiconductors, or polymers) from a molten state to a solid state, the heat transfer within the material can have a profound effect on the final properties of the solid. Equally, the heat transfer in a foodstuff is a determinant of its cooking, freezing, or processing time, while the size of equipment needed to heat or cool the liquid or gas stream in a chemical plant depends sensitively on the heat transfer within and between one stream and another.

For these reasons, there has been a great interest in the understanding and description of all three heat transfer processes. Among the three, that of thermal conduction is the simplest to describe in principle, since the empirical law of Fourier simply states that the heat transported by conduction per unit area in a particular direction is proportional to the gradient of the temperature in that direction. The coefficient of proportionality in this law is known as the thermal conductivity and denoted here by the symbol λ. Many important materials, whether made of pure chemical components or mixtures, are of uniform composition throughout, and for them the thermal conductivity is a true physical property of the material, depending often only on the temperature, pressure, and composition of the sample.

However, particularly in the solid state, the thermal conductivity can depend on the direction of the heat flow, for example, in the case of a molecular crystal.

It is also conventional to speak of the thermal conductivity of various types of composite materials such as bricks, glass–fiber insulation, carbon–fiber composites, or polymer blends. In this case, the thermal conductivity is taken to be the empirical constant of proportionality in the linear relationship between a measured heat transport per unit area and the temperature difference over a prescribed distance in the material. The thermal conductivity is not then, strictly, a property of the material, since it can often depend on a large number of parameters, including the history of the material, its method of manufacture, and even the character of its surface. However, this distinction between homogeneous and inhomogeneous materials is often ignored and leads to more than a little confusion, especially where intercomparisons among measurements are concerned.

The fact that in most practical situations all three heat transfer mechanisms are present greatly complicates the process of measurement of the thermal conductivity. Thus, much early work in the field is substantially in error, and it has been really quite difficult to devise methods of measurement that unequivocally determine the thermal conductivity. For that reason, the instruments to be described in the following sections often seem to be rather far removed from the apparent simplicity implied by Fourier's Law.

66.2 Fundamental Equations

The essential constitutive equation for thermal conduction relates the heat flux in a material to the temperature gradient by the equation

$$Q = -\lambda \nabla T \tag{66.1}$$

It is not possible to measure local heat fluxes and gradients; thus, all experimental techniques must make use of an integrated form of the equation, subject to certain conditions at the boundaries of the sample. All experiments are designed so that the mathematical problem of the ideal model is reduced to an integral of the one-dimensional version of Equation 66.1, which yields, in general,

$$Q_a = G\lambda \Delta T \tag{66.2}$$

in which G is constant for a given apparatus and depends on the geometric arrangement of the boundaries of the test sample. Typical arrangements of the apparatus, which have been employed in conjunction with Equation 66.2, are two flat, parallel plates on either side of a sample, concentric cylinders with the sample in the annulus and concentric spheres.

Techniques that make use of Equation 66.2 are known as steady-state techniques and they have found wide application, some of which are discussed later. They are operated usually by measuring the temperature difference ΔT that is generated by the application of a measured heat input Q_a at one of the boundaries. The absolute determination of the thermal conductivity, λ, of the sample contained between the boundaries then requires a knowledge of the geometry of the cell contained in the quanity G. In practice, it is impossible to arrange an exactly one-dimensional heat flow in any finite sample so that great efforts have to be devoted to approaching these circumstances, and then there must always be corrections to Equation 66.2 to account for departures from the ideal situation.

If the application of heat to one region of the test sample is made in some kind of time-dependent fashion, then the temporal response of the temperature in any region of the sample can be used to determine the thermal conductivity of the fluid. In these transient techniques, the fundamental differential equation that is important for the conduction process is

$$\rho C_P \frac{\partial T}{\partial t} = \nabla \cdot (\lambda \nabla T) \tag{66.3}$$

which arises from an elementary energy balance in the absence of any other processes and in which ρ is the density of the material and C_p its isobaric heat capacity. In most, but not all, circumstances, it is acceptable to ignore the temperature dependence of the thermal conductivity in this equation and to write

$$\frac{\partial T}{\partial t} = \frac{\lambda}{\rho C_p} \nabla^2 T = a \nabla^2 T \tag{66.4}$$

in which a is known as the thermal diffusivity.

Experimental techniques for the measurement of the thermal conductivity based on Equation 66.4 generally take the form of the application of heat at one surface of the sample in a known time-dependent manner, followed by detection of the temperature change in the material at the same or a different location. In most applications, every effort is made to ensure that the heat conduction is unidirectional so that the integration of Equation 66.4 is straightforward. This is never accomplished in practice, so some corrections to the integrated form of Equation 66.4 are necessary. The techniques differ among each other by virtue of the method of generating the heating, of measuring the transient temperature rise, and of the geometric configuration. Interestingly, in one geometric configuration only, it is possible to determine the thermal conductivity essentially independent of a knowledge of ρ and C_p, which has evident advantages. More usually, it is the thermal diffusivity, a, that is the quantity measured directly, so that the evaluation of the thermal conductivity requires further, independent measurements.

In the following sections, brief descriptions of the specific applications of these general principles are given. The examples chosen for study are intended to cover the full spectrum of materials and thermodynamic states, and in each case, attention is concentrated on a method that has proved most accurate and is widely used. The steady-state and transient techniques are considered separately.

66.3 Measurement Techniques

66.3.1 Steady-State Methods

The steady-state methods employed for the measurement of the thermal conductivity of fluids and solids have most often employed the geometry of parallel plates so that it is that configuration described here in two variants. Coaxial cylinder equipment has largely been used within the preserve of the research laboratory, with the apparatus of Tufeu and Le Neindre [1] an excellent example of the genre.

66.3.2 Steady-State Methods = Parallel-Plate Instrument

A schematic diagram of a guarded parallel-plate instrument is shown in Figure 66.1. The sample is contained in the gap between two plates (upper and lower) maintained at a distance d apart by spacers. A small amount of heat, Q_a, is generated electrically in the upper plate and is transported through the sample to the lower plate. Around the upper plate, and very close to it, is placed a guard plate. This plate

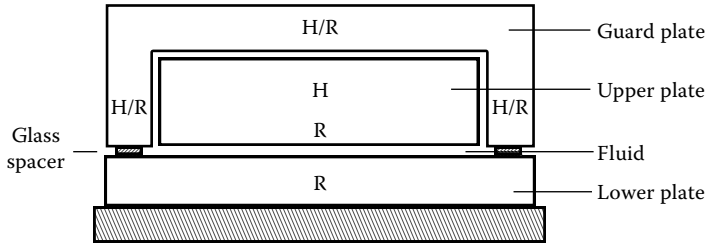


FIGURE 66.1 Schematic diagram of a guarded parallel-plate instrument. H, heater; R, resistance thermometer.

is, in many instruments, maintained automatically at the same temperature as the upper plate so as to reduce heat losses from the upper surface of the upper plate and to most nearly secure a one-dimensional heat flow at the edges of the sample.

The temperatures at the surfaces of the upper and lower plates are measured very precisely, as is the electric input of energy, so that the thermal conductivity can, in principle, be evaluated from the following equation:

$$Q_a = \frac{A\lambda\Delta T}{d} \tag{66.5}$$

where

ΔT is the measured temperature difference
A is the area of the upper plate

Whether the sample is a fluid or a solid, the electric energy generated in the upper plate is not all conducted to the lower plate. Thus, it is necessary in all cases to account for spurious heat losses and for all except opaque materials for the radiative transfer between the two surfaces. When the sample is transparent to radiation, this correction is straightforward and can be reduced by means of surface coating the plates to reduce their emissivity; but when the material adsorbs radiation, the problem is much more complicated and has been the subject of some controversy in the past, which has since been resolved (p. 147 of [2]). In many cases, the effort of performing absolute measurements cannot be justified so that the ratio A/d is determined by calibration with a material of known thermal conductivity.

66.3.3 Measurement on Fluids

Measurements with parallel-plate instruments on fluids have been performed for a considerable period of time. The technique has particular advantages in some special regions of thermodynamic space but requires great attention to detail if accurate results are to be obtained. In the most accurate instruments for fluids, the gap between upper and lower plates is kept as small as possible (perhaps as small as 0.2 mm). This has the benefit of reducing the effect of heat flows that are not normal to the heat surfaces, but more importantly, the small gap contributes to the reduction of the heat transferred by bulk convective motion of the fluid. Indeed, if very small temperature differences are employed by heating the top plate (to have a stable density gradient) *and* considerable care is taken to align the parallel plates normal to the earth's gravitational acceleration, then the effects of convective heat transfer can be rendered negligible (p. 154 of [2]). It is a fact of history that the necessary care with this instrument has been taken by only a few workers so that despite the fact that the instrument has been used in the temperature range from 4 to 800 K, and for pressures up to 250 MPa, only some of the measurements are reliable.

An example of what can be achieved with a parallel-plate instrument is provided by the work of Mostert [3] and Sakonidou [4] at the van der Waals laboratory in Amsterdam. They have used the technique near the critical point of a pure fluid or mixture where the extreme values of the compressibility make the fluid exceedingly prone to convection. In such circumstances, the small vertical extent of the fluid layer required for this technique avoids large density variations in the test layer and, combined with the fact that very small temperature differences (0.3 mK) can be employed, has enabled measurements of the thermal conductivity to be conducted to within 100 mK of the critical temperature at the critical density. Under these circumstances, the thermal conductivity of a pure fluid reveals an enhancement that is, in principle, infinite at the critical point itself.

The unique characteristics of the parallel-plate instrument mean that it is the method of choice for work near the critical region of a material. The arguments given earlier pertaining to the care required for reliable measurements militate against the production of commercial instruments.

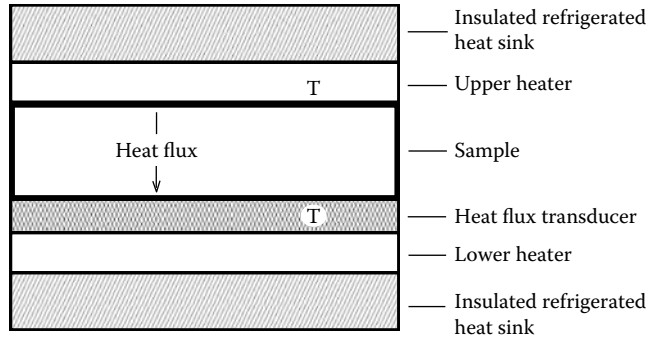


FIGURE 66.2 Schematic diagram of a heat flow meter instrument. T, thermocouple.

66.3.4 Measurements on Solids

A parallel-plate instrument of exactly the same type as has been described for fluids has been employed for solids. However, the spacing of the two plates is normally significantly greater for solids, owing to the difficulty of preparing very thin samples of solids. One essential difficulty with solid samples in this configuration is the contact between the two heater surfaces and those of the sample. Unlike the case for fluids, the contact cannot be made uniform at the molecular level. There is therefore always the possibility of an unaccounted interfacial heat transfer resistance. It seems likely that these considerations contribute to the wide differences between values reported for the same sample by different authors.

A more popular implementation of the parallel-plate configuration for solids is the so-called heat flow meter instrument, also described in the ASTM E1530 test method [5]. When applied to materials such as building insulation, the dimensions of this type of instrument can be very large.

In such instruments (see Figure 66.2), the upper heater plate is set at a higher temperature than the lower one. The hot and cold surface temperatures of the sample are measured with the two thermocouples permanently installed on the adjacent surface plates while a precalibrated heat flow transducer on the lower plate measures the magnitude of the heat flux through the sample. The thermal conductivity is calculated directly from Equation 66.5. In some cases, contact resistances (i.e., insulation) can be characterized and controlled by employing a pressurized gas in the sample chamber or applying a thermally conductive paste to the samples surfaces.

It should also be mentioned that there are many commercial steady-state instruments for measuring the thermal conductivity of solids that operate according to the guarded hot plate method. In the type of instrument, which complies with the ASTM C177 [6] and the ISO 8302 [7] test methods, there are two isothermal cold surface assemblies and a guarded hot plate, which is sandwiched between two identical specimens of the material. Electrically produced heat flux is dissipated through the hot plate and heat flows through both specimens creating a temperature gradient across them. The temperature drop and the heat flow from the hot plate surface are used to calculate the thermal conductivity value. The main difference between the heat flow meter instrument and the hot plate apparatus is that in the first one, the thermal conductivity value depends on the heat flux reference standards while the second one operates in an absolute way. Commercial instruments of both types are included in the listing of Table 66.1.

66.3.5 Transient Methods

There are rather more transient techniques that have achieved popularity than steady-state instruments. This is because transient techniques generally require much less precise alignment and dimensional knowledge and stability. Furthermore, some of the techniques have distinct advantages that arise from the speed of the measurement.

TABLE 66.1 Companies That Make Thermal Conductivity Instruments

Type of Instrument	Temperature Range	Supplier	Approximate Price (U.S. $)
Transient hot disk			
TPS 2500S	290–1,000 K	K-analys AB	25,000
Transient hot bridge	220–480 K	LINSEIS	45,000
Guarded hot plate			
GHP 456 Titan	110–1,000 K	NETZSCH	35,000
Heat flow meter			
HFM 300	290–350 K	LINSEIS	20,000
Hot-wire			
TCT 426	300–1,800 K	NETZSCH	Variable
Needle probe			
MTN01	240–350 K	HUKSEFLUX	8,000
Laser flash			
LFA 1000	290–1,800 K	LINSEIS	130,000
LFA 427	300–3,100 K	NETZSCH	Variable

Note: Prices are only indicative and depend on the operating temperature range and options of accessories.

66.3.6 Transient Hot-Wire Technique

In this technique, the thermal conductivity of a material is determined by observing the temporal evolution of the temperature of a very thin metallic wire (see Figure 66.3) after a step change in voltage has been applied to it. The voltage applied results in the creation of a line source of nearly constant heat flux in the fluid. As the wire is surrounded by the sample material, this produces a temperature field in the material that increases with time. The wire itself acts as the temperature sensor, and from its resistance change, its temperature change is evaluated, and this is related to the thermal conductivity of the surrounding material.

According to the ideal model of this instrument, an infinitely long line source of heat possessing zero heat capacity and infinite thermal conductivity is immersed in an infinite isotropic material, with physical properties independent of temperature and in thermodynamic equilibrium with the line source at time $t = 0$ at a temperature T_0. The heat transferred from the line source to the sample when a stepwise

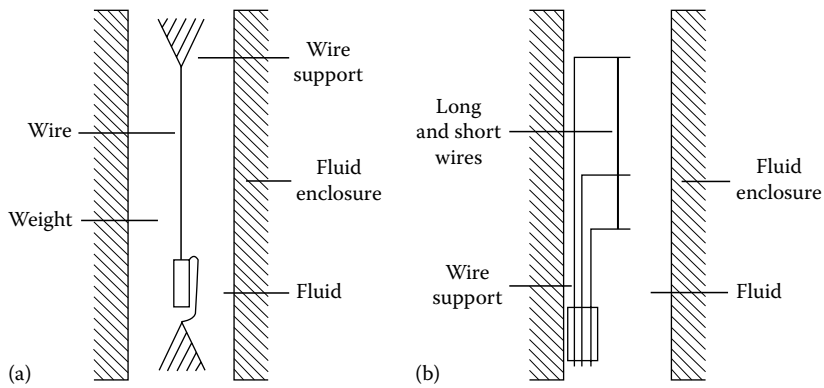


FIGURE 66.3 Schematic diagram of a transient hot-wire instrument for fluids: (a) hot wire is kept under constant tension by a weight; (b) hot wires and supports made of same material.

heat flux, q, per unit length is applied is assumed to be entirely conductive. Then the temperature rise of the material at a radial distance, r_0, which it transpires, is the same as the temperature rise at the surface of a wire of radius r_0, $\Delta T_i(r_0, t)$, which is given by

$$\Delta T_i(r_0,t) = T(r_0,t) - T_0 = \frac{q}{4\pi\lambda}\left[\text{in}\left(\frac{4at}{r_0^2 C}\right) + \frac{r_0^2}{4at} + \cdots\right] \tag{66.6}$$

In this equation, C is a known constant [8]. The equation suggests that, provided the radius of the wire is chosen small enough so that the second term on the right-hand side of Equation 66.6 is negligible, the thermal conductivity of the fluid can be obtained from the slope of the line ΔT_i, vs. ln t. Any practical implementation of this method of measurement inevitably deviates from this ideal model. The success of the technique, however, rests on the fact that by proper design, it is possible to construct an instrument that can operate very closely to the ideal model, while at the same time, small departures can be treated by a first-order analysis [8].

The roots of the transient hot-wire technique extend way back in time to the first heated-wire experiments in 1780. In the 1930s, the technique was used to measure the effective thermal conductivity of powders and solids [9]. However, its application to other materials was somewhat slower until in the late 1960s, the new technology associated with electronics made it possible to measure small, transient resistance changes with high accuracy in a period of less than 1 s. This development, pioneered by Haarman [10], made it possible to complete the transient heating process so quickly that, despite the inevitability of convective motion in the fluid from time zero, the inertia of the fluid ensures that the fluid velocity is sufficiently small so that there is no significant contribution to heat transfer. This fact prompted a rapid development of the measurement technique for fluids—first in gases and then in liquids—that was then followed by further developments in solids. The differences in the technique between solids and fluids are rather small; thus, some aspects of the instrumentation for liquids are briefly discussed, and the differences for solids are just outlined.

In the case of fluids, the instrumentation generally involves a wire some 7–25 μm in diameter (in order to reduce the correction owing to its heat capacity) and some 150 mm long. The wire is mounted vertically in a cylindrical cell containing the test sample. Often, a second wire differing only in length is employed to compensate automatically for effects at the ends of the wires via the electrical measurement system, but this can also be accomplished with potential taps [8]. In the more recent instruments [11], the wire supports are made of the same material so that they expand similarly to the wire when temperature rises, and hence the wire is always kept taut. Moreover, in this design, the wires are much shorter and placed one above the other. Whenever possible, platinum is used for the wire material because its resistance/temperature characteristics are well known and it can be readily obtained in the form of wires with a diameter as small as 5 μm. When the material under test is electrically conducting, it is necessary to insulate electrically the wire from the fluid. Several varieties of technique that enjoy different degrees of success depending on the range of conditions to be studied have been employed for this purpose. Near ambient temperature over a range of pressures, it has been found adequate to use a tantalum wire as the sensor that is electrolytically anodized to cover the wire with an insulating layer of tantalum pentoxide 100 nm thick [12]. Under more aggressive conditions, it has been necessary to employ ion-plating of the wire with a ceramic to secure the isolation [13]. In either case, the theory has been modified by the addition of a small correction.

In the case of solids, the need for the wire to be straight and vertical is removed by virtue of the rigidity of the material. Thus, Bäckström et al. [14] were able to employ a wire embedded as an arc within the compressed solid matrix of the material under study, particularly at very high pressures (up to 4 GPa). It should be noted that there are many standards (ASTM C1113 [15], EN 993-14/15 [16,17], ISO 8894 [18,19], and DIN 51046 [20]) that describe hot-wire test methods. The main problem arising in applying the technique to solids is the inevitable contact resistance between the samples and the wire.

A recent transient hot-wire design employs a silicone paste in order to avoid air gaps. The wires are placed inside a soft silicone layer that is squeezed between the two solids [21]. However, the last device is still available only for research laboratory use owing to difficulties in the sensor preparation. It is worth noting that this idea evolved from a transient hot-wire sensor originally employed for the measurement of the thermal conductivity of molten metals [22,23]. In that case, a 25 μm diameter platinum wire was placed over a sheet of pure green alumina, while its connections were printed on the substrate using platinum ink and a screen printing technique [24]. A second sheet of alumina was placed over them, and everything was hot-pressed in order to guarantee the best contact between the substrate and the platinum wire. Subsequent firing of the assembly produces a sufficiently robust sensor for it to be immersed in the molten metal.

The transient hot-wire technique has a unique advantage among transient methods that the thermal conductivity of the test material can be evaluated directly from the slope of the line relating the temperature rise of the wire to the logarithm of time. The heat capacity and density of the test material are required only to evaluate small corrections. Also advances in computer technology now allowed the use of a finite element method for the simulation of the exact geometry of the experimental setup, and therefore the uncertainty of the results is enhanced [21]. Furthermore, because the exact dimensions of the heating element and the cell are unimportant, the method avoids the intricate alignment problems of the parallel-plate technique while securing absolute measurements of the property. Despite these advantages and its wide application to measurements in gases, solids, and liquids, there has been no significant commercial development of an instrument of this kind, presumably because of the delicacy of the long thin wire in the case of devices for fluids, and the difficulty of sample and sensor preparation for solids. In Table 66.1, one of the few commercial hot-wire instruments is included.

66.3.7 Hot-Disk Instrument

A transient technique for which there is a commercial version suitable for solid materials is the transient hot-disk instrument shown in Figure 66.4.

The sensor in this case comprises a thin metal strip, often of nickel, wound in the form of a double-spiral in a plane. It is printed on, and embedded within, a thin sandwich formed by two layers of a material that is a poor electrical conductor but a good thermal conductor. This disk heater is then, in turn, placed either between two halves of a disk-shaped sample of solid material or affixed to the outside of the sample.

However it is configured, the essential measurement performed is the same as that for the hot-wire technique, and the temperature history of the sensor when subject to known electrical dissipation is inferred from its resistance charge. In the most recent version of the instrument, developed by Gustafsson [25], and also available commercially, the interpretation of the data is accomplished via a numerical solution of the differential equation rather than by some analytical

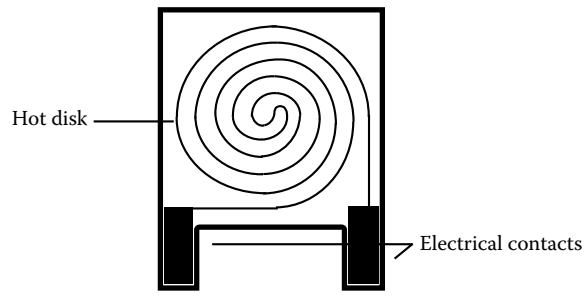


FIGURE 66.4 Schematic diagram of a transient hot-disk instrument.

approximation to it. The technique is used frequently for studies of polymer composites, glasses, superconductors, and insulating materials, and its use is defined by ISO 22007-2 [26].

66.3.8 Transient Hot Bridge

In this technique, which was developed at Physikalisch-Technische Bundesanstalt in Germany, the sensor is realized as a printed circuit foil of nickel between two polyimide sheets [27]. The sensor is itself placed symmetrically between two blocks of the solid material under test. The sensor itself consists of four identical strips arranged parallel to each other on the printed substrate. Each strip is itself composed of an identical short and long segment separated by a very small gap. Two of the strips are placed very close together either side of the central line of the sensor, and two are mounted some distance away from the central line but still symmetrically placed about it. All eight segments of strip are connected into the arms of a Wheatstone bridge so that the bridge is balanced at each and every homogeneous temperature of the sensor. However, when the same current is passed through the eight elements starting at time zero, there is an inhomogeneous heating of the inner and outer strips to an extent determined by the thermal properties of the test material. That differential heating is reflected in an out-of-balance of the bridge, which can be readily detected and related to the thermal conductivity and diffusivity of the test material [27]. The configuration has the advantages over some other techniques of automatically eliminating end effects in the sensor strips, homogeneous temperature drifts of the sample, thermoelectric emfs, and the need for external reference resistors.

66.3.9 Heated-Needle Probe

A further commercial device exists for the measurement of the thermal conductivity of granular materials such as powders and soils, natural materials such as rock and concrete, and, indeed, food. The probe is shown schematically in Figure 66.5, where it is seen that it consists of a thin, hollow, metallic needle (diameter 3 mm) containing an electric heater and a separate thermistor as a probe of the temperature history of the needle following initiation of a heat pulse [28]. The temperature history of the probe is generally interpreted with the aid of the equation appropriate to a transient hot-wire instrument but in a relative manner whereby its response is calibrated against known standards. This rather simplistic approach to the analysis of a somewhat complex cell inevitably restricts the accuracy that can be achieved, but does provide a measurement capability where no other technique is viable. It is often employed for measurements in inhomogeneous samples such as rocks or soils where it is simply the effective thermal conductivity that is required.

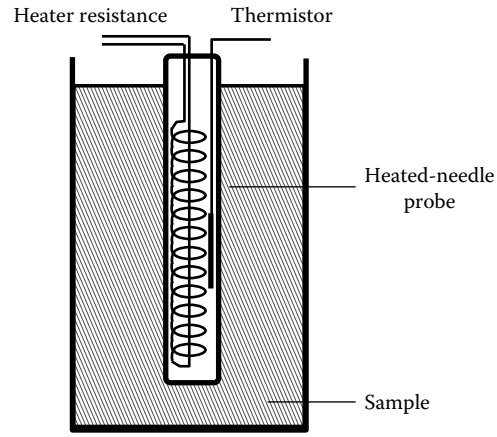


FIGURE 66.5 Schematic diagram of a transient heated-needle probe.

66.3.10 Laser-Flash Instrument

A final transient technique is that which has become known as the laser-flash technique developed originally for measurements in solids but occasionally used on liquids, particularly at high temperatures. The technique has been a standard test method since 1992 under the ASTM designation E1461 [29], and later also under the DIN EN 821 [30] and the ISO 18755 [31] standards.

Figure 66.6 contains a schematic diagram of the instrument as it is available today in a commercial form. The sample is illuminated on one face with a laser pulse of very short duration and high intensity. The absorption of the laser energy on the front face of the sample causes the generation of heat at that front surface, which is subsequently transmitted throughout the sample to the back face of the sample where the temperature rise is detected with an infrared remote sensor. The interpretation of measurements is based on a one-dimensional solution of Equation 66.4 subject to an initial condition of an instantaneous heat pulse at one location.

The temperature rise at the back face of a sample of thickness l and radius r is therefore given by [32]

$$\Delta T(l,t) = \frac{Q}{\rho C_p l \pi r^2}\left[1 + 2\sum_{n=1}^{\infty}(-1)^n \exp\left(-\frac{n^2\pi^2 at}{l^2}\right)\right] \tag{66.7}$$

where Q is the energy absorbed at the front surface at time zero. The thermal diffusivity of the sample, a, is then often deduced from the measurement of the time taken for the back face of the sample to reach one half of its maximum value. The technique has the very distinct advantage that it does not require physical contact between the test sample and the heater or detector. For this reason, it is a particularly appropriate technique for use at high temperatures or in aggressive environments.

However, there are a number of precautions that must be taken to ensure accurate results. First, the theory should be modified to account for non-unidirectional heat flow. Second, care must be taken to ensure that no radiation incident on the front face penetrates to the back face for transparent samples. Due care must also be taken to match the laser power to the system being studied so that there is neither fusion nor ablation at the front face that can distort the results. Finally, when the fluid state is studied, due care should be taken to eliminate convective heat transport. Seldom are all of these precautions adopted in routine work, so that some results obtained with the technique are of dubious validity. Nevertheless, the method has seen widespread application to a wide range of materials, including composites, polymers, glasses, metals, refractory materials, insulating solids, and coatings. Commercially available instruments operating according to this technique are listed in Table 66.1. Finally, the radial heat-flow method should also be mentioned. This is a transient technique in which the sample is heated and cooled continuously [33]. From the recording of the temperature gradient via thermocouples, the thermal diffusivity is obtained and thus the thermal conductivity is calculated. The advantage of the radial heat-flow method is that the measurements are fast and only small temperature gradients are necessary in the sample.

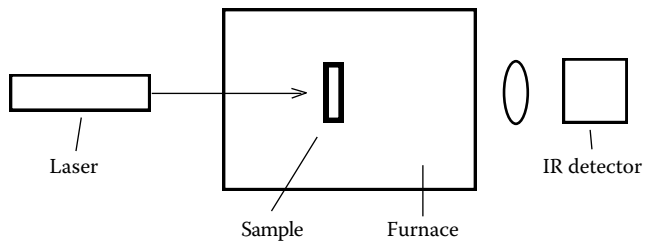


FIGURE 66.6 Schematic diagram of a laser-flash instrument.

TABLE 66.2 Addresses of Companies That Make Thermal Conductivity Instruments

TA Instruments, 159 Lukens Drive, New Castle, DE 19720, USA, Tel.: (302) 427-4000, Fax: (302) 427-4001
HUKSEFLUX Thermal Sensors B.V., Elektronicaweg 25, 2628 XG Delft, The Netherlands, Tel.: (31)-15-2142669, Fax: (31)-15-2574949
K-ANALYS AB, Salagatan 16F, SE-753 30 Uppsala, Sweden, Tel.: (46) 18 59 05 75, Fax: (46) 18 59 05 85
LINSEIS, Messgeraete GmbH, Vielitzerstr. 43, 95100 Selb, Germany, Tel.: (49) (0) 9287/880 0, Fax: (49) (0) 9287/704 88
NETZSCH, Gerätebau GmbH, Wittelsbacherstraβe 42, D-95100 Selb/Bayern, Germany, Tel.: (49) (9287) 88136, Fax: (49) (9287) 88144

66.4 Instrumentation

Table 66.1 lists various instruments for the measurement of the thermal conductivity of solids, while the addresses of the suppliers are shown in Table 66.2. As already mentioned, to our knowledge, there is no company that produces instruments specifically for the measurement of the thermal conductivity of fluids.

66.5 Appraisal

Naturally, the technique to be employed for the measurement of thermal conductivity depends on the type of sample to be studied, the range of conditions to be employed, and the required uncertainty. For fluid samples under most conditions, a variant of the transient hot-wire method must be the preferred technique. Under favorable conditions, an accuracy of ±0.3% can be achieved and a level of ±1% is possible under all but the most aggressive circumstances. Near the critical region of fluids, a parallel-plate instrument is essentially the only viable method. For molten materials at high temperature, while a variant of the hot-wire system has advantages, the laser-flash technique is very attractive, but an accuracy of no better than 10% is then to be expected.

For solids, the hot-disk or laser-flash techniques have many features that recommend them when the sample is amenable to appropriate preparation. In those cases, an accuracy of a few percent should be possible but is rarely attained. For samples such as rocks, the heated-needle probe is undoubtedly the only viable technique.

It should be emphasized again here that when a sample is inhomogeneous, the quantity determined is not the thermal conductivity of any element of it, but rather an effective value suitable for engineering purposes. It is not an intrinsic thermophysical property of the material.

References

1. B. Le Neindre, Contribution à l'étude expérimentale de la conductivité thermique de quelques fluides a haute température et à haute pression, PhD thesis, Université de Paris, Paris, France, 1969.
2. M. Sirota, Steady-state measurements for thermal conductivity, in A. Nagashima, G. V. Sengers, and W. A. Wakeham (eds.), *Experimental Thermodynamics. Vol. III. Measurement of the Transport Properties of Fluids*, London, U.K.: Blackwell Scientific Publications, 1991, Chapter 6.
3. R. Mostert, H. R. van der Berg, and P. S. van der Gulik, The thermal conductivity of ethane in the critical region, *J. Chem. Phys.*, 92, 5454–5462, 1990.
4. E. Sakonidou, The thermal conductivity of methane and ethane mixtures around the critical point, PhD thesis, University of Amsterdam, Amsterdam, the Netherlands, 1996.
5. ASTM Standard E1530, 1999 (2006), Standard test method for evaluating the resistance to thermal transmission of materials by the guarded heat flow meter technique, ASTM International, West Conshohocken, PA, 2006.

6. ASTM Standard C177, 1997 (2010), Standard test method for steady-state heat flux measurements and thermal transmission properties by means of the guarded-hot-plate apparatus, ASTM International, West Conshohocken, PA, 2010.
7. ISO 8302, Thermal insulation—Determination of steady-state thermal resistance and related properties—Guarded hot plate apparatus, Beuth Verlag GmbH, Berlin, Germany, 1996; M. J. Assael, K. D. Antoniadis, and W. A. Wakeham, Historical evolution of the transient hot-wire technique, *Int. J. Thermophys*., 31, 1051–1072, 2010.
8. M. J. Assael, C. A. Nieto de Castro, H. M. Roder, and W. A. Wakeham, Transient methods for thermal conductivity, in A. Nagashima, J. V. Sengers, and W. A. Wakeham (eds.), *Experimental Thermodynamics. Vol. III. Measurement of the Transport Properties of Fluids*, London, U.K.: Blackwell Scientific Publications, 1991, Chapter 7.
9. B. Stâlhane and S. Pyk, Ny Metod för Bestämning Av Värmelednings-Koefficienter, *Teknisk Tidskrift*, 61, 389–398, 1931.
10. J. W. Haarman, Thermal conductivity measurements with a transient hot-wire method, PhD thesis, Technische Hogeschool Delft, Delft, the Netherlands, 1969.
11. M. J. Assael, C.-F. Chen, I. Metaxa, and W. A. Wakeham, Thermal conductivity of suspensions of carbon nanotubes in water, *Int. J. Thermophys*., 25, 971–985, 2004.
12. A. Alloush, W. B. Gosney, and W. A. Wakeham, A transient hot-wire instrument for thermal conductivity measurements in electrically conducting liquids at elevated temperatures, *Int. J. Thermophys*., 3, 225–234, 1982.
13. Y. Nagasaka and A. Nagashima, Absolute measurements of the thermal conductivity of electrically conducting liquids by the transient-hot wire method, *J. Phys*., *E*, 14, 1435, 1981.
14. P. Andersson and G. Bäckström, Measurement of the thermal conductivity under high pressures, *Rev. Sci. Instrum*., 47, 205, 1976.
15. ASTM C1113, 1999 (2009) Standard test method for thermal conductivity of refractories by hot wire (platinum resistance thermometer technique), ASTM International, West Conshohocken, PA, 2009.
16. EN 993-14, Methods of test dense shaped refractory products—Determination of thermal conductivity by the hot-wire (parallel) method, European Committee for Standardization, Brussels, Belgium, 2005.
17. EN 993-15, Methods of test dense shaped refractory products—Determination of thermal conductivity by the hot-wire (cross-array) method, European Committee for Standardization, Brussels, Belgium, 1998.
18. ISO 8894-1, Refractory materials—Determination of thermal conductivity—Part 1: Hot-wire methods (cross-array and resistance thermometer), Beuth Verlag GmbH, Berlin, Germany, 2010.
19. ISO 8894-2, Refractory materials—Determination of thermal conductivity—Part 1: Hot-wire methods (parallel), Beuth Verlag GmbH, Berlin, Germany, 2007.
20. DIN 51046, Testing of ceramic materials: Determination of thermal conductivity up to 1600°C by the hot-wire method thermal conductivity up to 2 W/m/K, Deutsches Institut für Normung, Berlin, 1976.
21. M. J. Assael, K. D. Antoniadis, K. E. Kakosimos, and I. N. Metaxa, An improved application of the transient hot-wire technique for the absolute accurate measurement of the thermal conductivity of pyroceram 9606 up to 420 K, *Int. J. Thermophys*., 29, 445–456, 2008.
22. M. V. Peralta-Martinez, M. J. Assael, M. Dix, L. Karagiannidis, and W. A. Wakeham, A novel instrument for the measurement of the thermal conductivity of molten metals. Part I. Instrument's description, *Int. J. Thermophys*., 27, 353–375, 2006.
23. M. V. Peralta-Martinez, M. J. Assael, M. Dix, L. Karagiannidis, and W. A. Wakeham, A novel instrument for the measurement of the thermal conductivity of molten metals. Part II. Measurements, *Int. J. Thermophys*., 27, 681–699, 2006.
24. M. Dix, I. W. Drummond, M. Lesemann, M. V. Peralta-Martinez, W. A. Wakeham, M. J. Assael, L. Karagiannidis, and H. R. van den Berg, The thermal conductivity of liquid metals near ambient conditions, *Proc. 5th Asian Thermophysical Properties Conference*, Seoul, Korea, pp. 133–136, 1998.

25. S. E. Gustafsson, E. Karawacki, and M. N. Khan, Transient hot-strip method for simultaneously measuring thermal conductivity and thermal diffusivity of solids and fluids, *J. Phys. D*: *Appl. Phys*., 12, 1411, 1979.
26. ISO 22207, Plastics—Determination of thermal conductivity and thermal diffusivity—Part 2: Transient plane heat source (hot disc) method, Beuth Verlag GmbH, Berlin, Germany, 2008.
27. U. Hammerschmidt and V. Meier, A new pulse hot strip sensor for measuring thermal conductivity and thermal diffusivity of solids, *Int. J. Thermophys*., 27, 840–865, 2006.
28. J. Nicolas, Ph. André, J. F. Rivez, and V. Debaut, Thermal conductivity measurements in soil using an instrument based on the cylindrical probe method, *Rev. Sci. Instrum*., 64, 774–780, 1993.
29. ASTM Standard E1461, 1992 (2007), Standard test method for thermal diffusivity by the flash method, ASTM International, West Conshohocken, PA, 2007.
30. EN 821-2, Advanced technical ceramics—Monolithic ceramics—Thermo-physical properties—Part 2: Determination of thermal diffusivity by the laser flash (or heat pulse) method, European Committee for Standardization, Brussels, Belgium, 1997.
31. ISO 18755, Refractory fine ceramics (advanced ceramics, advanced technical ceramics)—Determination of thermal diffusivity of monolithic ceramics by laser flash method, Beuth Verlag GmbH, Berlin, Germany, 2005.
32. W. J. Parker, R. J. Jenkins, C. P. Butler, and G. L. Abbott, Flash method for determining thermal diffusivity, heat capacity and thermal conductivity, *J. Appl. Phys*., 32, 1679–1684, 1960.
33. G. S. Sheffield and J. R. Schorr, Comparison of thermal diffusivity and thermal conductivity methods, *Ceram. Bull*., 70, 102, 1991.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

67

Heat Flux

67.1 Heat Transfer Fundamentals **67**-2
Conduction • Convection • Radiation
67.2 Heat Flux Measurement..... **67**-4
67.3 Sensors Based on Spatial Temperature Gradient..... **67**-5
One-Dimensional Planar Sensors • Insert Gages
67.4 Sensors Based on Temperature Change with Time..... **67**-9
Semi-Infinite Surface Temperature Methods • Calorimeter Methods
67.5 Measurements Based on Surface Heating..... **67**-12
67.6 Calibration and Problems to Avoid **67**-12
67.7 Summary..... **67**-14
References..... **67**-14

Thomas E. Diller
Virginia Tech

Energy use is coming under increasing scrutiny as the effects of carbon dioxide on the environment are being factored into the cost. Consequently, thermal management is becoming a sophisticated science in modern society. It has become accepted that living spaces should be heated and cooled for maximum comfort of the occupants. Many industrial manufacturing processes require tight temperature control of material throughout processing to establish the desired properties and quality control. Examples include control of thermal stresses in ceramics and thin films, plasma deposition, annealing of glass and metals, heat treatment of many materials, fiber spinning of plastics, film drying, growth of electronic films and crystals, and laser surface processing.

Temperature control of materials requires that energy be transferred to or from solids and fluids in a known and controlled manner. Consequently, the optimal control of the heat transfer in equipment such as dryers, heat exchangers, boilers, condensers, and data management centers becomes crucial. The constant drive toward higher power densities in electronic, propulsion, and electric generation equipment continually pushes the limits of the associated cooling systems.

Although the measurement of temperature is common and well accepted, the measurement of heat flux is often given little consideration. Temperature is one of the fundamental properties of a substance. Moreover, because temperature can be felt by human senses, most people are familiar with its meaning. Conversely, heat flux is a derived quantity that is not easily sensed. How and where thermal energy is transferred is often equally or more important than the temperature. For example, the temperature of human skin can indicate the comfort level of the person, but has little connection with the energy being dissipated to the surroundings, particularly if evaporation is occurring simultaneously. Wind chill factor is a common example of the importance of convection heat transfer in addition to air temperature.

Controlling the thermal energy transfer in many systems is crucial to their optimum performance. Consequently, sensors that can be used to directly sense heat flux can be extremely important. It is amazing that heat flux sensors are not more widely used in modern technology. The subsequent material in this chapter is intended to help individuals understand and implement heat flux measurements that are best suited for the required applications.

67.1 Heat Transfer Fundamentals

The movement of thermal energy is known as "heat," and the rate of this transfer is commonly called "heat transfer." It is given the symbol q and has the units of watts. The heat transfer per unit area is termed the "heat flux" and is given the symbol q'' with the units of W m^{-2}. Although in some cases only the overall heat transfer from a system is required, often the spatial and temporal variations of the heat flux are important to performance enhancement. Methods for measuring these distributions of heat flux are identified and discussed in this chapter.

One of the most important principles concerning heat transfer is the first law of thermodynamics, which states that the rate of energy storage is equal to the overall energy transfer to and from a system. It includes all types of energy transfer across the system boundary, including the three modes of heat transfer—conduction, convection, and radiation. For the simple example shown in Figure 67.1, the transient energy balance for a control volume around the surface can be expressed as

$$mC\frac{\partial T}{\partial t} = q_{\text{convection}} + q_{\text{radition}} - q_{\text{conduction}} \tag{67.1}$$

where

m is the mass of the material included
C is the corresponding specific heat

The effect of the thermal capacitance (mC) of the material causes a time lag in the temperature response of the material (T) to a change in heat transfer. A short summary of important heat transfer principles follows, with many engineering textbooks available in the field that give additional details [1].

67.1.1 Conduction

Conduction encompasses heat transfer through stationary materials by electrons and phonons. It is related to the temperature distribution by Fourier's law, which states that the heat flux vector is proportional to and in the opposite direction of the temperature gradient:

$$\bar{q}'' = -k\bar{\nabla}T \tag{67.2}$$

The constant k is the thermal conductivity of the material. Measuring this temperature gradient is one of the basic methods for determining heat flux.

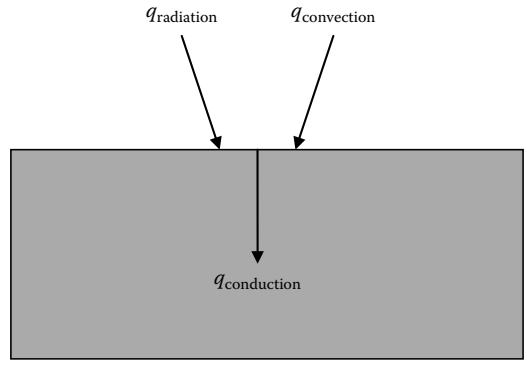


FIGURE 67.1 Illustration of energy balance.

For a homogeneous material in Cartesian coordinates, Equation 67.1 becomes

$$\frac{\partial T}{\partial t} = a\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}\right) \tag{67.3}$$

where a is the thermal diffusivity of the material with a density of ρ, $a = k/(\rho C)$. Measuring the transient temperature response of the system according to this equation is the second major method for determining the heat transfer.

As illustrated in Figure 67.1, convection and radiation are the modes of heat transfer at the boundary of a solid with conduction. They are usually the quantities of interest to measure with a heat flux sensor.

67.1.2 Convection

Although heat transfer by convection occurs by the same physical mechanisms as conduction, the fluid is free or forced to move relative to the surface. The fluid motion greatly complicates the analysis by coupling the heat transfer problem with the fluid mechanics. Particularly when the flow is turbulent, the fluid equations are generally impossible to solve exactly. Consequently, the heat transfer and fluid mechanics are commonly isolated by the introduction of a heat transfer coefficient, which encompasses all of the fluid flow effects.

$$q'' = h_{\mathrm{T}}(T_{\mathrm{r}} - T_{\mathrm{s}}) \tag{67.4}$$

The temperature of the fluid is represented by T_{r}, which for low-speed flows is simply the fluid temperature away from the surface. The recovery temperature is used for high-speed flows because it includes the effect of frictional heating in the fluid. The subscript T on the heat transfer coefficient, h_{T}, implies that the boundary condition on the surface is a constant temperature, T_s. Although other surface temperature conditions can be encountered, it is then important to carefully document the surface temperature distribution because it can have a profound effect on the values of h and q'' [2].

67.1.3 Radiation

Heat transfer by radiation occurs by the electromagnetic emission and adsorption of photons. Because this does not rely on a medium for transmission of the energy, radiation is very different from conduction and convection. Radiation has a spectrum of wavelengths dependent on the temperature and characteristics of the emitting surface material. Moreover, the surface properties are often dependent on the wavelength and angular direction of the radiation. Special coatings are sometimes applied to surfaces to control the absorption characteristics. For example, the surface of a radiation sensor is often coated with a high absorptivity paint or graphite.

Because the power emitted from a surface is proportional to the fourth power of the absolute temperature, radiation detectors are usually cooled sufficiently for the power emitted from the detector itself to be negligible. In this case, the temperature distribution of the sensor is not important and the heat flux is equal to the incident radiation:

$$q'' = \sigma T^4 \tag{67.5}$$

where

σ is the Stefan–Boltzmann constant

T is the temperature of the surroundings

This is a big advantage over convection measurements, where the temperature distribution on the surface has a big influence on the measurement.

67.2 Heat Flux Measurement

Most of the methods for measuring heat flux are based on temperature measurements on the surface or in the solid material. Usually this involves insertion of a device either onto or into the surface, which has the potential to cause both a physical disruption and a thermal disruption of the surface. As with any good sensor design, the goal for good measurements must be to minimize the disruption caused by the presence of the sensor. It is particularly important to understand the thermal disruption caused by the sensor because it cannot be readily visualized and because all heat flux sensors have a temperature change associated with the measurement. Consequently, wise selection of the sensor type and operating range is important for good heat flux measurements [2,3]. The following sections emphasize important factors in using the currently available heat flux sensors. They are grouped by the general type of sensor action. The commercially available gages are listed in Table 67.1 with the general characteristics of the system.

TABLE 67.1 Heat Flux Instrumentation

Manufacturer	Sensor	Description	Mounting Method
RdF	Micro-foil	Foil thermopile	Surface
Concept	Heat flow sensor	Wire-wound thermopile	Surface
MesoScribe Technologies	Heat flux transducer	Direct write thermal spray	Surface
ITI	Thermal flux meter	Thermopile	Surface
Vatell	HFM	Microsensor thermopile	Insert
Medtherm, Vatell	Gardon gage	Circular foil design	Insert
Medtherm, Vatell, Hukseflux	Schmidt-Boelter	Wire-wound thermopile	Insert
Medtherm	Coaxial thermocouple	Transient temperature	Insert
Medtherm	Null-point calorimeter	Transient temperature	Insert
Hallcrest	Liquid crystals	Temperature measurement kit	Surface

TABLE 67.2 Companies That Make Sensors for Heat Flux Measurement

Concept Engineering
43 Ragged Rock Road
Old Saybrook, CT 06475
(860) 388–5566
http://conceptengineering2000.com/

International Thermal Instrument Co.
P.O. Box 309
Del Mar, CA 92014
(858) 755–4436
http://www.iticompany.com/

Medtherm Corporation
P.O. Box 412
Huntsville, AL 35804
(205) 837–2000

Hukseflux Thermal Sensors
Elektronicaweg 25
2628 XG Delft
The Netherlands
31-15-2142669
http://www.hukseflux.com/

RdF Corporation
P.O. Box 490
Hudson, NH 03051-9981
(603) 882–5195
www.rdfcorp.com/

Vatell Corporation
P.O. Box 66
Christiansburg, VA 24073
(540) 961–3576
http://vatell.com/

LCR Hallcrest
1911 Pickwick Lane
Glenview, IL 60026
(8472) 998–8580
http://hallcrest.com/

MesoScribe Technologies, Inc.
7 Flowerfield, Suite 28
St. James, NY 11780
631-686-5710
http://www.mesoscribe.com/

Prices have not been included because there are usually many specific variations of gages with different costs. They are usually made individually rather than mass produced. Contact information for the manufacturers is given in Table 67.2. Many also have dealers and representatives located around the world.

67.3 Sensors Based on Spatial Temperature Gradient

The heat flux at the material surface can be found at a location if the temperature gradient can be determined at that position, as indicated in Equation 67.2. Because it is difficult to position temperature sensors with the requisite accuracy inside the material, sensors to measure heat flux are either applied on the surface or mounted in a hole in the material.

67.3.1 One-Dimensional Planar Sensors

A simple heat flux sensor concept for mounting on a surface is illustrated in Figure 67.2. The one-dimensional heat flux perpendicular to the surface is found from Equation 67.3 for steady-state conditions:

$$q'' = \frac{k}{\delta}(T_1 - T_2) \tag{67.6}$$

The thickness of the sensor δ and thermal conductivity k are not known with sufficient accuracy for any particular sensor to preclude direct calibrations of each sensor. An adhesive layer may also be required between the sensor and the surface to securely attach the sensor, which adds an additional thermal resistance and increases the thermal disruption. Temperature measurements on the sensor and on the surrounding undisturbed material are recommended to quantify this disruption. A new ASTM standard [4] gives guidance on the use of these sensors.

Although the temperature difference can be measured by any number of methods, the most commonly used are thermocouples. Thermocouples have the advantage that they generate their own voltage output corresponding to the temperature difference between two junctions. Consequently, they can be connected in series to form a thermopile that amplifies the output from a given temperature difference, as illustrated in Figure 67.3. Most any pair of conductors (e.g., copper–constantan) can be used for the legs of the thermopile, but the output leads should be of the same material so that additional thermocouple junctions are not created. The voltage output, E, is simply

$$E = NS_T(T_1 - T_2) \tag{67.7}$$

where

N represents the number of thermocouple junction pairs

S_T is the Seebeck coefficient or thermoelectric sensitivity of the materials, expressed in volts per degree Centigrade

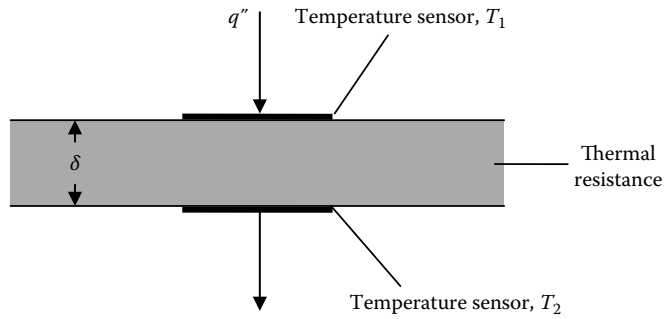


FIGURE 67.2 One-dimensional planar sensor concepts.

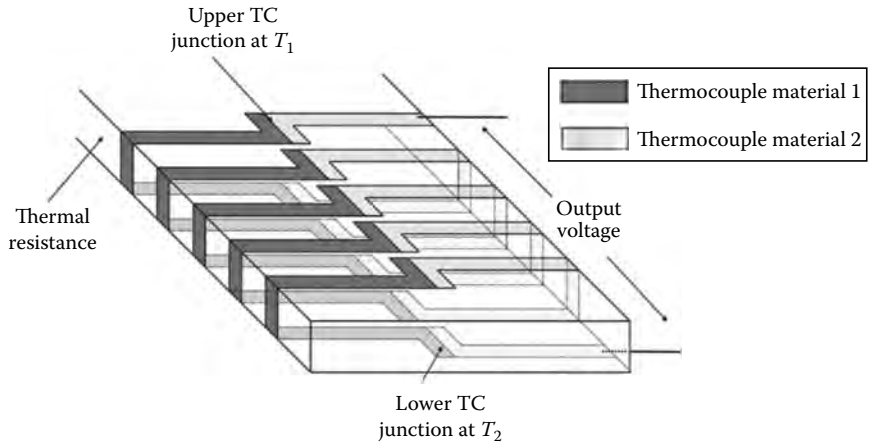


FIGURE 67.3 Thermopile heat flux sensor.

The corresponding sensitivity of the heat flux sensor is

$$S = \frac{E}{q''} = \frac{NS_T\delta}{k} \tag{67.8}$$

Although the sensitivity is determined in practice from a direct calibration, the last part of the equation can be used to determine the effects of different parameters for design purposes.

One successful design using a thermopile is the Micro-Foil HFS currently manufactured by RdF Corp. Thin pieces of two types of metal foil are alternately wrapped around a thin plastic (Kapton) sheet and butt-welded on either side to form thermocouple junctions, as illustrated in Figure 67.3. A separate thermocouple is included to provide a measure of the sensor temperature. The flexible, micro-foil sensors are 75–400 μm thick and can be glued to a variety of surface shapes, but are limited to temperatures below (250°C) and heat fluxes less than 100 kW m^{-2}. This covers many general-purpose industrial and research applications. The time response can be as fast as 20 ms, but transient signals can be attenuated unless the frequency of the disturbance is less than a few hertz. These sensors act much like a first-order system, which has 70% amplitude response to a sinusoidal input with a period six times the exponential time constant.

If the individual legs of the sensor are left in parallel instead of in series, an array of heat flux sensors is created. This was done by Hubble and Diller [5] using copper and constantan for the thermocouple materials. The time constants were between 100 and 200 ms when using a 50 μm thick layer of Kapton for the thermal resistance. The output from the 10-sensor array was arranged to give both the heat flux (from the differential temperature) and surface temperature at each location.

A high-temperature heat flux gage (HTHFS) has been developed by MesoScribe Technologies, Inc., which can be deposited directly onto test surfaces using a Direct Write Thermal Spray process. Up to 80 thermocouple pairs (Type K or N) are scribed around a layer of dielectric serving as the thermal resistance for the heat flux with a separate thermocouple giving the surface temperature. A layer of ceramic dielectric is first deposited on a metal surface to provide electrical isolation and can also be deposited over the gage to encapsulate it for better durability. All of the layers together are less than 0.4 mm thick, which gives the sensor a low profile. The sensor is sufficiently thick, however, for good survivability at high temperatures, and it gives a good output signal. The sensor can also be embedded in existing barrier coatings used for thermal, environmental, or erosion resistance.

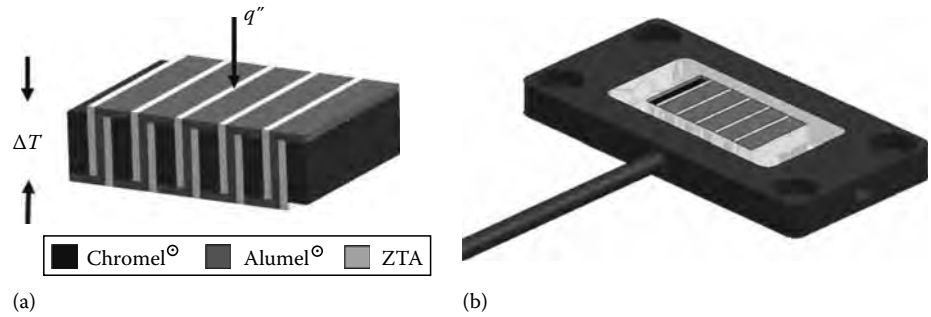


FIGURE 67.4 HTHFS: (a) schematic of the measuring element, (b) complete sensor.

A robust HTHFS is made by welding Chromel and Alumel strips [6] as illustrated in Figure 67.4. The thermal resistance between the top and bottom thermocouples is formed by the elements themselves. Ceramic (ZTA) strips are included to provide electrical insulation between elements. An additional thermocouple wire welded to the top of the gage provides the surface temperature, which is useful for interpreting the heat flux signal. The gage can be operated and cycled to over 1000 °C. The thickness is about 3 mm with a thermal time constant less than 1 s. Because of the all-metal construction, the thermal resistance is small. The measurement chip can also be mounted in an insert type housing that can be air or water cooled for gage temperature control.

A design using welded wire to form the thermopile across the thermal resistance layer is made by International Thermal Instrument Co. The sensors are about 3 mm thick with thermal time constants on the order of 1 s and an upper temperature limit of 300 °C. Typical applications include heat transfer in buildings and physiology. Sensors with higher sensitivity are made with semiconductor thermocouple materials for geothermal applications. Sensors are also available for operating temperatures up to 1000 °C.

Another technique for measuring the temperature difference across the thermal resistance layer is to wrap wire and then plate one side of it with a different metal. A common combination is constantan wire with copper plating. The resulting wire-wound sensor looks similar to the sensor shown in Figure 67.3, except that the constantan wire is continuous all around the sensor. A summary of the theory is given by Hauser [7], and a general review is given by van der Graaf [8]. Concept Engineering offers a range of these types of sensors at moderate cost. Because of the many wire windings around 2 mm thick plastic strips, the sensitivity to heat flux is high. The corresponding thermal resistance is also large and time constants are around 1 s. Temperatures are limited to about 150 °C. Measured heat flux levels are usually less than 1 kW m^{-2}, with applications including building structures, insulation, geothermal, and medicine.

Heat flux gages are normally mounted on a good heat sink to provide a pathway for the thermal energy to flow. Not all real applications provide such an ideal situation, however. Hubble and Diller [9] have shown with a simple modification of the signal how to use heat flux gages on any surface, including thermal insulators. Adding a second term to account for the energy storage in the sensor itself creates what is termed the "hybrid heat flux"

$$q'' = q''_{\text{dif}} + \frac{1}{2} q''_{\text{cal}} \tag{67.9}$$

The first term is the usual differential signal from the sensor, while the second term is the transient response representative of the signal from a calorimeter. This gives the correct heat flux whether the gage is mounted on a good heat sink (high-conductivity material) or a good insulator (low-conductivity material). It also increases the time response of the gage by about an order of magnitude. The differential heat flux is proportional to the usual temperature difference measured across the gage, and the calorimeter heat flux is proportional to the time rate of temperature change of the gage.

67.3.2 Insert Gages

Insert gages are mounted through a hole flush with the surface of the material. Their use is covered by a new ASTM standard [10]. One popular version uses a plated wire wrapped around a small anodized piece of aluminum that is potted into a circular housing, commonly known as a Schmidt-Boelter gage. Kidd and Nelson [11] have analyzed these gages to determine the effect of the plating relative to a true differential thermopile. The sensors are commercially available from Medtherm, Vatell, and Hukseflux in sizes as small as 1.5 mm diameter. There is also some ability to contour the surface of the sensor to match a curved model surface for complex test article shapes.

A thin thermopile sensor called the Heat Flux Microsensor (HFM) is manufactured by Vatell Corp. Because it is made with thin-film sputtering techniques, the entire sensor is less than 2 μm thick. The thermal resistance layer of silicon monoxide and the metal layers are sputtered directly onto the aluminum nitride surface of the gage. The resulting physical and thermal disruption of the surface due to the presence of the sensor is extremely small. They are best suited for heat flux values above 1 kW m^{-2}, with no practical upper limit. Because the sensor is so thin, the thermal response time is less than 10 μs [12], giving a good frequency response well above 1 kHz. A temperature measurement that is integrated into the sensor is very useful for determining the heat transfer coefficient. The high temperature and fast time response capabilities are useful for aerodynamic applications, combusting flows in engines and propulsion systems, and capturing high-speed events such as shock passage.

The circular foil or Gardon gage consists of a hollow cylinder of one thermocouple material with a thin foil of a second thermocouple material attached to one end. A wire of the first material is attached to the center of the foil to complete a differential thermocouple pair between the center and edge of the foil as illustrated in Figure 67.5. The body and wire are commonly made of copper with the foil made of constantan. Heat flux to the gage causes a radial temperature distribution along the foil as measured by this single thermocouple pair. These sensors are covered by an ASTM standard [13]. They are manufactured by two companies at moderate cost (Medtherm and Vatell) and are often used as secondary standards for measurement of radiation. The biggest problems with the circular foil gages arise when they are used with any type of convection heat transfer. It has been shown analytically and experimentally that the output is altered for convective heat transfer because of the distortion of the temperature profile in the foil from the parabolic profile of radiation [14]. Because the amount of error is a function of the gage geometry, the fluid flow, and the heat transfer coefficient, it is difficult to correct reliably. The errors become particularly large when the sensor is used in a flow that has a shear flow component, which encompasses most convection situations. Consequently, great care must be used to keep the temperature difference across the gage small if Gardon gages are used to measure convective heat transfer.

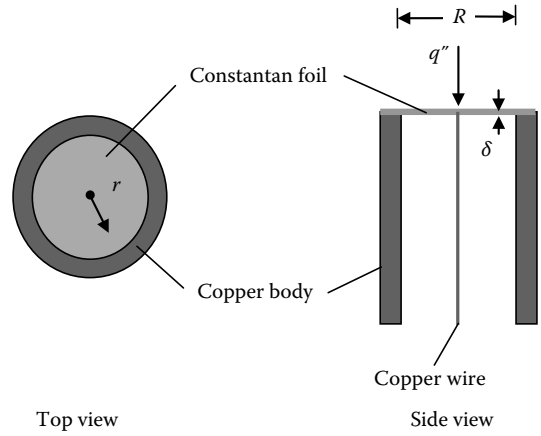


FIGURE 67.5 Circular foil heat flux gage.

When insert heat flux gages are used in high heat flux situations, such as combustors, water cooling is usually supplied through the body of the sensor to keep the temperature from exceeding material limits. Because of the resulting temperature mismatch of the gage and surrounding material in which it is mounted, a water-cooled gage is not recommended for convection heat transfer measurements. It is also important for a water-cooled gage to ensure that condensation does not occur on the sensor face.

Although most heat flux sensors are designed to measure the total heat flux, sensors have been developed to separate convection from radiation. The most common method is to put a transparent window over the sensor to eliminate convection to the sensor face. Because the resulting sensor only measures radiation, it is termed a radiometer. The field of view is limited, however, in these radiometers and must be included in the interpretation of results. In a dirty environment where the transmission of the window could be degraded, air is blown across the face of the window to keep the particles away from the sensor. Applications include use in high-ash boilers and gas turbine combustors.

67.4 Sensors Based on Temperature Change with Time

Equations 67.1 through 67.3 give the form of the relationship between the unsteady response of temperature and surface heat transfer. If the thermal properties of the wall material are known along with sufficient detail about the temperature history and distribution, the heat transfer as a function of time can, in principle, be determined. Although temperature sensors are available from manufacturers, the necessary data manipulation must be done by the user to obtain heat flux. There are two types of solutions used to reduce the temperature history to heat flux. These are discussed separately as the semi-infinite solution and calorimeter methods. In addition, a variety of methods for measuring the required temperature history are discussed.

67.4.1 Semi-Infinite Surface Temperature Methods

An important technique for short-duration heat flux tests is to measure the surface temperature history on a test object with a fast-response temperature sensor. For short enough times and sufficiently thick material, it can be assumed that the transfer is one-dimensional and that the thermal effects do not reach the back surface of the material. Equation 67.3 reduces to a one-dimensional, semi-infinite solution, which is simple to implement for this case. For example, the surface temperature for a step change of heat flux at time zero is

$$T_s - T_i = \frac{2q_o''\sqrt{t}}{\sqrt{\pi k\rho C}} \tag{67.10}$$

The substrate properties are the thermal conductivity, density, and specific heat, represented by the product $k\rho C$. T_i is the uniform initial temperature of the substrate and T_s is the surface temperature as a function of time. The effects of both energy storage in the material and conduction through the material are included.

Data analysis of the measured temperature record to find the corresponding heat flux can be performed by several methods. The simplest is to use the analytical solution with each sampled data point to recreate the heat flux signal. The most popular equation for this conversion is that attributed to Cook and Felderman [15] for uniformly sampled data:

$$q''(t_n) = \frac{2\sqrt{k\rho C}}{\sqrt{\pi\Delta t}} \sum_{j=1}^{n} \frac{T_j - T_{j-1}}{\sqrt{n-j} + \sqrt{n+1-j}} \tag{67.11}$$

This can easily be implemented for digital data with a short computer program to perform the summation for the measured points. Numerical solutions are used to account for changes in property values if there are large temperature changes. One problem is the inherent instability of using a temperature record to infer the heat flux. This is akin to taking the derivative of the temperature signal, which greatly amplifies any noise present in the signal and is particularly important when making detailed measurements of time-resolved heat flux. There are several methods to increase the stability of the signal, such as that reported by Oldfield [16].

A good criterion for the time limit on the test is the time before 1% error occurs in the heat flux:

$$t = 0.3\frac{L^2}{a} \tag{67.12}$$

where

a represents the thermal diffusivity of the substrate material
L is the substrate thickness

For a typical ceramic substrate (MACOR), the corresponding minimum thickness for 1 s of test time is 1.6 mm. High-conductivity materials, such as metals, have much larger required thicknesses.

The required surface temperature can be measured with a number of different techniques. Two broad categories are point measurements using thermocouples or resistance temperature devices (RTDs) and optical methods that allow for simultaneous measurement of temperatures over the entire surface. They all require substantial effort from the user to initiate the test procedure and reduce the data to find the heat flux.

Point temperature measurements for determining convective heat flux are often made with thin-film RTDs. A metallic resistance layer is sputtered, painted and baked, or plated onto the surface. Because the resulting thickness of the sensor is less than 0.1 μm, the response time is a fraction of 1 μs, and there should be no physical or thermal disruption of the measured temperature due to the sensor. Most researchers develop techniques for instrumenting models themselves. An alternative is to deposit RTDs onto a thin plastic sheet and then glue that to the surface [17]. This may require consideration of the different properties of the multiple layers of plastic, adhesive, and model. Transient flow facilities provide an easy method for quickly initiating the flow and heat transfer, as required by this transient method. However, the model can also be injected into the flow or the flow can be quickly diverted to provide the fast initiation of heat flux. The method is used for basic measurements applicable to gas turbine engines, rockets, internal combustion engines, and high-speed aerodynamics [2].

A special type of thermocouple called coaxial thermocouples [18] is specifically made for transient surface temperature measurement. It has one thermocouple wire inside the second thermocouple material with an insulating layer in between. One end is mounted into a metal sheath for press fitting into the surface material for testing. A thin thermocouple is formed by combining the two metals right at the end of the assembly. Response times are typically 1 ms or less, which although slower than the thin-film sensors is sufficient for most applications. The cost per sensor is moderate, and they are available from Medtherm Corp. The average properties of the sensor should be matched with those of the substrate material.

Null-point calorimeters [19] are a further extension of the semi-infinite surface temperature method. They are designed for the measurement of extremely high levels of heat flux (over 1000 kW m^{-2}). To protect the thermocouple and wires, they are mounted in a cavity behind the surface. The geometry of the null-point calorimeter is designed, however, for the thermocouple to measure a temperature that would match the surface temperature of a semi-infinite material so that Equation 67.11 can be used for data reduction [19]. Medtherm is the current supplier of null-point calorimeters.

Optical methods give the opportunity to measure the entire temperature field over the surface of interest. Consequently, much data can be collected over each test, but interpretation to obtain

quantitative heat flux values is more challenging than measurements with point sensors. The visual display of the temperature over the surface is very appealing and can enhance the interpretation of data by seeing the distribution of temperature on a surface in addition to its numerical value.

One popular optical method for measuring temperature is to record the color change using liquid crystals. These are specially prepared molecules that reversibly change their color reflection through several distinct colors as a function of temperature, typically in the range of 25 °C–40 °C. The best for transient measurements are the chiral-nematic form, which has been microencapsulated to stabilize their properties. A variety of types can be obtained from Hallcrest. They can easily be spray-painted onto a blackened surface for testing. Setting the lighting for reproducible color, temperature calibration, image acquisition, and accurately establishing the starting temperature are crucial steps [20]. The basic materials are cheap, but the associated equipment is expensive, including a high-quality video camera, lighting system, calibration system, computer hardware, and software for image processing.

As previously discussed for radiation heat transfer, all surfaces emit radiation with an intensity and wavelength distribution that can be related to the surface temperature. The advent of high-speed infrared scanning radiometers has made it feasible to record the transient temperature field for the determination of the transient heat flux distribution [21]. A coating is usually applied to establish a known high-emissivity surface. To convert the measured radiation emission to surface temperature, the radiation field of the surroundings is also required because reflections can be important.

Thermographic phosphors emit radiation in the visible spectrum when illuminated with ultraviolet light. The intensity of emission at specific wavelengths can be related to the temperature over a wide range of surface temperatures [22]. The potential high-temperature applications are particularly appealing. A CCD camera is required to record the transient optical images, and calibration is challenging.

The temperature-sensitive paint method uses a Europium complex as the luminophore with a mercury-xenon arc lamp for illumination [23]. A CCD camera is used to collect the images as a function of time.

67.4.2 Calorimeter Methods

A *calorimeter* is a device for measuring the amount of absorbed thermal energy. A slug calorimeter [24] uses the rate of temperature change in a known mass of material to determine the heat flux at the surface. It is used as an insert gage mounted flush with the surface of a material and isolated thermally from the material. This is done with a small air gap around the calorimeter and insulation on the back side where the thermocouple for temperature sensing is attached. Although there is a temperature gradient through the material of the slug, after a short transient, the rate of temperature change in the gage is uniform. When exposed to a constant heat flux, the rate of temperature change will be constant and can be related to the heat transfer by

$$q = mC\frac{dT}{dt} + q_{\text{losses}} \tag{67.13}$$

where the product of mass (m) and specific heat (C) is the thermal capacitance of the slug. The heat losses should be kept small unless they can be quantified. One way to determine the losses is to monitor the rate of temperature cooling after the heat source is removed. Although slug calorimeters are rarely used currently, they can provide an important standard in the calibration process.

A similar device is the differential flame thermometer [25], which consists of two inconel plates separated by a layer of insulation with a thermocouple mounted on each plate. Plates are typically 120 mm square and the insulation is 19 mm thick. The resulting time constant for the system is about 90 s. The application is in a radiation-dominated furnace or combustor to measure the heat flux and steady-state temperature. Inverse heat transfer methods are used to convert the two temperature records into heat flux. It should be noted that this device is different from a standard heat flux gage because it is suspended in the furnace or environment rather than mounted on or in a surface.

67.5 Measurements Based on Surface Heating

For research on convective heat transfer, electric heating provides an easy method of controlling and measuring the heat flux from a surface. A combination of guard heaters and insulation allows control of the heat losses to give an accurate knowledge of the heat flux leaving the surface based on the total electrical power supplied once steady-state conditions have been established. The temperatures of the surface and fluid are used in Equation 67.4 to give the heat transfer coefficient for the surface:

$$h = \frac{q''}{T_r - T_s} \tag{67.14}$$

The experimental techniques can be grouped according to the two boundary conditions that normally result: constant heat flux and constant wall temperature. In constant heat flux cases, the experiment usually runs until a steady-state temperature distribution is achieved. For constant wall temperature, manual or active control of the power is required to stabilize the system at the prescribed constant temperature.

For constant heat flux conditions, a thin metal sheet is often used as the heater to provide uniformity of resistance (and heat flux) and to minimize the lateral conduction. Several centimeters of good insulation as the backing material usually insures acceptable heat loss from the back surface. Unfortunately, the response time for this arrangement is typically very long, depending on the surface convection and substrate properties. With a surface heat transfer coefficient of 100 $W/m^2 \cdot K$ and a good substrate insulator, the time from a cold start to reach steady-state conditions (within 2%) is typically over 1 h. The time limitation is the result of transient heat transfer into the insulation backing layer.

Surface temperature measurements are not needed to determine the heat flux, but they are required for the determination of the heat transfer coefficient. Although any of the methods for measuring temperature are possible candidates, only thermocouples, RTDs, infrared cameras, and liquid crystals are commonly used. Thermocouples give point measurements and must be carefully mounted to avoid losses along the wires. Resistance elements can be used to both supply the heating and to measure the average temperature of the element. Infrared radiation measurements can give the entire surface distribution. Liquid crystals can also give the surface distribution and can even measure the transient response under the right conditions [26].

The design and operation of constant surface temperature experiments are substantially different than those for constant heat flux. Instead of measuring the temperature that results from an imposed heat flux, the electrical power that is required to maintain a set temperature is measured. This typically requires a control system to maintain the temperature constant spatially and/or temporally, depending on whether steady or unsteady measurements are desired. Time-averaged measurements of spatial distributions of convection heat transfer are usually made with segmented plates and heaters. If the plates are made of a thick, high-conductivity material, the surface will be nearly isothermal, even if the heat transfer coefficient has large spatial variations. Some thermal isolation is required between segments, often provided by insulation strips. A good set of PID automatic temperature controllers can greatly speed the stabilization process. When combined with thin-film heaters deposited on insulation, time-resolved measurements are even possible.

67.6 Calibration and Problems to Avoid

Calibration of heat flux is a complicated issue because some heat flux sensors respond to different modes and conditions of heat flux differently. For example, a sensor calibrated by radiation can have a substantially different response to the same amount of heat flux in convection.

For the low heat fluxes seen in building applications, the guarded hot-plate method for calibration has been well established [27]. The sensor to be calibrated is placed between a heated plate and a cooled plate.

The electrical power into the heated plate is measured at steady state to provide the known heat flux. The National Institute of Standards and Technology (NIST) maintains calibration devices for this range of conduction heat transfer through one-dimensional planar sensors. Calibrated insulation samples are readily available to check other guarded hot-plate calibrators against their standards.

Most insert gages are calibrated with radiation. There are two basic approaches. The first is to use a system that provides an absolute comparison with a primary standard. The second uses another heat flux sensor as a secondary standard for the comparison. There are three methods for absolute radiation calibration of heat flux sensors as illustrated in the facilities at NIST [28]. They are all designed for water-cooled sensors that are kept at room temperature for the calibration. One uses an electrical substitution radiometer as the primary standard in a narrow-angle facility. The second is a spherical cavity that uses the temperature of the wall as the standard along with the Stefan–Boltzmann law (Equation 67.5) for the blackbody radiation to the sensor. The heat flux sensor should be located at the bottom of the sphere or the sphere should be in a vacuum to minimize the effects of convection. The third method is insertion into a graphite tube blackbody cavity with a pyrometer on the opposite side of the center partition for the primary standard. All of these methods calibrate the sensors based on the incident radiation, q''_{inc}.

$$S_{inc} = E/q''_{inc} \tag{67.15}$$

The voltage output is E and the sensitivity is S in units of voltage per heat flux. The emissivity of the surface ε is required to calculate the radiation that is actually absorbed by the sensor.

$$q''_{abs} = \varepsilon q''_{inc} \tag{67.16}$$

The gage surface is assumed to be gray with the absorptivity equal to the emissivity, ε, and both values constant with angle and wavelength for the conditions of use. Here it is assumed that the radiation emitted from the gage is negligible compared to the incident because the gage is water cooled and kept at a temperature near 300 K.

The manufacturers use secondary standards to calibrate heat flux gages by direct comparison either simultaneously or sequentially with matching radiation fluxes. The same can be done for convection [29]. The resulting sensitivity corresponds to the absorbed radiation. Consequently, the usual incident radiation calibration must be adjusted to correspond to the convection sensitivity by the emissivity of the original surface coating:

$$S_{conv} = S_{inc}/\varepsilon \tag{67.17}$$

As with many other measurements, the major problem with heat flux measurement is the error caused in the heat flux by the disruption of the sensor itself. For sensors based on the spatial temperature gradient methods, a larger signal implies a larger temperature difference and a larger temperature disruption. For the second type of sensors based on the transient temperature change, the surface temperature is changing while the measurement occurs. The larger the temperature change, the easier the determination of the sensor heat flux, but the larger the error from the sensor temperature disruption. In convection, the effect of the surface temperature disruption directly affects the apparent heat flux measured by the gage [29]. Therefore, it is imperative in convection measurements to keep the thermal disruption of the sensor to a minimum.

There are many situations where the heat flux is desired on surfaces with elevated temperatures. As previously discussed, there are several gages that are capable of operating at up to 1000 °C. Because of changes in thermal properties of the gage materials with temperature, calibration at the actual gage temperature is required. One method is to place the sensor on a heated surface and a water-cooled gage

on a cold surface and measure the radiation from the hot surface to the cold surface in enclosure [30]. If the side walls are insulated and adiabatic, the heat leaving the hot surface will match that received by the cold plate. This system allows sensors to be characterized up to 900 °C.

67.7 Summary

There are a large number of heat flux sensors available for use in a wide variety of applications. Those commercially available have been listed in Table 67.1, and the information for contacting the manufacturers is given in Table 67.2. The sensors that have not yet been commercialized are sometimes still available from the labs where they were developed.

The differential temperature devices provide a direct readout of the heat flux over the surface of the sensor. With the proper choice of sensor for the application and care in measurement method, the results are easily interpreted and used. Alternatively, the transient temperature methods can provide more surface details, but the output is a surface temperature history that must be analyzed to obtain the corresponding heat flux. Although certain components of these systems are off-the-shelf, more work is required of the user to interpret the results. Issues of calibration and errors have been addressed briefly here. More detail on all aspects of heat flux measurement can be obtained from the manufacturers and the references listed.

References

1. F. Incropera, D. DeWitt, T. Bergman, and A. Lavine, *Fundamentals of Heat and Mass Transfer*, 7th edn., New York: John Wiley & Sons, 2011.
2. T. E. Diller, Advances in heat flux measurement, in J. P. Hartnett et al. (eds.), *Advances in Heat Transfer*, Vol. 23, Boston, MA: Academic Press, 1993, pp. 279–368.
3. N. R. Keltner, Heat flux measurements: Theory and applications, Chapter 8, in K. Azar (ed.), *Thermal Measurements in Electronics Cooling*, Boca Raton, FL: CRC Press, 1997, pp. 273–320.
4. ASTM E2684-09, Standard test method for measuring heat flux using surface-mounted one-dimensional flat gages, *Ann. Book ASTM Standards*, 15.03, 2009.
5. D. O. Hubble and T. E. Diller, Development and evaluation of the time-resolved heat and temperature array, *Journal of Thermal Science and Engineering Applications*, 2, 2010, 031003, 9.
6. A. R. Gifford, D. O. Hubble, C. A. Pullins, S. T. Huxtable, and T. E. Diller, A durable heat flux sensor for extreme temperature and heat flux environments, *AIAA Journal of Thermophysics and Heat Transfer*, 24, 2010, 69–76.
7. R. L. Hauser, Construction and performance of in situ heat flux transducers, in E. Bales et al. (eds.), *Building Applications of Heat Flux Transducers*, ASTM STP 885, 1985, pp. 172–183.
8. F. Van der Graaf, Heat flux sensors, in W. Gopel et al. (eds.), *Sensors*, Vol. 4, New York: VCH, 1989, pp. 295–322.
9. D. O. Hubble and T. E. Diller, A hybrid method for measuring heat flux, *ASME Journal of Heat Transfer*, 132, 2010, 031602, 8.
10. ASTM E2683-09, Standard test method for measuring heat flux using flush-mounted insert temperature-gradient gages, *Annual Book ASTM Standards*, 15.03, 2009.
11. C. T. Kidd and C. G. Nelson, How the Schmidt-Boelter gage really works, *Proc. 41st Int. Instrum. Symp.*, Research Triangle Park, NC: ISA, 1995, pp. 347–368.
12. D. G. Holmberg and T. E. Diller, High-frequency heat flux sensor calibration and modeling, *ASME J. Fluids Eng.*, 117, 1995, 659–664.
13. ASTM E511-07, Measurement of heat flux using a copper-constantan circular foil heat-flux gage, *Annual Book ASTM Standards*, 15.03, 2009.
14. C. H. Kuo and A. K. Kulkarni, Analysis of heat flux measurement by circular foil gages in a mixed convection/radiation environment, *ASME Journal of Heat Transfer*, 113, 1991, 1037–1040.

15. W. J. Cook and E. M. Felderman, Reduction of data from thin film heat-transfer gages: A concise numerical technique, *AIAA Journal*, 4, 1966, 561–562.
16. M. L. G. Oldfield, Impulse response processing of transient heat transfer gauge signals, *Journal of Turbomachinery*, 130, 2008, 021023.
17. S. M. Guo, C. C. Lai, T. V. Jones, M. L. G. Oldfield, G. D. Lock, and A. J. Rawlinson, The application of thin-film technology to measure turbine-vane heat transfer and effectiveness in a film-cooled, engine-simulated environment, *International Journal of Heat Fluid Flow*, 19, 1998, 594–600.
18. C. T. Kidd, C. G. Nelson, and W. T. Scott, Extraneous thermoelectric EMF effects resulting from the press-fit installation of coaxial thermocouples in metal models, *Proc. 40th Int. Instrum. Symp.*, Research Triangle Park, NC: ISA, 1994, pp. 317–335.
19. ASTM Standard E598-77, Standard method for measuring extreme heat-transfer rates from high-energy environments using a transient null-point calorimeter, *Annual Book of ASTM Standards*, 15.03, 1988, pp. 381–387.
20. G. Wagner, M. Kotulla, P. Ott, B. Wegand, and J. von Wolfersdorf, The transient liquid crystal technique: Influence of surface curvature and finite wall thickness, *Journal of Turbomachinery*, 127, 2005, 175–182.
21. J.-W. Ahn, R. Maingi, D. Mastrovito, and A. L. Roquemore, High speed infrared camera diagnostic for heat flux measurement in NSTX, *Review of Scientific Instruments*, 81, 2010, 023501.
22. A. H. Khalid and K. Kontis, Thermographic phosphors for high temperature measurements: Principles, current state of the art and recent applications, *Sensors*, 8, 2008, 5673–5744.
23. I. Kurits and M. J. Lewis, Global temperature-sensitive paint system for heat transfer measurements in long-duration hypersonic flows, *Journal Thermophysics Heat Transfer*, 23, 2009, 256–266.
24. ASTM Standard E457-72, Standard method for measuring heat-transfer rate using a thermal capacitance (slug) calorimeter, *Annual Book of ASTM Standards*, 15.03, 1988, 299–303.
25. N. R. Keltner, J. V. Beck, and J. T. Nakos, Using directional flame thermometers for measuring thermal exposure, *Journal of ASTM International*, 7, 2010, 102280.
26. T. J. Praisner and C. R. Smith, The dynamics of the horseshoe vortex and associated endwall heat transfer–Part 1: Temporal behavior, *Journal of Turbomachinery*, 128, 2006, 747–754, 2006.
27. M. Bomberg, A workshop on measurement errors and methods of calibration of a heat flow meter apparatus, *Journal of Thermal Insulation and Building Environments*, 18, 1994, 100–114.
28. A. V. Murthy, B. K. Tsai, and R. D. Saunders, Radiative calibration of heat-flux sensors at NIST: Facilities and techniques, *Journal of Research of National Institute of Standards and Technology*, 105, 2000, 293–305.
29. A. Gifford, A. Hoffie, T. Diller, and S. Huxtable, Convection calibration of Schmidt-Boelter heat flux gages in stagnation and shear air flow, *ASME Journal of Heat Transfer*, 132, 2010, 031601, 9.
30. C. A. Pullins and T. E. Diller, In situ high temperature heat flux sensor calibration, *International Journal of Heat and Mass Transfer*, 53, 2010, 3429–3438.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

68

Resistive Thermometers

68.1 Introduction **68**-1
68.2 How an RTD Works? **68**-1
R versus *T* Relationship of Various Metals • Resistance versus Temperature Relationship of a PRT
68.3 RTD Construction **68**-4
Sensing Element • Protective Housing • Lead Wires
68.4 Classifications of RTDs **68**-7
68.5 Applications **68**-7
Selection • Common Sources of Error • Interchangeability • Insulation Resistance • Stability • Repeatability • Hysteresis • Stem Conduction • Calibration/Interpolation • Lead Wire • Self-Heating • Time Response • Thermal EMF
68.6 Calibration and Maintenance **68**-19
68.7 Conclusion **68**-19
Defining Terms **68**-20
References **68**-21

Jim Burns
Burns Engineering, Inc.

68.1 Introduction

Temperature, the most frequently measured process variable on the planet, can be measured in various ways. Many temperature measurement devices incorporate a resistive material that when exposed to an environment, the resistance will change in response to the temperature in a predictable manner. This style of sensor is referred to as a resistance temperature detector (RTD). The resistive materials used are nickel, copper, and most commonly platinum. RTDs designed with platinum provide the most stable resistance versus temperature relationship and therefore can achieve accuracies capable for use as a primary reference standard.

The resistive material is "packaged" in a wide variety of approaches to achieve the performance required and to support the installation constraints of the measurement point in the process of concern. There are many factors that should be considered in the design, selection, installation, integration, and monitoring of an RTD to ensure that the actual measurement meets the needs of the process.

This chapter discusses the principle of operation of an RTD, the general construction methods, common applications, how to select an RTD, interfacing an RTD with measurement instruments, and finally an overview of the types of errors that must be taken into consideration when using an RTD.

68.2 How an RTD Works?

Common RTD sensing elements constructed of platinum copper or nickel have a unique, and repeatable and predictable resistance versus temperature relationship (R versus T) and operating temperature range. The R versus T relationship is defined as the amount of resistance change of the

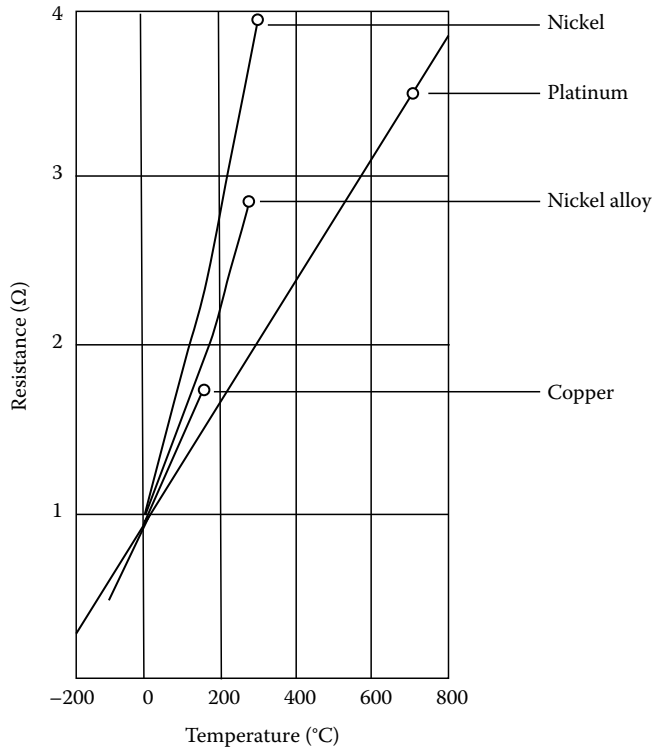


FIGURE 68.1 *R* versus *T* relationship for various metals.

sensor per degree of temperature change. Figure 68.1 shows the resistance relationship for the most common metals used for RTDs and the temperature ranges.

68.2.1 *R* versus *T* Relationship of Various Metals

Platinum is a noble metal and has the most stable resistance to temperature relationship over the largest temperature range. Nickel elements have a limited temperature range because the amount of change in resistance per degree of change in temperature becomes very nonlinear at temperatures over 572 °F (300 °C). Copper has a very linear resistance to temperature relationship; however, copper oxidizes at moderate temperatures and cannot be used over 302 °F (150 °C).

Platinum is the best metal for RTDs because it follows a very linear resistance to temperature relationship and it follows the *R* versus *T* relationship in a highly repeatable manner over a wide temperature range. The unique properties of platinum make it the material of choice for temperature standards over the range of –272.5 °C to 961.78 °C.

The basic differentiator between metals used as resistive elements is the linear approximation of the *R* versus *T* relationship between 0 °C and 100 °C and is referred to as alpha, α. Equation 68.1 defines α; its units are Ω/Ω/°C.

$$\alpha = \frac{R_{100} - R_0}{100 R_0} \tag{68.1}$$

where

R_0 is the resistance of the sensor at 0 °C

R_{100} is the resistance of the sensor at 100 °C

Pure platinum has an alpha of 0.003925 Ω/Ω/°C and is used in the construction of laboratory-grade RTDs. Conversely, two widely recognized standards for industrial RTDs IEC60751 [1] and ASTM E-1137 [2] specify an alpha of 0.00385 Ω/Ω/°C. Before these standards were widely adopted, several different alpha values were used. It is still possible to find older probes that are made with platinum that have alpha values of 0.003916 Ω/Ω/°C and 0.003902 Ω/Ω/°C.

These different alpha values for platinum are achieved by doping, basically carefully introducing impurities into the platinum. The impurities introduced during doping become embedded in the lattice structure of the platinum and result in a different *R* versus *T* curve and hence alpha value. To characterize the *R* versus *T* relationship of any RTD over a temperature range that represents the planned range of use, calibration must be performed at temperatures other than 0 °C and 100 °C. Two common calibration methods are the fixed-point method and the comparison method.

Fixed-point calibration, used for the highest accuracy calibrations, uses the triple point, freezing point, or melting point of pure substances such as water, zinc, tin, and argon to generate a known and repeatable temperature. These cells allow the user to reproduce actual conditions of the ITS-90 temperature scale. Fixed-point calibrations provide extremely accurate calibrations (within ±0.001 °C). A common fixed-point calibration method for industrial-grade probes is the ice bath. The equipment is inexpensive, easy to use, and can accommodate several sensors at once. The ice point is designated as a secondary standard because its accuracy is ±0.005 °C (±0.009 °F), compared to ±0.001 °C (±0.0018 °F) for primary fixed points.

In *comparison calibrations,* commonly used with secondary standard platinum resistance thermometers (SPRTs) and industrial RTDs, the thermometers being calibrated are compared to calibrated thermometers by means of a bath whose temperature is uniformly stable. Unlike fixed-point calibrations, comparisons can be made at any temperature between –100 °C and 500 °C (–148 °F to 932 °F). This method might be more cost-effective since several sensors can be calibrated simultaneously with automated equipment. These electrically heated and well-stirred baths use silicone oils and molten salts as the medium for the various calibration temperatures.

68.2.2 Resistance versus Temperature Relationship of a PRT

After determining the platinum resistance thermometer's (PRT) resistance at various temperature points, the *R* versus *T* curve can be defined. Once this has been done, any temperature can be determined from the measured resistance. These equations are multivariable polynomials, and although complex, the various forms of the equations are typically incorporated into transmitters, handheld readouts, and temperature-related electronic devices to solve this challenge.

The two most common curve-fitting techniques are the ITS-90 [3] and Callendar–Van Dusen equations. On January 1, 1990, the International Temperature Scale of 1990 (ITS-90) became the official international temperature scale. Two reference functions are used to define the temperature coefficient for an ideal SPRT: one for temperatures below 0 °C and the other for temperatures above 0 °C. When a PRT is calibrated on the ITS-90, the coefficients determined in the calibration are used to describe a deviation function that represents the difference between the resistance of the standard PRT and the reference function at all temperatures within the range. Using the calibration coefficients and the deviation functions, the SPRT can be used to determine any temperature from its measured resistance. Because ITS-90 equations are complex, computer software is necessary for accurate calculations.

Not all PRT users need to follow the complex equations and computer programs associated with ITS-90. As a rule of thumb, if the minimum required uncertainty of measurement is less than 0.1 °C, one probably will want to use ITS-90. For uncertainty of measurements greater than 0.1 °C, the effect of the change in scales is relatively small. Callendar–Van Dusen equations are interpolation equations that describe the temperature versus resistance relationship of industrial PRTs. These second- and fourth-order equations can be programmed easily into many electronic controllers.

The equation for an RTD with a 100 Ω R_0:

For the temperature range of 0 °C–850 °C (32 °F–1562 °F) is

$$R(t) = R_0(1 + At + Bt^2) \tag{68.2}$$

For the temperature range –200 °C to 0 °C (–392 °F to 32 °F):

$$R(t) = R_0[1 + At + Bt^2 + C(t - 100)t^3] \tag{68.3}$$

where

$R(t)$ is the resistance of the PRT at temperature t
t is the temperature in °C
R_0 is the nominal resistance of the PRT at 0 °C
A, B, and C are constants based on calibration coefficients: α, δ, β
(These coefficients and constants are generally provided by the calibration laboratory)
Expressions for the constants A, B, and C are

$$\boxed{A = \alpha + \frac{\alpha \cdot \delta}{100}} \quad \boxed{C = -\frac{\alpha \cdot \beta}{100^4}} \quad \boxed{B = -\frac{\alpha \cdot \delta}{100^2}} \tag{68.4}$$

Expressions for the constants α, δ, and β are

$$\boxed{\alpha = \frac{R_{100} - R_0}{100 \cdot R_0}} \quad \boxed{\delta = \frac{t_h - \frac{R_{th} - R_0}{R_0 \cdot \alpha}}{\left(\frac{t_h}{100} - 1\right)\left(\frac{t_h}{100}\right)}}$$

$$\boxed{\beta = \frac{t_l - \left(\frac{R_{tl} - R_0}{R_0 \cdot \alpha} + \delta\left(\frac{t_l}{100} - 1\right)\left(\frac{t_l}{100}\right)\right)}{\left(\frac{t_l}{100} - 1\right)\left(\frac{t_l}{100}\right)^3}} \tag{68.5}$$

where

t_h is the high temperature in the calibration range (e.g., 260 °C)
t_l is the low temperature in the calibration range (e.g., –196 °C)
R_{th} is the resistance of the sensor at the high temperature
R_{tl} is the resistance of the sensor at the low temperature

68.3 RTD Construction

There are many different types of RTDs for different applications. Generally, an RTD consists of three major components:

- Sensing element
- Protective housing
- Lead wires

Figure 68.2 shows the construction of a typical industrial RTD; the tip is cut away to show the position of the sensing element.

The selection of these three components is critical to the performance of the probe, and careful selection is required to ensure that the probe is suitable for the application.

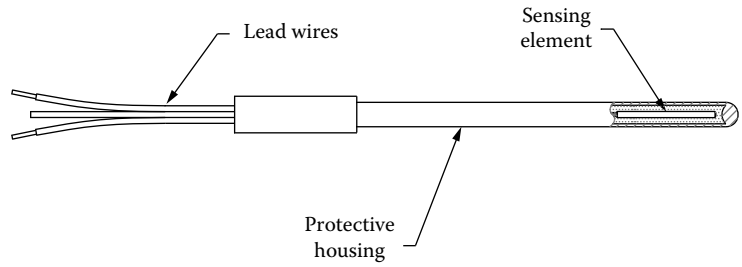


FIGURE 68.2 Construction of a typical industrial RTD.

68.3.1 Sensing Element

All RTD elements are in essence a length of wire (whether coiled, wound, or deposited through vapor deposition as in the case of thin film) that becomes more resistive at higher temperature. But the resistance of these wires can also be affected by physical deformation or strain. The two main causes of strain in RTDs come from thermal expansion of the wire relative to the electrically insulating support structure and from deformation of the wire due to shock or vibration. All element designs attempt to minimize the effect of strain on the wire, but there is a trade-off between strain from thermal expansion versus shock and vibration. This can be understood simplistically as the more firmly the wire is attached to the electrical insulator, the less likely it is able to move under shock and vibration conditions, but it is more likely to be stretched with the thermal expansion and contraction of the insulator.

There are three different categories of elements; each has their own particular characteristics:

Strain Free Elements consist of platinum wire loosely coiled over a support structure so the element is free to expand and contract with temperature, but it is very susceptible to shock and vibration as the loops of platinum can sway back and forth causing deformation (see Figure 68.3).

Wire coil wound elements (or coil-in-the-hole elements) are coils of platinum placed in ceramic tubes that are filled with ceramic powder to support the coils, as in Figure 68.4. The powder supports the coil against some shock and vibration, and allows some movement for thermal expansion and contraction.

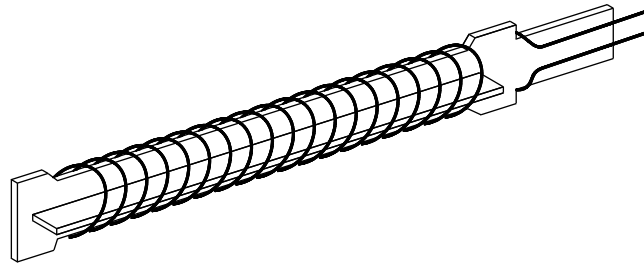


FIGURE 68.3 Strain-free constructions.



FIGURE 68.4 Wire coil wound.

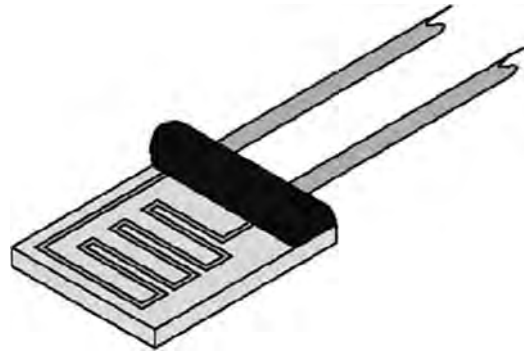


FIGURE 68.5 Thin-film elements.

Thin Film elements are small squares of ceramic that have a thin film of platinum deposited on them, Figure 68.5. Some platinum is removed to form a zigzag patterned electrical path. The thin film of platinum does not move with shock and vibration but due to the small mismatch in coefficient of expansion between the platinum pattern and ceramic, thin films become less accurate over large temperature ranges. This limits their useful range.

A fourth type called externally wire wound is still used, but thin films and wire coil wound elements are generally preferred, and there are few instances where this older design is better than modern thin films or wire coil wound elements.

68.3.2 Protective Housing

The protective housing has to ensure there is adequate thermal connection between the element and the measurement environment of interest, and at the same time protect the element from its soundings. Stainless steel probes are among the most common; a stainless steel tube closed at one end houses the element with the electrical connections coming out the other end. Many types of mounting fittings are available for various uses. Other types of probes use high-temperature plastics such as PTFE to encase the element.

The working temperature range of the probe is one of the criteria that will influence or limit the choices regarding the sheath material. For example, if stainless steel is used at elevated temperatures, the oxides formed can cause contamination of the platinum sensing element adversely affecting the accuracy of the probe.

How the element and lead wires are supported inside the protective housing is also critical. The materials used have to be capable of the temperature range, provide a good thermal path to the element, and isolate the element from the ambient vibrations without putting excessive strain on the element. Materials used include ceramic powder, heat sink compounds, and silicone RTVs.

68.3.3 Lead Wires

When selecting the lead wire materials and the number of lead wires, consideration has to be given to the environment in which the probe will reside as well as the instrumentation to which it is going to interface. The element is basically a length of wire that changes in resistance with temperature. In order to measure the resistance of the element, it must be wired to the measurement equipment in such a way that the resistance of the leads does not adversely affect the measurement. This is done in three ways, using two, three, or four wire connections. See the section on *Common Sources of Error* to understand the considerations that must be taken into account with regard to specifying the number of lead wires.

68.4 Classifications of RTDs

Though many types of element wire can be used to make RTDs, platinum has become the standard in both laboratory and industrial settings. Platinum RTDs are often called platinum resistance thermometers (PRTs). RTDs utilizing other metals for the sensing element are only used in special applications or for measurements outside the PRT temperature range.

The highest accuracy of all PRTs is the SPRTs. This accuracy is achieved at the expense of durability and cost. The SPRT elements are wound from reference-grade platinum wire. Internal lead wires are usually made from platinum while internal supports are made from quartz or fuse silica. The sheaths are usually made from quartz or sometimes Inconel, depending on temperature range. Larger diameter platinum wire is used, which drives up the cost and results in a lower resistance for the probe (typically 25.5 Ω). SPRTs have a wide temperature range (–200 °C to 1,000 °C) and approximately accurate to ±0.001 °C over the temperature range. SPRTs are only appropriate for laboratory use.

Another classification of laboratory PRTs is secondary SPRTs. They are constructed like the SPRT, but the materials are more cost-effective. SPRTs commonly use reference-grade, high–purity, smaller diameter platinum wire, metal sheaths, and ceramic-type insulators. Internal lead wires are usually a nickel-based alloy. Secondary SPRTs are limited in temperature range (–200 °C to 500 °C) and are approximately accurate to ±0.03 °C over the temperature range.

Industrial-grade PRTs are designed to withstand industrial environments. They can be almost as durable as a thermocouple. Depending on the application, industrial PRTs can use thin film elements or coil wound elements. The internal lead wires can range from PTFE insulated stranded nickel-plated copper to silver wire, depending on the sensor size and application. Sheath material is typically stainless steel; higher-temperature applications may demand Inconel. Other materials are used for specialized applications.

68.5 Applications

Sensor assemblies can be categorized into two groups by how they are installed or interface with the process: immersion or surface mounted.

Immersion sensors take the form of an SS tube and some type of process connection fitting (Figures 68.6 and 68.7). They are installed into the process with sufficient immersion length to ensure good contact with the process medium and reduce external influences. A variation of this style includes a separate thermowell that provides additional protection for the sensor. These



FIGURE 68.6 Direct immersion styles.

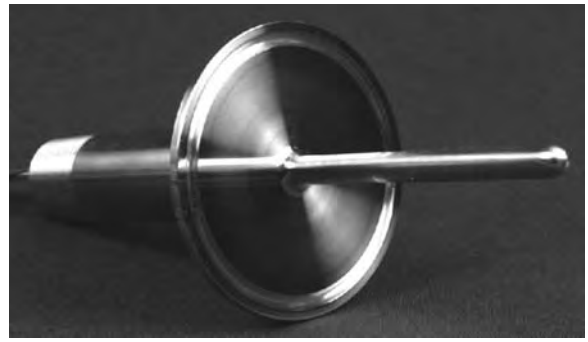


FIGURE 68.7 Direct immersions with sanitary process connection.

styles are used to measure fluid or gas temperatures in pipes and tanks. Most sensors have the sensing element located at the tip of the stainless steel tube. An averaging-style RTD, however, can measure an average temperature of air in a large duct. This style of immersion RTD has the sensing element distributed along the entire probe length and provides an average temperature. Lengths range from 3 to 60 ft.

Surface-mounted sensors (Figure 68.8) are used when immersion into a process fluid is not possible due to configuration of the piping or tank, or the fluid properties may not allow an immersion-style sensor. Configurations range from tiny cylinders to large blocks that are mounted by clamps, adhesives, or bolted into place (Figure 68.9). Most require the addition of insulation to isolate them from cooling or heating effects of the ambient conditions to ensure accuracy.

Other applications may require special water proofing or pressure seals. A heavy-duty underwater temperature sensor is designed for complete submersion under rivers, cooling ponds, or sewers. Steam autoclaves require a sensor that is sealed from intrusion by steam during the vacuum cycle process.

Immersion sensors generally have the best measurement accuracy because they are in direct contact with the process fluid. Surface-mounted sensors are measuring the pipe surface as a close approximation of the internal process fluid.

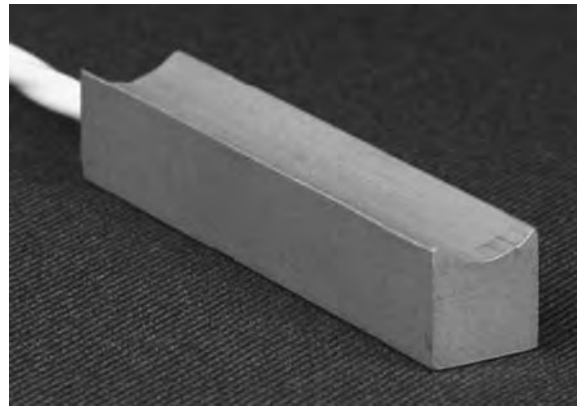


FIGURE 68.8 Surface mount for a pipe.



FIGURE 68.9 Surface mount for flat surfaces.

68.5.1 Selection

Selection of a PRT can be achieved by considering each of five criteria:

- Placement
- Protection
- Performance
- Price
- Service life

Placement refers to how the sensor will be mounted in or on the process. Several factors influence which style is selected to achieve the desired measurement accuracy and longest service life. An immersion-style sensor usually provides the best accuracy but does require cutting a hole in the pipe or tank to allow for installation. A surface-mount sensor is the easiest to install and of lowest cost; however, some accuracy is sacrificed. Placement considerations involve determining where in the piping or tank an entry can be made to accommodate an RTD and provide enough immersion length for an accurate measurement. The ideal immersion length to avoid external thermal influences is equal to 10 sensor or thermowell diameters plus the sensitive length (the element length). For a direct immersion, 1/4 ft diameter RTD with a 1.0 in. element length, the minimum immersion would be 3.5 ft. A 1/4 ft diameter sensor with a thermowell should have at least 4.5 ft of immersion in the process. For pipes less than 4 ft and other situations where sufficient immersion cannot be obtained, there are other solutions. Installing an elbow allows the RTD to be aligned parallel to the flow and extend down the pipe. Other choices may be to use a surface mount RTD, or a modified immersion sensor that mounts flush to the inside of the process vessel.

Protection refers to the items that are used to shield the sensor from the ambient conditions and from the process environment. External to the process, a connection head is commonly used to provide a transition from the RTD leads to the facility wiring. In addition, it can provide protection from explosive atmospheres and keep water and other contaminants away from the sensor. Any of these ambient conditions can, over time, cause sensor failure.

To maintain protection from the process, a thermowell (Figure 68.10) can be used to protect the sensor from corrosive, erosive, high-velocity flow, high pressure, and vibration. Another advantage of using a thermowell is that RTDs that need periodic calibration can be easily removed from the protective thermowell without draining the process.



FIGURE 68.10 RTD and thermowell assembly.

Performance includes factors such as probe accuracy, installation errors, long-term drift, and erosion or corrosion considerations. Accuracy of the sensor and accuracy of the measurement are most times quite different. Sensor accuracy is determined mostly by the manufacturing interchangeability and the style of sensing element. The wire wound style has the widest temperature range and lowest drift, and a thin film element may have the same interchangeability but larger drift characteristics. Temperature range is often restricted by manufacturers to minimize the effect of drift on the sensor accuracy. Measurement accuracy includes the sensor accuracy and the installation affects. In addition, time response can be a large factor in the measurement accuracy. A sensor that does not keep pace with the process temperature change may be several degrees off from the actual temperature during thermal transients.

Price, or from the user's perspective, the cost is always a consideration. For any measurement point, the costs for maintenance, calibration, acquisition, and accuracy must be determined. The initial cost of the RTD can be insignificant if a poor-quality inaccurate device is selected. A poor-quality RTD will waste energy, require frequent replacement, require more frequent calibration, and can reduce product quality. A more expensive high-quality RTD will alleviate all of these problems and related expenses, resulting in the lowest life cycle cost for the measurement point.

Service life of an RTD varies dramatically based on how and where the sensor is used. Two main factors that affect the life span are temperature cycling and vibration. All industrial-grade RTDs will drift over time and require periodic calibration to determine if they are still within tolerance. Vibration and wide temperature cycles can accelerate the drift rate possibly leading to premature failure. Selecting the right RTD and following the selection criteria will ensure accuracy, long service life, and trouble-free temperature measurement.

68.5.2 Common Sources of Error

To fully understand the accuracy of a temperature measurement, it is imperative that all the sources of errors are understood and where possible quantified. Accuracy is not a statement of the correctness of a temperature sensor independent of its system. It is the total accuracy of the sensor system, and process for which one needs to measure temperature. This accuracy value is a summation of the sensors'

interchangeability value, as well as all the errors sources associated with the system, the measurement location, and the installation of the RTD. The method that is used for calculating the accuracy of the system is by calculating the root-sum-square (RSS) of all the sources of error and variation of the probe, and the errors related to the system. Generally, the sensors' interchangeability value is the largest component of the accuracy analysis and can be reduced to near zero by matching the sensor with a transmitter. As the temperature range of use expands, the errors associated with stability, repeatability, and stem conduction will begin to play a larger role in the system accuracy at higher temperatures.

With all measurement devices, there are common sources of error that have to be accounted for. In PRTs, these sources of error are interchangeability, insulation resistance, stability, repeatability, stem conduction, hysteresis, calibration and interpolation, lead wire resistance, self-heating, rime response, and thermal EMF.

68.5.3 Interchangeability

Interchangeability refers to the "closeness of agreement" in the *R* versus *T* relationship of a PRT to a predefined nominal *R* versus *T* relationship. Since it is not possible to manufacture a PRT that exactly replicates a predefined nominal *R* versus *T* relationship, a tolerance band about the relationship must be established, and this is what is called the "interchangeability" of the PRT.

The two most widely used industrial PRT standards in the United States are ASTM E1137 and IEC 60751. Fortunately, both of these standards have the same nominal *R* versus *T* relationship, but they do not have the same interchangeability tolerance. These standards use the term "tolerance" to describe the permissible variation in the *R* versus *T* relationship, define tolerance as a function of temperature in °C, and offer multiple tolerance levels—referring to them by either a tolerance "class" or tolerance "grade." Table 68.1 shows the interchangeability tolerances for these two industry standards while Figure 68.11 shows a comparison of the interchangeability levels graphically.

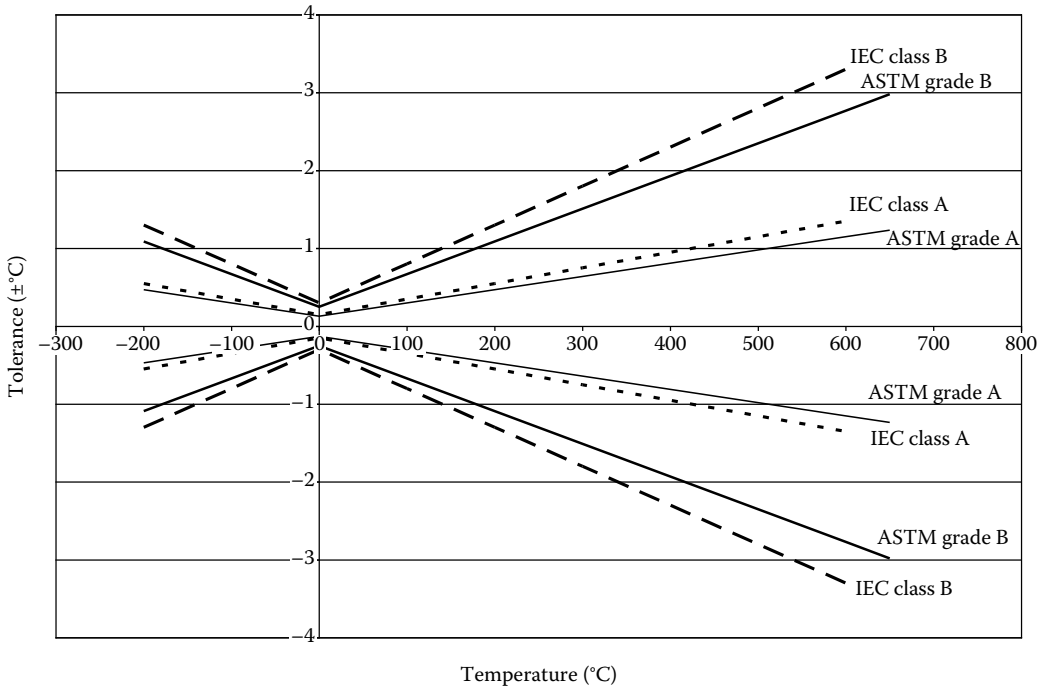


FIGURE 68.11 Standard PRT interchangeability.

TABLE 68.1 Standard PRT Interchangeability Equations

Standard	Tolerance Nomenclature	Defining Equation[1]
ASTM E1137	Grade A	± [0.13 + 0.0017 | *t* |]
ASTM E1137	Grade B	± [0.25 + 0.0042 | *t* |]
IEC 60751	Class A	± [0.15 + 0.002 | *t* |]
IEC 60751	Class B	± [0.3 + 0.005 | *t* |]

Note: \| *t* \| = the value of temperature in °C without regard to sign.

It is worth noting that many times manufacturers will define interchangeability in terms of a percent, such as 0.06%. In these cases, the percent refers to a percent of resistance valid only at 0 °C. For example, 0.06% interchangeability is ±.06 Ω (±.15 °C) for a 100 Ω PRT. This is not enough information to know what the tolerance is at other temperatures. The user should be provided with either a statement that the tolerance meets one of the industry standard classes or grades, or a table showing the tolerances, or an equation similar to the ones shown in Table 68.1 that define the tolerance over the entire rated temperature range of the PRT.

68.5.4 Insulation Resistance

Insulation resistance (IR) refers to the electrical resistance between the sensing circuit and the metallic sheath of a PRT. It is important for the sensing element circuit to be insulated from the sheath because electrical leakage can cause an error when measuring the resistance of the sensing element. Any error in measuring the resistance will translate to an error in the indicated temperature.

Resistance is a parameter that cannot be measured directly. The typical method used to measure the resistance of an industrial PRT is to apply a constant current and measure the voltage drop to determine the resistance. If a portion of the applied current has the opportunity to "leak out" of the circuit, through a low IR, then a false resistance reading will be obtained for the sensing element.

ASTM E1137 and IEC 60751 specify a minimum IR value at ambient temperature (25 °C) as well as at elevated temperatures. Table 68.2 shows the minimum IR values required by these standards and the estimated temperature error values calculated for a 100 Ω PRT. It is easy to see by examining this table that the minimum IR requirement decreases as the temperature increases. This represents the typical behavior of an industrial PRT and is a reason why the ambient temperature IR requirements are generally much higher than necessary for accurate operation at ambient temperature.

TABLE 68.2 PRT Insulation Resistance Requirements

Standard	Rated Temperature (°C)	Minimum IR (MΩ)	Test Voltage (VDC)	Estimated Error for 100 Ω PRT (°C)
ASTM E1137	25	100	10–50	0.0003
ASTM E1137	300	10	10–50	0.013
ASTM E1137	650	2	10–50	0.17
IEC 60751	25	100	10–100	0.0003
IEC 60751	100–300	10	10	0.013
IEC 60751	301–500	2	10	0.12
IEC 60751	501–850	0.5	10	1.0

68.5.5 Stability

Stability refers to the ability of a PRT to maintain its *R* versus *T* relationship over time as a result of thermal exposure. Both ASTM E1137 and IEC 60751 address resistance change due to thermal exposure; however, the criteria are somewhat different. ASTM E1137 essentially states that the PRT must meet the resistance tolerance for a Grade A or Grade B sensor as applicable, after 4 weeks (672 h) of exposure to the maximum rated temperature. IEC 60751 states that the PRT can change at 0 °C by the equivalent of 0.15 °C for a Class A sensor and 0.30 °C for a Class B sensor after 250 h at maximum rated temperature. Many manufacturers state stability specifications more like the IEC 60751 standard, stating a maximum change at 0 °C as a result of exposure to maximum rated temperature for a given period of time. These specifications typically do not require the PRT to remain within the original *R* versus *T* tolerance after the exposure.

Typically three or four temperatures including the maximum rated temperature would be adequate to estimate stability. The graph in Figure 68.12 gives a realistic representation of the stability of a typical industrial PRT at five different exposure temperatures.

By examining Figure 68.12, a few generalizations can be made about the stability of this industrial PRT. These generalizations hold true for the behavior of most industrial PRTs; however, the shape of the curve will be unique based on the specific PRT design. If a better understanding is needed regarding a specific PRT, the manufacturer of that PRT should be consulted.

1. The resistance change due to exposure to room temperature and below can be considered negligible. Test results on several industrial PRTs have shown that the ice point resistance shift following 25,000+ h at 23 °C is less than 0.01 °C. This is a very small change in comparison to other potential error sources.
2. The magnitude of the change increases with increasing exposure temperature and is a maximum at maximum rated temperature. For example, the change after 1,000 h will be larger for an exposure temperature of 500 °C than an exposure temperature of 400 °C.
3. The magnitude of the change is not linear with temperature and cannot be extrapolated beyond the upper temperature limit. For example, the change caused by exposure to 550 °C cannot be reliably predicted based on the changes observed due to 500 °C or 400 °C exposure.
4. The rate of change over time is fairly linear and remains fairly constant. For example, the change after 2,000 h can be conservatively approximated as twice the amount of change as was present after 1,000 h for the same exposure temperature.

Given the earlier generalizations along with the information illustrated in Figure 68.12, a reasonably accurate estimation of stability can be made for the hypothetical example of a PRT at 350 °C for 1 year. The stability can be estimated by looking at the graph and estimating a change after 1,000 h for a 350 °C exposure. According to the graph, this would be approximately 0.1 °C. Multiplying this number

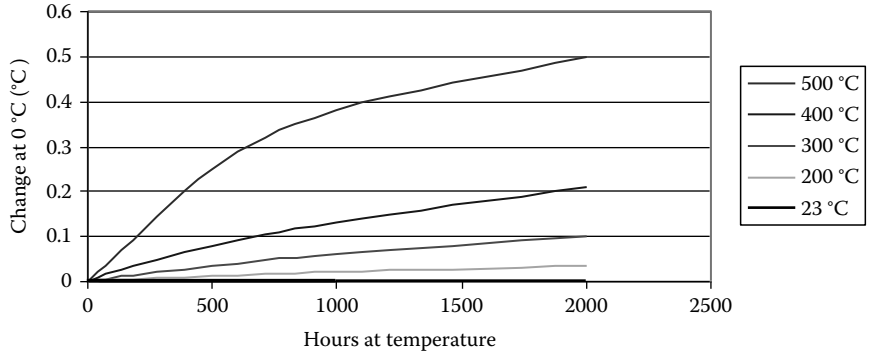


FIGURE 68.12 Example of industrial PRT stability.

by 9 to determine the cumulative effect over 9,000 h gives a result of 0.9 °C maximum change at 0 °C after 1 year of exposure to 350 °C. This is likely to be a more accurate estimate than extrapolating the IEC 60751 requirement of 0.15 °C for Class A (or 0.30 °C for Class B) per 250 h out to 9,000 h, which would result in an estimated value of 5.4 °C for Class A (or 10.8 °C for Class B).

68.5.6 Repeatability

Repeatability refers to the ability of a PRT to maintain its *R* versus *T* relationship when measured under the *same test condition* after experiencing thermal cycling throughout a specified temperature range. Industrial specification IEC 60751 refers to this as "Effect of temperature cycling" and states that the PRT can change at 0 °C by the equivalent of the class tolerance after 10 thermal cycles between the PRT's upper and lower temperature limits. The standard also contains a note to the effect that improved performance may be achieved for certain thermometers when used over narrower temperature ranges, but no specific requirements are given. ASTM E1137 does not contain any requirement for thermal cycling performance other than to state that the PRT must be capable of continuous operation over the specified temperature range. Neither of these industry standards provides substantial information about the performance of a PRT as a result of thermal cycling.

A practical and useful approach to understanding PRT repeatability would be a set of specifications that state the change due to a varying number of cycles over varying temperature ranges. Typically three or four temperature ranges including the maximum rated range and cycles up to 1,000 would be sufficient. The graph in Figure 68.13 gives a realistic representation of the repeatability of a typical industrial PRT that has a temperature rating of 0 °C–500 °C, over four different thermal cycle temperature ranges.

From Figure 68.13, several generalizations can be identified regarding repeatability. The generalizations hold true for the behavior of most industrial PRTs; however, the actual values, maximum temperature ranges, and so on will all be unique based on the specific PRT design. If a better understanding is needed regarding a specific PRT, the manufacturer of that PRT should be consulted.

1. A disproportionate amount of change can occur during the first few thermal cycles. Figure 68.13 starts at 10 thermal cycles, which align with a typical specification such as "0.1 °C maximum change after 10 cycles from 20 °C to 500 °C."
2. There will be some temperature range for which there is virtually no difference in repeatability between 10 cycles and 1,000 cycles. In the case of this example, that temperature range is 20 °C–200 °C.

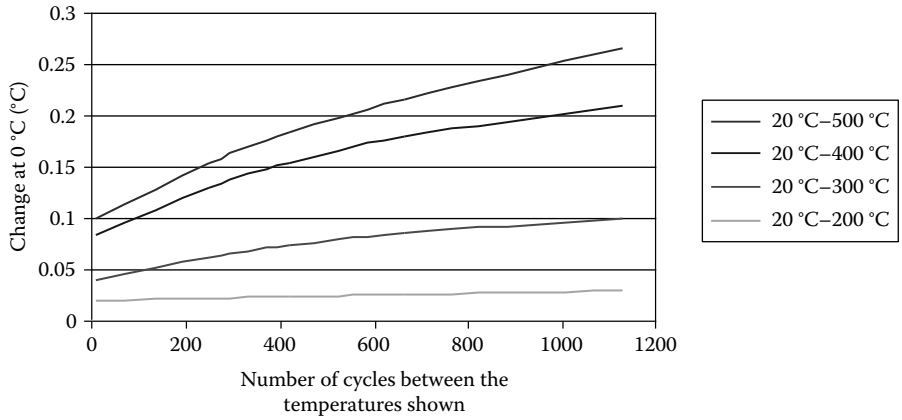


FIGURE 68.13 Example of industrial PRT repeatability.

3. The maximum nonrepeatability will occur due to cycling over the maximum rated temperature range of the PRT. Smaller ranges will result in smaller changes.
4. The change at intermediate temperature ranges and numbers of cycles can be estimated by interpolating between known values. For example, the repeatability after 500 cycles from 20 °C to 350 °C could be estimated as approximately 0.13 °C.
5. The change can be estimated as linear with the number of cycles for any given temperature range. For example, the change after 2,000 cycles can be conservatively estimated as two times the change after 1,000 cycles. This method should result in a conservative estimate for most high-quality PRTs.

Given the information contained in Figure 68.13, a reasonable and useful estimate could be made for the repeatability of the PRT under various conditions. An application example would be the use of a PRT in a process that must heat from room ambient to 350 °C for 1 h and then cool back to room ambient temperature. Assuming this process runs three times per day for an entire year, the error due to repeatability could be estimated from Figure 68.13 as the value shown at 1,095 cycles (3/day × 365 days/year) in the region between the 300 °C line and the 400 °C line. From Figure 68.13, an estimate of 0.16 °C could be made. Keep in mind that this error is separate from the error that would be created due to the 1,095 h of heat soak at 350 °C: that error is considered in the stability component of the overall error.

68.5.7 Hysteresis

In PRTs, thermal hysteresis results in a difference in resistance at a given temperature based on the thermal history to which the PRT was exposed. More specifically, the resistance of the PRT will be different when the temperature is approached from an increasing direction versus a decreasing direction, and the magnitude of the difference will depend on the magnitude of the temperature excursion and the design of the PRT.

Industrial specification IEC 60751 refers to this as "Effect of hysteresis" and states that the PRT resistance shall be measured at the middle temperature after exposure to the lower limit temperature and after exposure to the upper limit temperature. The acceptance criterion that is given is that the resistance at the midpoint cannot change by more than the tolerance value at the test temperature for the respective tolerance class tested. It does not state that both measurements must meet the resistance tolerance; therefore, it is possible that the hysteresis effect could result in the sensor producing an out-of-tolerance measurement. ASTM standard E1137 does not contain any requirement for thermal hysteresis, however; a test method for hysteresis is contained in ASTM standard E644, which would allow testing over ranges less than the maximum rated range of the PRT. While the requirement of the IEC 60751 standard is fairly well defined, it does not provide any information about hysteresis when the range is less than the maximum, and neither of these industrial PRT standards provides information about the hysteresis performance at temperatures other than the midpoint temperature where hysteresis is considered to be at its maximum.

A practical and useful approach to understanding a PRT's hysteresis would be if the hysteresis was specified as a percent of span to which the PRT was exposed.

Figure 68.14 shows a hypothetical representation of what could be considered a very well-behaved hysteresis loop for a typical industrial PRT. The hysteresis is a maximum at the midpoint temperature and a minimum at each end point. While there is still some variability expected in the measurements taken at the end points, this variability is caused by the PRT repeatability, not by hysteresis.

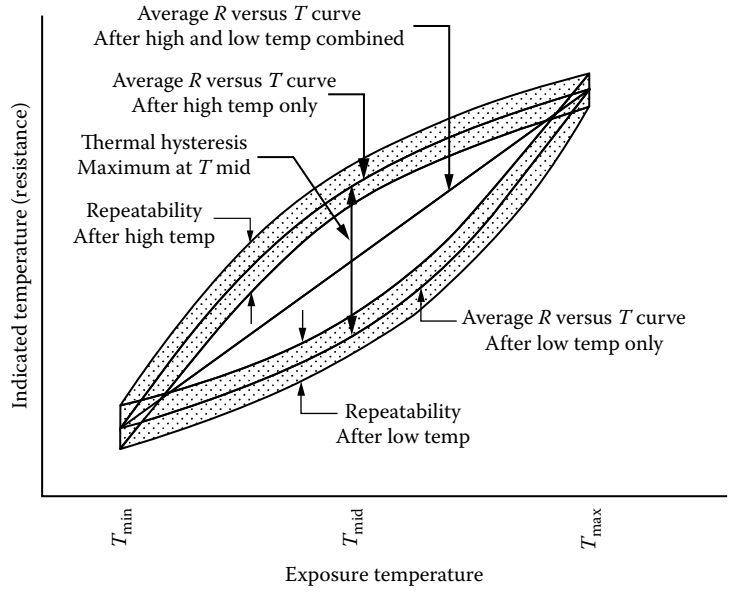


FIGURE 68.14 Hypothetical well-behaved hysteresis loops for industrial PRT.

One simplified method to estimate the hysteresis error at any temperature within the range of use is to perform a linear interpolation between the midpoint temperature and the temperature of interest. Since no known equation exists for this estimate, one is offered here for the purpose of this chapter and it is given as

$$\text{Hysteresis (°C)} = H\left(\frac{T_{\max} - T_{\min}}{100}\right)\left(1 - \left|\frac{\left(T_{\max} + T_{\min} - 2T\right)}{\left(T_{\max} - T_{\min}\right)}\right|\right) \tag{68.6}$$

where

H is the hysteresis error coefficient in % of temperature span
$T_{\max}$ is the maximum temperature of span
$T_{\min}$ is the minimum temperature of span
T is the temperature where hysteresis is being determined
(all temperatures are in °C)

To provide a context for the order of magnitude of the hysteresis error, several examples were created. Table 68.3 shows the estimated hysteresis error in °C calculated using this method for PRTs with varying specifications over several common temperature ranges.

By examining the values in Table 68.3, it is easy to see that the hysteresis error can be nearly negligible or approaching as much as ±1 °C.

68.5.8 Stem Conduction

Stem conduction is the error that results from the PRT sheath conducting heat into or out of the process of interest. ASTM offers a test method to compare PRT performance in the lab, specifying a maximum error of 0.13 Ω (0.25 °C) when the PRT is immersed 2 in. into an ice bath. International Electrotechnical Commission (IEC) defines the "minimum immersion depth" to be that where a 1 °C error is produced in

TABLE 68.3 Hysteresis Error Estimates for Common Temperature Ranges

	Temperature Range: −200 °C–600 °C			Temperature Range: 0 °C–400 °C			Temperature Range: 0 °C–200 °C			
Hysteresis Spec	0.02% of span	0.05% of span	0.10% of span	0.02% of span	0.05% of span	0.10% of span	0.02% of span	0.05% of span	0.10% of span	
	Error (±°C)									
Temperature (°C)	−200	0	0	0	—	—	—	—	—	—
	−100	0.04	0.10	0.20	—	—	—	—	—	—
	0	0.08	0.20	0.40	0	0	0	0	0	0
	100	0.12	0.30	0.60	0.04	0.10	0.20	0.04	0.10	0.20
	200	0.16	0.40	0.80	0.08	0.20	0.40	0	0	0
	300	0.12	0.30	0.60	0.04	0.10	0.20	—	—	—
	400	0.08	0.20	0.40	0	0	0	—	—	—
	500	0.04	0.10	0.20	—	—	—	—	—	—
	600	0	0	0	—	—	—	—	—	—

an 85 °C water bath. The test methods are an effective method to compare sensors, although the thermal dynamics of the installed sensor can produce different stem conduction results than those found from the lab test method. The specific installation, heat transfer conditions, and the temperature difference—process to ambient—can have a significant influence on the stem condition error.

To manage this error,

- Strive to achieve the ideal immersion depth of 10× the sensor diameter plus the element sensitive length. For a 0.25 ft diameter sheath with a 1 ft element length, the ideal immersion depth is 3.4 ft.
- Select a sensor design that is tip sensitive. These are typically smaller diameter, specifically in the element area.
- Insulate the portion of the sensor external to the process.

68.5.9 Calibration/Interpolation

These are errors that occur due to calibration uncertainty at the cal points or between cal points due to propagation of uncertainty, or curve fit errors. The ASTM and IEC standards do not address this error source beyond the requirement of the resistance tolerance limits. For most industrial sensor uses, these errors are less significant. To minimize these errors, select a calibration laboratory with low calibration uncertainties and a PRT with low repeatability and hysteresis error values.

68.5.10 Lead Wire

To provide a practical temperature measuring instrument, it is necessary to attach extension leads to the platinum sensing element. If not properly selected for the application, the number and type of extension leads used can introduce significant errors into the temperature measurement. See Figure 68.15 for the different connection methods.

The two-lead-wire circuit has no mechanism to compensate or eliminate the resistance of the lead wires from the resistance of the element. As the temperature increases, the resistance of the lead wires will also increase. Unlike the platinum element, the resistance in the lead wires increases in a nonlinear manner, which will result in inaccurate measurements. The two-lead-wire designs are only useful where extension lead resistance is insignificant to the sensing element, either short leads or large sensing resistance, and an accurate temperature measurement is not required.

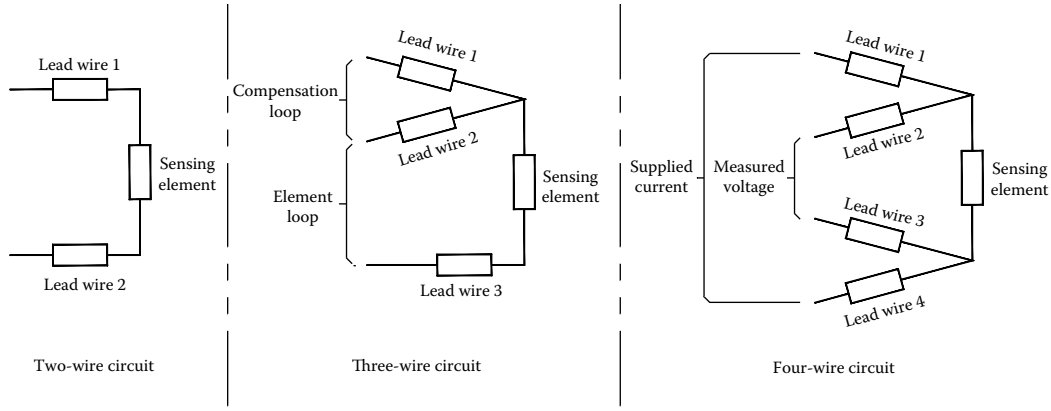


FIGURE 68.15 Temperature sensor lead wire configurations.

The three-wire configuration provides a compensation loop that can be used to subtract the lead wire resistance from the measurement to provide a value for just the element resistance. The resistances of two different electrical paths are used. The first is through the element like a two-wire measurement. The second is to one end of the element and back without going through the element to approximate the lead resistance. The measurement is obtained by subtracting the second electrical path from the first. Complete compensation in this method is predicated on the resistance of each of the leads being exactly equal. Unfortunately, this is seldom the case, and steps in the design and application of a three-wire sensor must be taken to maintain the accuracy of the measurement.

The method that offers the highest accuracy temperature measurement is the four-wire scheme. Having two wires either side of the element allows for a measurement current to be passed across the element via two of the leads and a voltage measurement to be taken through the other two leads. The resistance is simply then calculated using Ohm's law. The impedance of the meter taking the voltage measurement is sufficiently high to prevent any current from flowing through the leads being used for the voltage measurement. The current flowing through the other two leads and the element is the same throughout the circuit. Therefore, by measuring the voltage across the element and dividing it by the known current yields just the resistance of the element.

In applications where the sensor must be at a long distance from the measurement equipment, temperature transmitters are often used. Temperature transmitters are capable of measuring the resistance of the RTD probe and converting the measurement to a signal that can be transmitted longer distances, typically a 4–20 mA signal. It is not uncommon for measurement equipment, even respected brands, to take short cuts when it comes to lead wire connections. Be especially wary if the manual instructs the user to use jumpers to unused terminals. This will almost always cause the resistance of at least one lead to be included in the measurement. Measurement equipment that requires internal settings to be changed specifically for two-, three-, or four-wire measurement is fewer suspects. It is good practice to verify how the measurement is being taken by switching between settings and/or disconnecting a lead to verify that the output is congruent with the connections.

68.5.11 Self-Heating

Self-heating errors are caused by the inability of the element to dissipate the heat generated by the required power applied via the measuring current. The ASTM Standard requires the error to be a maximum of 1 °C when 33 mW is applied in 25 °C water. IEC requires a maximum error of 0.05 °C in 25 °C water when the maximum operating current is applied. These test methods are good laboratory comparison methods. For PRTs installed with proper immersion in a process, the operating current is 1 mA or less,

so the power (I^2R) for a 100 Ω PRT is also small (0.02–0.39 mW). Lager errors can occur in sensors with resistances in the 500–1,000 Ω range, or when the process exhibits poor heat transfer conditions such as still air or low-pressure gases.

68.5.12 Time Response

A time response-related error can be produced during temperature transients if the PRT cannot respond to the change in temperature fast enough. During steady-state or near-steady-state operation, this error is zero. ASTM and IEC do not define this error, although there is a test method to characterize a PRTs' response time for comparison purposes. When it is important to monitor transient conditions, this error can be minimized by selecting a sensor with a faster lab-tested response time and assessing the related rate of change to the process to best match the sensors' time response performance.

68.5.13 Thermal EMF

Thermal EMF errors are also known as the thermocouple effect. This error is caused by various wire compositions, wire connections in homogeneity of materials, and temperature gradients within the PRT. The ASTM and IEC Standards offer guidelines at high sensing currents—although when an EMF influence exists, it will have a greater effect at lower currents near the standard operating current. This error occurs primarily in direct current systems. To minimize thermal EMF errors, select a PRT with a low specified EMF. Also, using alternating current circuitry and appropriately selected transmitters can eliminate EMF influences.

68.6 Calibration and Maintenance

Testing programs such as periodic calibration or performance testing are essential to verify the accuracy and general health of PRTs. Some industrial PRTs are factory-calibrated to a temperature such as at the ice point, but PRT users might want to calibrate them at other temperatures, depending on their application. The periodic calibration results can be compared to prior or original calibration values to help determine if it is necessary to repair or replace the instrument, or if calibration is required more often.

An RTD's stability depends on its working environment. High temperatures can cause drift and contamination of the platinum wire. Shock, vibration, and rough handling will put stress in the platinum wire and change its characteristics, and eventually cause damage that may not be recoverable. Humidity can get into the sensor and reduce the IR. Under extreme operating conditions, a sensor should be calibrated more frequently. If several calibrations are completed without a significant change, then the time between calibrations can be extended based on the user's quality system guidelines; at least once a year is common.

68.7 Conclusion

Measuring temperature using RTDs is extremely common. The sensor design and packaging approaches have evolved over the years, yet the basic technology remains generally unchanged. The user or specifier of RTD-based measurements need to be aware that the values obtained and confidence on those results are very much dependent on the sensors used. Managing the sources of error, establishing a procedure to monitor the health and performance of the sensor, and properly installing the sensor into the system are necessary to ensure that the measurement result will meet the needs of the process control. The process performance characteristics should drive the decisions regarding element design, sensor packaging, sensor capability, calibration, and electronic interface needs.

Defining Terms

Accuracy: The degree of agreement between an actual measurement and its reference standard.

Alpha (α): The temperature coefficient of resistance of a PRT over the range 0 °C–100 °C. For example, α for a standard platinum resistance thermometer (SPRT) is 0.003925 Ω/Ω/°C.

Callendar–Van Dusen: An interpolation equation, which provides resistance/temperature values for PRTs.

Calibrate: To check, adjust, or determine an RTD's accuracy by comparing it to a standard.

Calibration, Calibration: A calibration technique commonly used with industrial PRTs. The PRT and a temperature standard are both measured at a stable temperature, then compared.

Calibration, Fixed Point: A calibration technique that uses the triple point, freezing point, or melting point of pure substances to generate a known and repeatable temperature. The sensor's resistance is measured at one of these points.

Connection Head: A housing that provides a terminal block for a sensor assembly, with threads for attaching it to a protection tube or conduit.

Defining Fixed Points: Specific temperatures that create the basis for the ITS-90.

Fixed Point: A reproducible temperature of equilibrium between the phase changes in a material.

Immersion Error: An error caused by heat conduction or radiation, both between the resistance thermometer element and environment external to the measurement system (also known as stem conduction error).

Interchangeability: The ability to interchange one sensor for another, without affecting the overall accuracy of the system.

Interchangeability Errors: Results from the manufacturing tolerances of the sensing elements and several of the performance characteristics including hysteresis repeatability and calibration errors.

Primary Standard: (Or SPRT) A platinum resistance thermometer that meets the requirements for establishing calibrations according to the ITS-90. This highly accurate instrument is intended for laboratory environments and is accurate to 0.001 °C.

PRT: Platinum resistance thermometer (PRT) is an RTD that uses platinum wire in the sensing element.

Response Time: The time required for the output of the sensor to change by 63.2% of the step change in temperature.

RSS: Root-sum-square. A mathematical method to quantify various errors of a system.

RTD: Resistance temperature detector.

SPRT: A PRT that has its temperature resistance relationship determined by reference to a primary standard of temperature.

Self-Heating Error: Caused because the current that measures sensor resistance also heats the sensor. This causes the temperature reading to be higher than the actual temperature.

Span: The algebraic difference between the upper and lower range values of an instrument.

Stability: The state of being resistant to change or deterioration.

Stem Conduction Error: Heat transfer along the sheath or lead wires of a probe, or along the length of a thermowell. This results in the sensor being at a different temperature than the medium.

Thermowell: Closed sleeve for protecting temperature sensors from the rigors of industrial processes while permitting accurate measurement.

References

1. IEC, *Industrial Platinum Resistance Thermometer Sensors*, IEC 60751:2008 Genéve, Suisse: Bureau Central de la Commission Electrotechnique Internationale, 2008.
2. ASTM Standards, Standard Specification for Industrial Platinum Resistance Thermometers, Standard E 1137/E 1137M-08, 2008.
3. NIST Technical Note 1265, Guidelines for Realizing the International Temperature Scale of 1990 (ITS-90), National Institute of Standards and Technology, 1990.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

69

Thermistor Thermometers

69.1 Principles of Operation **69**-1
69.2 Calibration Equations **69**-4
69.3 Instrumentation **69**-5
69.4 Limitations in performance **69**-7
Self-heating • Stray Thermal Influences • Settling Time • Thermistor Stability • Resistor Stability • Lead Resistances • Offset voltages • Insulation Resistance
69.5 Choice of Thermistor **69**-10
69.6 Thermistors in Other Applications **69**-11
NTC Thermistors • PTC Thermistors • Silicon Thermistors
References **69**-14

Rod White
Measurement Standards Laboratory of New Zealand

Meyer Sapoff
MS Consultants

Thermistors are temperature-sensitive semiconducting devices. There are several types used in a wide variety of temperature measurement, switching, and sensing applications. This section will focus primarily on negative temperature coefficient (NTC) thermistors designed specifically for temperature measurement. While NTC thermistors are used over a very wide range of temperatures, from 1 K to 1000°C, we will focus on the range −80 °C to 300 °C, where they are most commonly used and offer genuine advantages over platinum resistance thermometers for some applications. A brief overview of other types of thermistors and other applications for thermistors is given at the end of the chapter.

NTC thermistors are suited to a number of specialized thermometry applications because they are smaller, faster, and more sensitive than almost all other contact thermometers. However, they are very nonlinear and have a narrow temperature range. Traditionally, thermistors have also had a reputation for instability, but thermistors are now readily available for temperature ranges within −40 °C to 60 °C with stabilities of a few tenths of a millikelvin per year. Their short-term performance approaches that of standard platinum resistance thermometers, but at a much lower cost. Thermistors are also available with selected performance characteristics, including interchangeability to 0.05 °C, aged and selected for stability, as well as qualification for military and space applications.

Reviews of the properties and applications of thermistors can be found in [1,2] and, with simpler overviews, in [3,4].

69.1 Principles of Operation

NTC thermistors are extrinsic semiconductors manufactured from mixtures of metal oxides heated to high temperatures to form a polycrystalline ceramic. Ordinarily, the individual oxides have a large energy gap between the vacant conduction band and the fully occupied valence band, which makes them good electrical insulators. The mixture of oxides, however, has intermediate electronic states that make the ceramic a semiconductor. As the temperature increases, increasing numbers of electrons jump to higher-energy states enabling conduction in the conduction bands, and, by creating holes, enabling

conduction in the valence band. The total number of mobile charge carriers, n, depends on the energies of the electrons as determined by the Boltzmann distribution and on the density of electronic states near the band edges [5]:

$$n \propto T^{3/2} \exp\left(\frac{-E_g}{2kT}\right) \tag{69.1}$$

where
E_g is the excitation energy for the carriers
k is Boltzmann's constant
T is temperature (K)

The leading $T^{3/2}$ term in Equation 69.1 is due to the density of states at the band edges and tends to be compensated by a similar term in the carrier mobilities (due to the density of states for phonons that scatter the carriers). The combination of effects gives an overall resistance–temperature behavior that is roughly approximated by

$$R(T) = R_0 \exp\left(\frac{E_g}{2kT}\right) \tag{69.2}$$

where R_0 is a constant. Usually, this equation is rewritten in the form

$$R(T) = R(T_0)\exp\left[\beta\left(\frac{1}{T} - \frac{1}{T_0}\right)\right], \tag{69.3}$$

where T_0 is some convenient reference temperature, often 298.15 K (25 °C). The parameter β is a characteristic of the thermistor material with typical values in the range 2000–6000 K. Figure 69.1 plots the

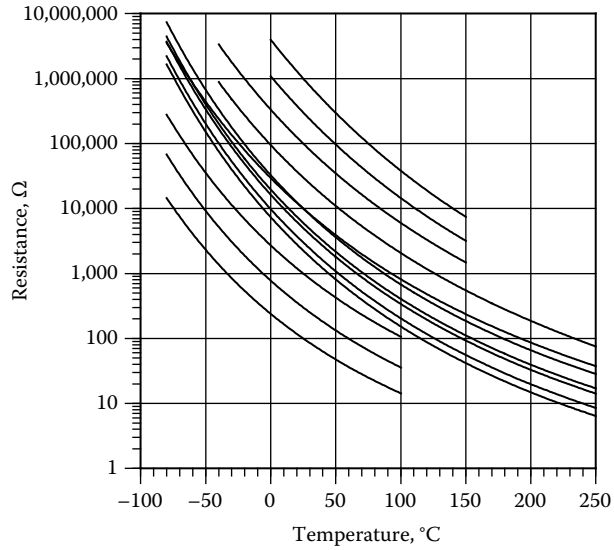


FIGURE 69.1 The resistance–temperature characteristics for a range of thermistors showing the typical resistance and temperature ranges.

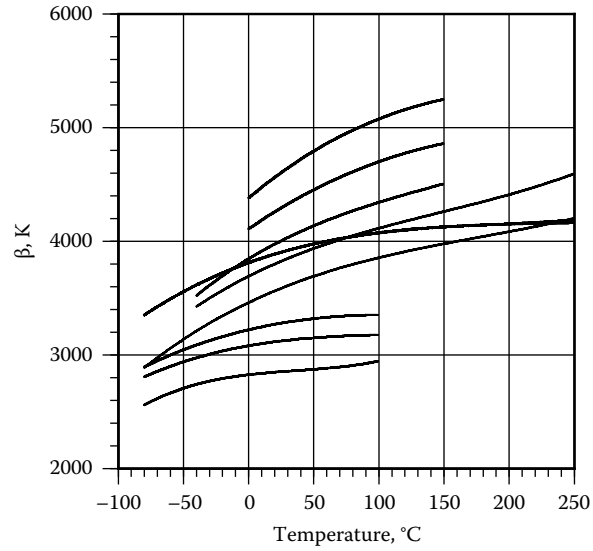


FIGURE 69.2 The β values for the thermistors of Figure 69.1. Some of the thermistors have the same β value.

resistance–temperature characteristic for a range of commercially available thermistors. Figure 69.2 plots the β values for the same set of thermistors and shows the typical variation in β with temperature.

The sensitivity, S, of the thermistor is defined as the relative change in resistance for a change in temperature and is given by

$$S = \frac{1}{R}\frac{dR}{dT} = \frac{d\ln(R)}{dT} = -\frac{\beta}{T^2}. \tag{69.4}$$

This number is useful for evaluating measurement errors, as will be shown later. Figure 69.3 shows the sensitivities for the thermistors of Figure 69.1. At room temperature, they range between about −0.03/°C

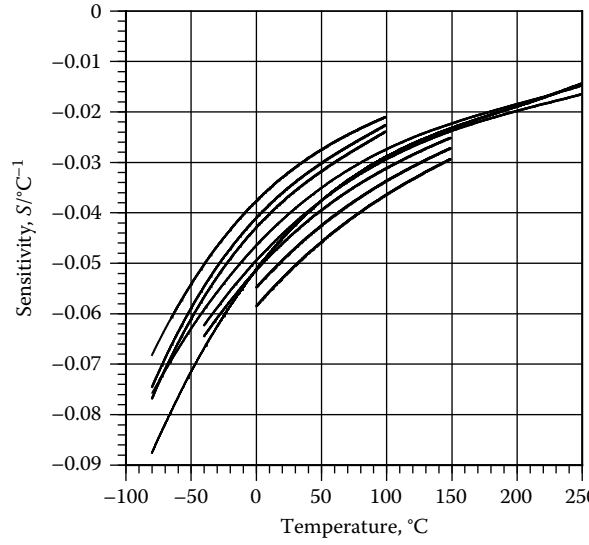


FIGURE 69.3 The sensitivities, the fractional change in resistance per degree Celsius, as a function of temperature for the thermistors of Figure 69.1.

and −0.05/°C. That is, thermistors are about 10 times more sensitive to temperature than platinum resistance thermometers: the sensitivity for platinum resistance thermometers is more nearly constant and approximately equal to the alpha value, $\alpha = 3.85 \times 10^{-3}$/°C.

69.2 Calibration Equations

One of the notable features of Figure 69.2 is that the β value for a thermistor varies with temperature, which is characteristic of extrinsic semiconductors. This means that (3) is only useful as a calibration equation for temperature ranges of no more than 20 °C–30 °C. For high-accuracy or wide temperature ranges, more complex equations are required. Although a wide range of calibration equations have been explored, for example, [6–10], there are two extended series expansions of Equation 69.3 that are most commonly used:

$$\frac{1}{T} = A + B\ln\left(\frac{R}{R_0}\right) + C\left[\ln\left(\frac{R}{R_0}\right)\right]^2 + D\left[\ln\left(\frac{R}{R_0}\right)\right]^3 + \cdots \tag{69.5}$$

and

$$\ln\left(\frac{R}{R_0}\right) = A + \frac{B}{T} + \frac{C}{T^2} + \frac{D}{T^3} + \cdots \tag{69.6}$$

where R_0 is usually a convenient reference resistance, for example, 1 Ω or 1 kΩ, depending on the measurement units. The number of terms in each series, N, is chosen according to the temperature range and the accuracy required, with more terms required as the range or accuracy increases. Most commonly, the equations are used with four terms, but occasionally up to five terms. Equation 69.5 has the advantage of giving temperature directly from the measured resistance and tends to be used more frequently. Both equations are linear in their coefficients so are amenable to least-squares fitting using the linear algorithms found in many spreadsheet and graphing applications. Figure 69.4 summarizes the performance of Equation 69.5 when different numbers of terms are used. Note that the $N = 2$ case corresponds to Equation 69.3.

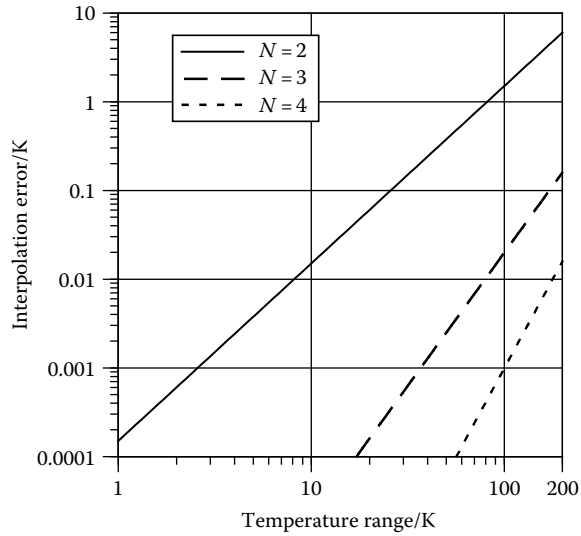


FIGURE 69.4 The typical standard deviation of interpolation errors versus the number of terms used in the calibration equation. The lines are indicative only: the errors tend to be greater for temperatures below 0 °C, and for thermistors with a high β value.

Among a number of equations investigated by Steinhart and Hart [6] is an alternative calibration equation,

$$\frac{1}{T} = A + B\ln\left(\frac{R}{R_0}\right) + D\left[\ln\left(\frac{R}{R_0}\right)\right]^3 \tag{69.7}$$

which is, unfortunately, in common use and recommended by many thermistor manufacturers. The original recommendation for Equation 69.7 was based on a numerical error [7], and use of Equation 69.7 should be discouraged. The equation has some peculiar attributes that make its performance depend on the $R(25\ °C)$ value of the thermistor, the unit of measurement, R_0, and whether thermistors are connected in series or parallel [10,11]. Under very specific conditions, its performance can approach, but never exceed, that of Equation 69.5 with $N = 4$.

69.3 Instrumentation

For the highest accuracy and for temperature ranges of more than a few degrees, the thermistor resistance is best measured directly using a good-quality digital voltmeter in combination with either Equation 69.5 or 69.6, for example, [12]. Care should be taken to account for lead resistance effects and offset voltages and to make sure that the self-heating of the thermistor due to the sensing current is not excessive (see the discussion of errors later). With care in the design of the measurement, selection of the thermistor, use, and calibration, accuracies approaching 0.001 °C are achievable. Thermistor-based thermometers are commercially available with this level of accuracy.

For narrow temperature ranges, and the highest resolution, purpose-built measurement circuits can be advantageous, and an example is shown in Figure 69.5.

The output voltage of the circuit in Figure 69.5 is

$$V_{out} = \frac{R(T)R_{in} - rR_f}{R_{in}\left[r + R(T)\right]}V_{ref}, \tag{69.8}$$

where the resistors are as indicated. The condition for zero output voltage is

$$\frac{R_{in}}{R_f} = \frac{r}{R(T_0)} \tag{69.9}$$

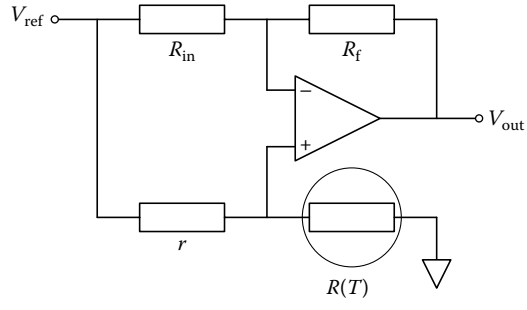


FIGURE 69.5 A simple narrow-range thermistor thermometer using a half-bridge formed by the thermistor, $R(T)$, and the linearizing resistor, r. The operational amplifier with the resistors R_{in} and R_f provides buffering, gain, and offset features.

where T_0 is the nominal operating temperature of the thermometer. The output voltage is a nonlinear function of the thermistor resistance, and this nonlinearity is used to compensate for the thermistor nonlinearity. By replacing $R(T)$ in Equation 69.8 by 69.3, and differentiating with respect to T, it is found that the nonlinearity is largely eliminated if r is chosen so that [13]

$$r = \frac{\beta - 2T_0}{\beta + 2T_0} R(T_0). \tag{69.10}$$

The key feature of the linearization circuit is the resistance r in series with the thermistor (the half-bridge in Figure 69.5). The same scheme can be applied when a single thermistor is used in more complex resistor networks by recognizing that r corresponds to the Norton or Thévenin equivalent resistance seen by the thermistor. Near room temperature, r is often found to be about $0.6 \times R(T_0)$. With the linearizing resistor, r, set according to Equation 69.10 and the gain setting resistors R_{in} and R_f set to the ratio given by Equation 69.9, the output of the thermometer in Figure 69.5 is approximately

$$V_{out}(T_0 + \Delta T) \approx -\frac{\beta + 2T_0}{2T_0^2}\left[\Delta T - \frac{1}{12}\frac{\beta^2}{T_0^4}\Delta T^3\right] V_{ref}. \tag{69.11}$$

The residual nonlinearity is almost purely a cubic function of temperature, as shown in Figure 69.6, with maximum errors of about ±0.1 °C for a 30 °C range. Equation 69.11 shows that the residual nonlinearity scales as cube of the temperature range. Sapoff and Oppenheim [14] describe refinements of this basic linearization technique and give a wide range of examples for different applications.

One of the big advantages of purpose-built circuits like that of Figure 60.5 is, with care and appropriate signal conditioning, a resolution of better than 10 μK can be achieved relatively easily. This enables thermistors to be used for narrow-range high-resolution applications such as differential thermometry, calorimetry, and precision temperature control.

For temperature ranges wider than a couple of tens of degrees, linearization of the thermistor response becomes more difficult. Practical solutions include the use of logarithmic amplifiers [15,16] and oscillators [17], which work well for thermistors with a nearly constant β value. Some commercial thermistor thermometers use this technique and yield accuracies of a few tenths of a degree over a 100 °C range. Another solution is to expand the linearizing scheme of Figure 69.5 to resistance networks in which two or more thermistors are used in combination with linearizing resistors [18] (see Figure 69.7 for example). Commercial thermistor networks optimized for specific temperature ranges are available with linearities as good as ±0.03 °C for a 40 °C temperature range and ±0.2 °C for a 100 °C temperature range.

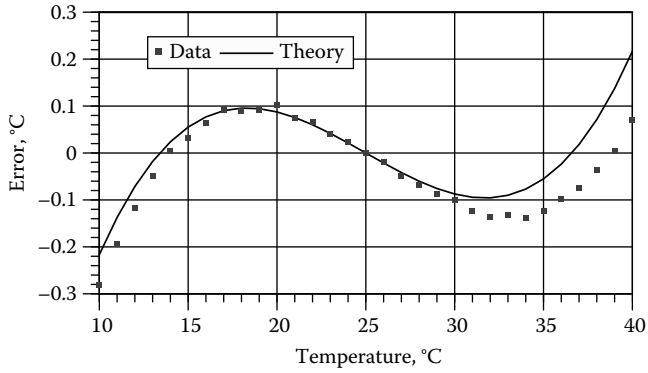


FIGURE 69.6 The nonlinearity for the thermistor thermometer of Figure 69.5.

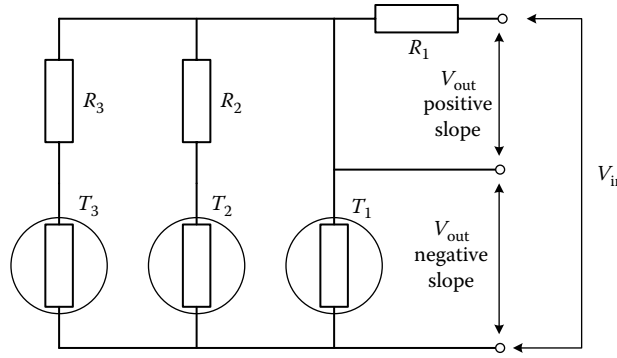


FIGURE 69.7 A three-element linear thermistor network.

A variation on the multithermistor scheme is to sum the outputs from several single-thermistor networks. While superficially the same as the type of network illustrated in Figure 69.7, the approach uses additional resistors and creates extra degrees of freedom in the design, which enables further reduction in the nonlinearities. A three-thermistor network of this design has peak nonlinearities of about 16 mK over a 100 °C range [19], more than 10 times better than the network of Figure 69.7.

69.4 Limitations in Performance

This section describes in detail the most significant errors affecting thermistor thermometers. To illustrate the typical magnitude of the errors, numerical examples are included.

69.4.1 Self-heating

In order to measure a resistance, a current must be passed through the thermistor, which leads to Joule heating and a small temperature increase called the self-heating error. The error, ΔT, is proportional to the power dissipated and the thermal resistance between the thermistor and its environment:

$$\Delta T = \rho I^2 R(T) = \rho \frac{V^2}{R(T)}, \tag{69.12}$$

where

I and V are the sensing current and the voltage across the thermistor, respectively
ρ is the thermal resistance between the thermistor and its surroundings

The thermal resistance depends in part on the dimensions of the thermistor and the material from which it is made, as well as the thermal conductivity and, if fluid, the velocity and viscosity of the medium in which the thermistor is immersed. Self-heating can be a serious problem for measurements made over a wide temperature range. If a constant sensing current is used, the power dissipated (I^2R) at low temperatures becomes large, and if constant voltage excitation is used, the power at high temperatures (V^2/R) becomes a problem.

The thermal resistance is often expressed as the dissipation constant, which is the reciprocal of the thermal resistance. The dissipation constant measures the power required to raise the thermistor temperature 1 °C and is usually reported in units of mW/°C. Typical values of the dissipation constants range from 0.05 to 20 mW/°C in stirred fluids, but are very dependent on the environment so may vary as much as 10–100 times between still air and stirred water. For a 10 kΩ thermistor with a dissipation constant of 10 mW/°C driven with a sensing current of 100 μA, the self-heating error is 0.01 °C.

In many applications, the thermal resistance may fluctuate due to turbulence in the air or stirred fluids. This causes the self-heating error to fluctuate and can limit the resolution of thermistors, when used, for example, in precision temperature controllers or differential thermometers.

69.4.2 Stray Thermal Influences

The generally high thermal resistance between the thermistor and the environment means that thermistors are prone to stray thermal influences from either infrared radiation or heat conducted along the lead wires to the thermistor. This effect may be exacerbated by the poor thermal design of many commercial thermistor probes. The "stem conduction" or "immersion errors" tend to fall exponentially with immersion depth, and a useful rule of thumb is that the error is reduced by a factor of 100 for each 10 diameters of immersion [20]. For example, to measure temperatures near 120 °C (100 °C different from ambient temperature) with accuracy of 0.01 °C, a stainless-steel sheathed probe may need to be immersed as much as 20 diameters.

Stray thermal effects are a particular problem when thermistors, and other contact thermometers, are used to measure air and surface temperatures. For air temperature measurements, the leads should be long and thermally anchored to an object at a temperature as close as practical to that of the air, so that the leads adjacent to the thermistor also come to thermal equilibrium with the air and do not provide a source of heating or cooling. Polished-metal cylinders can be used as radiation shields to reduce the influence of infrared radiation sources such as incandescent lamps, sunlight, warm electronics, heaters, and people [21]. With surface temperatures, a sufficient length of the leads must be thermally anchored to the surface to ensure that the thermistor is at the surface temperature. Radiation shielding is often required.

69.4.3 Settling Time

One of the major advantages of thermistors is their small size and fast response time. The responses times are normally characterized by the 1/*e* time constant, which is the constant, τ in the model of the temperature error, ΔT, due to a step change in temperature:

$$\Delta T = \Delta T_{\text{step}} \exp\left(\frac{-t}{\tau}\right), \tag{69.13}$$

where

t is the time following the step change

ΔT_{step} is the step change to which the thermistor is responding

This model suggests that the error is reduced by a factor of 100 for each five time constants. For example, a thermistor with a time constant of 100 ms will have settled to within 1% of a step change within 500 ms. The time constants for thermistors range from about 20 ms for ultra-small beads to more than 200 s for large disk types and are very strongly correlated with size and thermal resistance.

Bead thermistors with environmental protection such as glass or epoxy coatings, or a stainless steel sheath, typically have time constants in the range of 1–20 s. Often, larger and more complex probes can exhibit settling behavior that is more complex than that suggested by Equation 69.13 and may be characterized by two or more time constants. For these probes, the settling behavior may be more usefully characterized by the 10%–90% settling time, and many industrial and military applications use such specifications.

As with the dissipation constants, the time constants are very strongly dependent on the medium into which the thermistors are immersed. A thermistor in a well-stirred low-viscosity fluid may be 10–100 times faster than the same thermistor in still air.

69.4.4 Thermistor Stability

There are several effects causing instability in thermistors [22], including mechanical cracking of the thermistor body with temperature cycling, drift at high temperatures due to ingress of atmospheric gases, changes in the atomic and crystallographic structure, and changing contact resistance between the leads and the thermistor body. The most stable thermistors are bead thermistors encapsulated in glass, closely followed by epoxy encapsulated beads. Within the range –20 °C to 60 °C, selected and pre-aged thermistors are often stable to better than a few tenths of a millikelvin per year, and thermistors used as high as 120 °C can be stable to a few millikelvin [23,24]. Thermistors with $R(25\ °C)$ resistances in the range 2–10 kΩ appear to be the most stable. For further information on thermistor stability, see Ref. [25–29].

69.4.5 Resistor Stability

A thermistor-based thermometer is only as stable as the other components in the instrument. The effect of the instability of the other resistors in Figure 69.5, for example, can be inferred from the zero-output condition (Equation 69.9), so that the apparent change in the thermistor resistance due to changes in the other three resistors is

$$\frac{\Delta R(T)}{R(T)} = \frac{\Delta r}{r} + \frac{\Delta R_{\mathrm{f}}}{R_{\mathrm{f}}} - \frac{\Delta R_{\mathrm{in}}}{R_{\mathrm{in}}}, \tag{69.14}$$

and hence, from Equation 69.4, the apparent change in measured temperature caused by changes in the resistances is

$$\Delta T = -\frac{T^2}{\beta}\left(\frac{\Delta r}{r} + \frac{\Delta R_{\mathrm{f}}}{R_{\mathrm{f}}} - \frac{\Delta R_{\mathrm{in}}}{R_{\mathrm{in}}}\right). \tag{69.15}$$

The effects of resistor drift on the observed temperature can be estimated using this equation. If the changes in the resistors are due to changes in the ambient temperature, T_a, then the apparent temperature change is

$$\Delta T = -\frac{T^2}{\beta}\left(\alpha_r + \alpha_{R\mathrm{f}} - \alpha_{R\mathrm{in}}\right)\Delta T_a, \tag{69.16}$$

where α_r, $\alpha_{R\mathrm{f}}$, and $\alpha_{R\mathrm{in}}$ are the temperature coefficients of the three resistors.

Equation (69.16) shows that the effects of ambient temperature are reduced by the ratio of the temperature coefficients. Good-quality resistors have temperature coefficients of 0.0005%/°C or less, being about 1/10,000 that of thermistors. Hence a 10 °C change in the temperature of the circuit will cause an apparent change of about 1 mK in the measured temperature.

69.4.6 Lead Resistances

The high sensitivity and high resistance of thermistors mean that for many measurements, a two-wire resistance measurement is satisfactory and provides a useful simplification. However, the lead resistances are often forgotten and can become a problem at higher temperatures when the thermistor resistance is low. The error due to a lead resistance, R_{L}, is

$$\Delta T = -\frac{T^2}{\beta}\frac{R_{\mathrm{L}}}{R(T)}. \tag{69.17}$$

The error increases very rapidly with temperature due to the falling resistance of the thermistor and the T^2 temperature dependence. For a 10 kΩ thermistor with a β value of 3600 K and a 1 Ω lead resistance, the error at 25 °C (~300 K) is 2.5 mK, while at 50 °C, the error increases to 7.4 mK. Where lead resistances are a problem, a four-wire measurement should be used [21].

69.4.7 Offset Voltages

Typically, a thermistor resistance measurement involves a voltage measurement, the accuracy of which may be limited by the resolution of a voltmeter, or the input-offset voltage and input bias currents of an operational amplifier, and/or thermoelectric effects. The error in a temperature measurement due to an error, ΔV, in the voltage measurement is

$$\Delta T = \frac{T^2}{\beta}\frac{\Delta V}{I_0 R(T)}, \tag{69.18}$$

where I_0 is the sensing current. For a 10 kΩ thermistor with a β value of 3600 K, a 100 μV error, and a sensing current of 100 μA, the error at 25 °C (~300 K) is 2.5 mK.

69.4.8 Insulation Resistance

At low temperatures, the thermistor resistance becomes very large, often greater than 10 MΩ. In such applications, care should be given to the insulation on the connecting leads to ensure that the insulation resistance does not shunt the measuring current. Some pigmented plastics can be very poor in this respect. The temperature error due to an insulation resistance R_{ins} shunting the thermistor is

$$\Delta T_{\text{ins}} \approx -\frac{T^2}{\beta}\frac{R(T)}{R_{\text{ins}}}, \tag{69.19}$$

which increases in proportion to the thermistor's resistance (low temperatures). A 10 kΩ thermistor shunted by 10 MΩ will exhibit an error of 25 mK.

69.5 Choice of Thermistor

The temperature range of commercially available thermistors varies from 1 K to 1000 °C with large differences in the R(25 °C) resistance, β values, the metal oxides used, and the encapsulation and sheath assemblies. The most stable thermistors for temperature measurement have a much restricted range: glass-encapsulated thermistors are most useful in the range from about −40 °C to 100 °C, but some are usable from −80 °C to 250 °C. Epoxy-encapsulated thermistors have an upper temperature limit of about 150 °C.

Thermistors are made in many different forms. Bead thermistors are supplied with leads attached during the manufacturing process but may be further encapsulated in glass or epoxy, as well as diode-like forms suitable for automatic insertion into printed circuit boards. Other forms include flakes, disks, surface-mount, and rod types, which may be supplied with or without leads. When supplied without leads, they have metallized surfaces for the contacts.

There are many large companies manufacturing thermistors, and comparing specifications from different suppliers can be very helpful to highlight limitations in the specifications, manufacturing tolerances, and sheath assemblies that might be relevant to any application. Most of the manufacturers also

provide extensive application information. Thermistors are available to meet military specifications, space qualification, and can be selected for long-term stability and interchangeability. Military standard MIL-PRF-23648 [30] not only provides specifications for thermistors, but also gives the measurement conditions for the various electrical and thermal properties of thermistors.

69.6 Thermistors in Other Applications

The term thermistor is applied generically to any semiconducting device with a high temperature coefficient of resistance. While NTC thermistors are the best of the devices for thermometry, once other fields of application are considered, other types of thermistors, including positive temperature coefficient (PTC) ceramic and silicon thermistors, become useful.

69.6.1 NTC Thermistors

When used for thermometry, thermistors are operated with a low current to avoid self-heating errors. However, for many applications, NTC thermistors are deliberately operated at very high currents to exploit the thermal feedback that occurs [31]. If a thermistor is operated at a nominal ambient temperature T_a and with a sensing current I_0, then the thermistor temperature can be written as

$$T = T_a + \rho P + \rho I_0^2 R(T_a) \exp\left[\beta\left(\frac{1}{T} - \frac{1}{T_a}\right)\right], \tag{69.20}$$

where P is any additional power applied to the thermistor (e.g., microwave or optical) and ρ is the thermal resistance between the thermistor and its environment. The essential feature of the model equation is that any power applied to the thermistor, either through Joule heating $I_0^2 R(T)$ or external power P, causes the temperature of the thermistor to rise and, hence, its resistance to fall. Equation 69.20 can be solved numerically for T and then the voltage $I_0 R(T)$ plotted against the current to give a voltage–current characteristic for the thermistor. Example curves are shown in Figure 69.8.

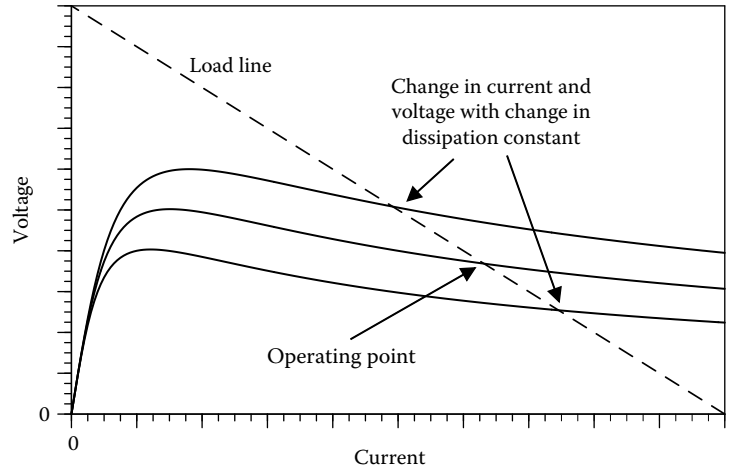


FIGURE 69.8 The voltage–current characteristic for a thermistor operated at high currents and with three different values of the dissipation constant. The load line is the voltage–current characteristic for a Thévenin voltage source. The intersection of the load line with the thermistor curves gives the operating point for the thermistor when driven by the source.

There are three distinct operating regimes in Figure 69.8. At low currents, the power dissipated is small compared with the dissipation constant, and the self-heating effect is negligible. Under these conditions, the thermistor resistance depends only on the temperature, and the voltage across the thermistor is directly proportional to the current. Consequently, at low currents, the ratio of the voltage to the current is equal to the zero-power resistance of the thermistor. This is the regime used for temperature measurement, as described in the previous sections.

As the current increases, the effects of self-heating become more evident, the thermistor temperature rises, its resistance begins to decrease, and the slope of the voltage–current characteristic decreases. This trend continues until the slope of the voltage–current characteristic is zero. The temperature, voltage, and current at the peak of the curve are [31]

$$T_{\text{peak}} = \frac{\beta - \sqrt{\beta^2 - 4\beta T_a}}{2}, \tag{69.21}$$

$$V_{\text{peak}} = \sqrt{R(T_{\text{peak}})\rho\left(T_{\text{peak}} - T_a\right)}, \tag{69.22}$$

and

$$I_{\text{peak}} = \sqrt{\frac{\rho(T_{\text{peak}} - T_a)}{R(T_{\text{peak}})}}. \tag{69.23}$$

Beyond the peak voltage, the slope of the curve continues to decrease, and the thermistor exhibits a negative resistance characteristic, that is, dV/dI_0 is negative.

Also shown in Figure 69.8 is a straight line corresponding to the voltage–current characteristic of a Thévenin voltage source. The characteristic is usually called a "load line" and has the equation

$$V = V_{\text{Thev}} - IR_{\text{Thev}} \tag{69.24}$$

where V_{Thev} and R_{Thev} are the voltage and resistance in the Thévenin equivalent circuit for the circuit around the thermistor. The load line intersects with the thermistor's characteristic at the points indicated. Applications based on self-heated thermistors work by detecting changes in the operating point of the thermistors. For example, Figure 69.8 shows three voltage–current curves for the same thermistor but with different values of the dissipation constant. The changes in voltage and current between the points on the load line indicate the circuit's response to the changes in the dissipation constant. The circuit could be used in a wide variety of applications where, by measuring changes in the dissipation constant, the circuit can be used for leak detection, flow measurement, and level detection.

Once the operating point of the thermistor is chosen, the sensitivity of the circuit in the vicinity of the operating point can be expressed as a partial derivative of the voltage to all of the other variables in the circuit. For example,

$$\Delta V = \mu_{T0}\Delta T_0 + \mu_{I0}\Delta I_0 + \mu_\delta \Delta\rho + \mu_{V0}\Delta V_0 + \mu_P \Delta P \cdots \tag{69.25}$$

where the sensitivity coefficients $\mu_{T0}, \mu_{I0}, \mu_\rho, \mu_{V0}, \mu_P \ldots$, are the respective derivatives of the voltage while all other variables are constant. For example,

$$\mu_{T0} = \left.\frac{\partial V}{\partial T_0}\right|_{\rho, I, P, \ldots = \text{constant}}. \tag{69.26}$$

In some cases, mathematical expressions can be determined for the sensitivity coefficients [31]. The key to using the thermistor as a sensor is to design the circuit so that the sensitivity of the circuit to the variable of interest is as large as practicable while all other sensitivity coefficients are as small as practicable. It is not possible to have all other coefficients equal to zero, so the circuit is often designed as a differential system using two identical thermistor subcircuits with one measuring the variable of interest and the other serving as a reference. In this way, unwanted influences affect both subcircuits equally and the differential output signal does not change. Note too that the load line in Figure 69.8 could be placed to intersect the curve at one, two, or three points. In some cases, the circuit will exhibit switching behavior as it jumps from one operating point to another in response to external changes.

Sapoff [31] cataloged a wide range of applications for thermistors and describes many examples of circuits using different operating principles:

- Variation in dissipation constant and time constant: vacuum manometers, flow meters, velocity meters, thermal conductivity, leak detectors, gas chromatography, liquid level, surge suppression, filament protection, overload protection, and switching.
- Variation in the load line (over a fixed current–voltage curve): oscillator amplitude and/or frequency control, gain control, level stabilization, equalization, voltage regulators, signal limiters, signal expansion and compression, switching.
- Variation in ambient temperature: principally thermometry including measurement, control, and alarm applications.
- Power measurement: bolometry for optical and microwave measurement, ac-to-dc conversion, and true rms measurement.

69.6.2 PTC Thermistors

The PTC thermistor refers to another family of semiconducting ceramic devices with an extremely variable temperature coefficient that is dominated by a temperature region in which the resistance rises extremely rapidly [32]. This makes the thermistors ideally suited to switching applications, and typically they are manufactured in families with specified switching temperatures. Figure 69.9 shows a sample of the resistance–temperature characteristics for a family of PTC thermistors.

At low temperatures, the resistance is low, and the thermistor exhibits a small NTC of resistance on the order of –1%/°C. As the temperature increases, the temperature coefficient of resistance becomes positive and the resistance begins to rise, and then rise rapidly. At the switching or "threshold" temperature,

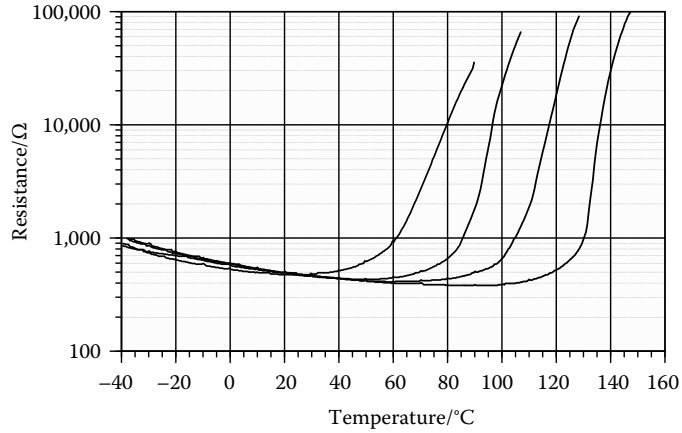


FIGURE 69.9 A sample of the resistance–temperature characteristics for a family of PTC thermistors with nominal switching temperatures of 75 °C, 95 °C, 115 °C, and 135 °C.

the rate of rise becomes very rapid and the resistance–temperature characteristic becomes very steep, sometimes as high as 100%/°C. At temperatures above the switching range, the resistance reaches a maximum value beyond which the temperature coefficient becomes negative again.

The resistance–temperature characteristic of the PTC thermistor can be exploited in a number of ways. A simple application is to use the thermistor as a heating element. The power dissipated is $V^2/R(T)$, and as the temperature rises in response to the heat, the applied power falls so that the thermistor can be used as a temperature controller or simply to limit the current flowing in the circuit. If the joule heating is sufficient to maintain the thermistor temperature when in its high-resistance state, the thermistor may be used for delaying the switching of start-up mechanisms in electronic equipment such as the auxiliary windings in electric motors. When used as a liquid level or flow sensor, the thermistor is forced back into its low-resistance state by changing its dissipation coefficient. These simple examples show that the thermistor can have a wide range of different behaviors, which makes them suitable for heating, current limiting, motor start, overcurrent protection, temperature control, alarm, time delay, level, and flow sensing applications.

PTC thermistors are available in many different forms with the form usually designed for specific applications. The switching temperature can be varied between about 70 °C and 240 °C by altering the chemical composition of the ceramic. As with the NTC thermistors, most manufacturers provide extensive literature supporting their selection and application.

69.6.3 Silicon Thermistors

Silicon PTC thermistors, also known as "silistors" or "sensistors," are typically single-crystal silicon devices that have been heavily doped to give the silicon a resistance–temperature characteristic similar to that of a metal [33]. The resistance–temperature characteristic is predominantly linear with an additional weak quadratic dependence that give the device an almost constant sensitivity of about 0.7%/°C for all temperatures. It is about twice that of platinum resistance thermometers. The main advantages of the silicon thermistors are the higher sensitivity, higher resistance (typically 100 Ω to 5 kΩ, but some up to 40,000 kΩ), and very much lower cost (below US$1 in large numbers). Although less accurate and less stable than NTC thermistors and platinum resistance thermometers, the low cost and ease of use make them the sensor of choice in many low-accuracy high-production applications.

The most common types of silicon PTC have the same package as small glass diodes with diameters of 1.8–2.5 mm and lengths of 3.8–7.5 mm, and the molded epoxy package with diameters of 3.6–6.0 mm and lengths of 10.4–15.0 mm. Such units also have axial leads, so are suited for electronic applications with automated component insertion. Epoxy-coated chips and disks with radial leads are also available, with diameters of approximately 3 mm. Silicon PTC thermistors that comply with MIL-T-23648 have an operating temperature range of −55 °C to 125 °C. Commercial versions can be obtained with a range of −65 °C to 300 °C.

References

1. Sachse, H. B., *Semiconducting Temperature Sensors and Their Applications*, John Wiley & Sons: New York, 1975.
2. Hyde, F. J., *Thermistors*, Iliffe Books: London, U.K., 1971.
3. McGee, T. D., *Principles and Methods of Temperature Measurement*, John Wiley & Sons: New York, 1988.
4. Michalski, L., Eckersdorf, K., and McGhee, J., *Temperature Measurement*, John Wiley & Sons: Chichester, U.K., 1991.
5. Sze, S. M., *Physics of Semiconductor Devices*, 2nd edn., John Wiley & Sons: New York, 1981.
6. Steinhart, J. S. and Hart, S. R., Calibration curves for thermistors, *Deep Sea Research*, **15**, 497–503, 1968.

7. Bennett, A. S., The calibration of thermistors over the range 0–30°C, *Deep Sea Research*, **19,** 157–163, 1971.
8. Sapoff, M., Siwek, W. R., Johnson, H. C., Slepian, J., and Weber, S., The exactness of fit of resistance-temperature data of thermistors with third-degree polynomials, In *Temperature: Its Measurement and Control in Science and Industry*, J. F. Schooley (ed.), Vol. 5 American Institute of Physics: New York, pp. 875–887, 1982.
9. Siwek, W. R., Sapoff, M., Goldberg, A., Johnson, H. C., Botting, M., Lonsdorf, R., and Weber, S., A precision temperature standard based on the exactness of fit of thermistor resistance-temperature data using third degree polynomials, In *Temperature: Its Measurement and Control in Science and Industry*, J. F Schooley (ed.), Vol. 6, American Institute of Physics: New York, pp. 491–496, 1992.
10. Chen, C-C, Evaluation of resistance–temperature calibration equations for NTC thermistors, *Measurement*, **42,** 1103–1111, 2009.
11. White, D. R., Some mathematical properties of the ITS-90, In *Temperature: Its Measurement and Control in Science and Industry*, C. W. Meyer (ed.), Vol. 8, American Institute of Physics: New York, 2012.
12. Mangum, B. W., The triple point of succinonitrile and its use in the calibration of thermistor thermometers, *Review of Scientific Instruments*, **54**, 1687–1692, 1983.
13. Beakley, W. R., The design of thermistor thermometers with linear calibration, *Journal of Scientific Instruments*, **28**, 176–179, 1951.
14. Sapoff, M. and Oppenheim, R. M., The design of linear thermistor networks, *IEEE International Convention Record*, **Part 8**, 12–24, 1964.
15. Patranabis, D., Ghosh, S., and Bakshi, C., Linearizing transducer characteristics, *IEEE Transactions on Instrumentation and Measurement*, **IM-37**, 66–69, 1988.
16. Seingold, D. H., *Non-linear Circuits Handbook*, Analog Devices: Norward, MA, 1976.
17. Natarajan, S., Widely Linear temperature-to-frequency converters, *IEEE Transactions on Instrumentation and Measurement* **IM-24**, 235–239, 1975.
18. Trolander, H. W., Case, D. A., and Harruff, R. W., Reproducibility, stability and linearization of thermistor resistance thermometers, In *Temperature: Its Measurement and Control in Science and Industry*, H. H. Plumb (ed.), Vol. 4, Instrument Society of America: Pittsburgh, PA, pp. 997–1009, 1972.
19. Renneberg, C. and Lehmann, T., Analogue circuits for thermistor linearization with Chebyshev-optimal linearity error, *Proceedings of 18th European Conference on Circuit Theory and Design*, Seville, Spain, pp. 910–913, 2007.
20. White, D. R. and Jongenelen, C. L., The immersion characteristics of industrial PRTs, *International Journal of Thermophysics*, **31**, 1685–1695, 2010.
21. Nicholas, J. V. and White, D. R., *Traceable Temperatures*, John Wiley & Sons: Chichester, U.K., 2001.
22. Zurbuchen, J. M. and Case, D. A., Aging phenomena in nickel-manganese oxide thermistors, In *Temperature: Its Measurement and Control in Science and Industry*, J. F. Schooley (ed.), Vol. 5, American Institute of Physics: New York, pp. 889–896, 1982.
23. Strouse, G. F., Meyer, C. W., and Miller, W. W., Drift study of thermistors used between 90°C and 150°C, In *Temperature Its Measurement and Control*, C. W. Meyer (ed.), Vol. 8, American Institute of Physics: New York, 2012.
24. Vaughn, C. D., Gartenhaus, J., and Strouse, G. F., NIST calibration uncertainties of thermistors thermometers over the range from −50°C to 90°C, *Proceedings of NCSL Workshop and Symposium*, Washington, DC, 2005.
25. Siwek, W. R., Sapoff, M., Goldberg, A., Johnson, H. C., Botting, M., Lonsdorf, R., and Weber, S., Stability of NTC thermistors, In *Temperature: Its Measurement and Control in Science and Industry*, J. F. Schooley (ed.), Vol. 6, American Institute of Physics: New York, pp. 497–502, 1992.
26. Wise, J. A., Stability of glass-encapsulated disc-type thermistors, In *Temperature: Its Measurement and Control in Science and Industry*, J. F. Schooley (ed.), Vol. 6, American Institute of Physics: New York, pp. 481–484, 1992.

27. Edwards, T. J., Observations on the stability of thermistors, *Review of Scientific Instruments*, **54**, 613–617, 1983.
28. LaMers, T. H., Zurbuchen, J. M., and Trolander, H., Enhanced stability in precision interchangeable thermistors, In *Temperature: Its Measurement and Control in Science and Industry*, J. F. Schooley (ed.), American Institute of Physics: New York, pp. 865–873, 1982.
29. Wood, S. D., Mangum, B. W., Filliben, J. J., and Tillet, S. B., An investigation of the stability of thermistors, *Journal of Research of NBS*, **83**, 247–263, 1978.
30. DSCC, General specifications for Resistors, Thermal (thermistors) Insulated, MIL-PRF-23648, Jan 23, 2009.
31. Sapoff, M. and Oppenheim, R. M., Theory and application of self-heated thermistors, *Proceedings of IEEE*, **51**, 1292, 1963.
32. Fabian, J., Heating with PTC thermistors, *EDN Products Edition*, **41**(12A), 10, 1996.
33. Michalski, L., Eckersdorf, K., and McGhee, J., *Temperature Measurement*, John Wiley & Sons: Chichester, U.K., 1986.

70

Thermocouple Thermometry

70.1 Introduction **70**-2
70.2 Terminology **70**-2
70.3 Simplest Thermocouple Thermometer **70**-2
70.4 Thermoelectricity **70**-3
Absolute Seebeck Effect • Relative Seebeck Characteristics
70.5 Basic Thermocouple Circuits **70**-5
Thermoelectric Circuit Schematics • Seebeck's Circuit • Dual-Reference-Junction Circuit • Reference Temperature Compensation • Types of Optional Extension • Single-Reference-Junction Circuit • Differential Thermocouple Thermometer
70.6 Functional Model of Thermoelectric Circuits **70**-9
Basic Thermoelectric Circuit Element • Law of Seebeck emf • Corollaries from the Seebeck Law • T/X Sketch
70.7 Applications of the Functional Model **70**-13
Dual-Reference-Junction Circuit • Single-Reference-Junction Circuit • Paralleled Thermoelements
70.8 Thermocouple Characteristics **70**-17
Standardized Thermocouple Types • Color Codes • Thermocouple Selection • Standardized Tolerances • Other Thermocouple Types
70.9 Thermocouple Hardware **70**-23
Material Sources • Quality Assurance • Forms of Thermoelements • Insulation • Connectors • Thermocouple Junction Styles • Specialized Thermocouple Assemblies
70.10 Quality Assurance by Users **70**-28
Thermoelement Identification • Active Tests • Thermocouple Calibration
70.11 Thermocouple Instrumentation Systems **70**-34
Signal Transmission • Thermocouple Instruments
70.12 Thermocouple Thermometry Practice **70**-37
Temperature Ranges • Thermometry Process • Distinctive Thermocouple Noise Problems • Thermocouple Modes of Failure
70.13 Information Sources for Thermoelectric Theory and Thermometry **70**-41
Internet • Standards, Reports, and Books • Technical Papers • Trade Publications • Workshops and Symposia
70.14 Caveat **70**-43
70.15 Summary **70**-43
References **70**-43

R.P. Reed
Proteun Services

70.1 Introduction

The self-generating thermocouple is one of the more widely used (but most misunderstood) sensors for routine and specialty thermometry because of its versatility, simplicity, ease of use, and low cost. It employs the Seebeck thermoelectric effect that, unlike the related Peltier and Thomson thermoelectric effects, alone converts heat to emf [1–4]. T.J. Seebeck first recognized the phenomenon in 1821. However, as a very early pioneer of electrical phenomena, he misunderstood the effect that bears his name. Seebeck monitored the magnetic symptom of current rather than voltage. He never applied it to thermometry. Seebeck and other early users could apply the effect only in closed circuits, rather than in modern current-suppressed "open-circuit" mode. Seebeck's discovery enabled George Ohm to use thermocouples as the essential steady current source to develop his law of resistance in 1826. A. C. Becquerel recognized in 1826 that Seebeck's discovery could be used for pyrometry, but C. S. M. Pouillet, in 1836, was the first to use the Seebeck effect for applied thermometry.

70.2 Terminology

Thermocouple: Any pair of electrically conducting thermoelectrically different material legs in Ohmic contact at a junction is a *thermocouple*. The individual legs are *thermoelements*. Thermoelements can be solid or liquid, normal-, semi-, or superconductors. Seebeck emf occurs only in nonisothermal thermoelements, not junctions.

Junction: Any electrically conducting *interface* between dissimilar thermoelements is a thermoelectric *junction*. Junctions provide electrical connectivity. Junctions contribute no Seebeck emf. Therefore, calibration is of thermoelement segments, not of junctions. Junctions define thermoelement endpoint temperatures and thereby "sense" temperature. Electrical connections between like thermocouple legs are *splices*, not junctions.

Sensitivity: The sensitivity, emf versus temperature, of each individual leg material is expressed by an *absolute Seebeck coefficient*. Net emf between thermally paired dissimilar thermoelements is described by the more familiar *relative Seebeck coefficient*.

Thermocouple states: Segments of thermoelements are *isothermal* where they are at uniform temperature. Multiple junctions are *isothermal* for thermometry while they are acceptably near the same temperature. A thermoelement is *thermoelectrically homogeneous* along any segment where its Seebeck coefficient is uniform. It is *inhomogeneous* for thermometry where spatial nonuniformity of the absolute Seebeck coefficient could significantly degrade thermometry. Significant hidden *inhomogeneity* resulting from use is a commonplace and usually unrecognized problem in practical thermometry and calibration.

70.3 Simplest Thermocouple Thermometer

Correctly understood, thermocouple thermometry is simple and its many significant, distinctive, and obscure problems are easily avoidable. However, from simplistic descriptions of circuits and principles and its elementary applications, a need for deeper understanding may not be apparent. Commonplace concealed problems cause unrecognized thermometry error.

Elementary thermocouple thermometry can seem as easy as is the measurement of electrical quantities by a multimeter. A digital multimeter that can be used with a thermocouple probe alternately to monitor temperature as well as voltage, current, resistance, and other variables is shown in Figure 70.1. This meter, complete with a thermocouple probe of Type **K** material, presently sells as a consumer product in the Electrical Department of some neighborhood hardware stores for the price of a gallon of paint.



FIGURE 70.1 The simplest thermocouple thermometer. Inexpensive multimeters with a thermocouple temperature sensor obscure the need for detailed understanding of the principles of thermoelectricity for modern thermometry circuits. (Note shorted input and unplugged thermocouple. See Section 70.10.3.6.)

This illusion of ultimate operational simplicity may suggest incorrectly that it is unnecessary to understand the essential principles of the diverse thermoelectric circuits now in use and for thermometry.

Critical applications, modern circuits, and inconsistent international thermocouple conventions make it more important than ever to understand the distinctions between electrical and thermoelectric circuits. Thermometry errors of only a few degrees in energy, process, manufacturing, medical, and research fields annually cost millions of dollars in lost yield, fuel cost, performance bonuses, and so on. Consequence of hidden error also occurs as incorrect interpretation of data, failure of objective, equipment damage, personal injury, or even loss of life.

This chapter explains why for realistic thermometry understanding of the authentic principles and conventions is essential. It offers useful tools for recognizing and avoiding commonplace significant obscure measurement errors that are unique to modern practical thermoelectric thermometry.

70.4 Thermoelectricity

The physical cause of thermoelectric phenomena is complex in detail as applied to real materials. It is explained by modern theory confirmed by experiment [1,2]. Seebeck properties of many materials can be predicted. For reliable thermometry, it is not essential to understand the underlying physics. For application, it is sufficient to understand a few very simple but general thermoelectric principles that are fully consistent with physical theory and experiment.

70.4.1 Absolute Seebeck Effect

The absolute Seebeck thermoelectric effect, the occurrence of a net emf between two points of a conductor because their temperatures are different, results from the temperature-dependent distribution of charge carriers [1–5]. Seebeck emf occurs *without* dissimilar materials. It is *not* a junction phenomenon, nor is it the same as Volta's contact potential. It is related to but is not the result of the associated Peltier or Thomson thermoelectric effects that are insignificant in thermocouple thermometry.

To practice and to fully apply the simple principles, it also is not necessary to understand calculus. Nevertheless, the application principles are developed as follows. The *absolute Seebeck coefficient* expresses the measurement sensitivity (volts per unit of temperature) of the Seebeck effect. It is defined over any *thermoelectrically homogeneous* region, x, along a slender conducting material

$$\sigma(T) = \lim_{\Delta T \to 0} \frac{\Delta E}{\Delta T} = \frac{dE}{dT} \tag{70.1}$$

where

ΔE is the net Seebeck emf across the homogeneous segment
ΔT is the temperature difference between ends of that region

The Seebeck coefficient, $\sigma(T)$, is a physical transport property of *all* individual electrically conducting materials.

From Equation 70.1, the net emf is

$$\Delta E = \int_{T_1}^{T_2} \sigma(T) dT = E(T_2) - E(T_1) \tag{70.2}$$

where ΔE is simply the difference of emf between any pair of points of a *homogeneous* segment, separated by any distance, x, between which the temperature difference is $\Delta T = (T_2 - T_1)$. Note that the net Seebeck emf for such a *homogeneous* segment depends only on the temperatures at the two segment endpoints and *not* on the local values of temperature or temperature gradients, dT/dx, between them, nor on the distance between those two points. An overall inhomogeneous thermoelement is composed of locally homogeneous segments. Nevertheless, for general thermometry, each entire thermoelement must be homogeneous so that only endpoint temperatures, not temperature distribution, matter.

The Seebeck coefficient $\sigma(T)$ is a nonlinear function of temperature. For meaningful thermometry, it must be a function of temperature alone, independent of time or location along a thermoelement. Although it is vulnerable to environmental effects, for accurate thermometry, it must not depend on such environmental variables as strain, pressure, or magnetic field during measurement. The coefficient, $\sigma_M(T)$, is an absolute Seebeck *coefficient* for a particular material, **M**. The corresponding *source* voltage within a single material is an *absolute Seebeck emf*, $E_M(T)$. The absolute Seebeck emf does physically exist, but it is *directly* observable over only a limited cryogenic temperature range near 0 K where it can be related to a superconductive thermoelement reference that has, over the range, a negligible Seebeck coefficient. The absolute Seebeck coefficient at higher temperatures is usually determined indirectly [1–5]. The *Thomson coefficient*, $\tau(T)$, of the individual material is measured. The Seebeck coefficient is determined from a thermodynamic Kelvin relationship as follows:

$$\sigma = \int_0^{T_{abs}} \left(\frac{\tau}{T_{abs}} \right) dT \tag{70.3}$$

70.4.2 Relative Seebeck Characteristics

The difference between the Seebeck emfs of two thermoelements, of materials **A** and **B**, with one end of each at temperature, T_m, and the other ends both at temperature, T_r, is their relative *Seebeck emf*, $E_{AB}(T)$, expressed as

$$E_{AB}(T_m, T_r) = E_A(T_m, T_r) - E_B(T_m, T_r) \tag{70.4}$$

The corresponding relative Seebeck coefficient at T_m for a pair is

$$\sigma_{AB}(T_m) = \sigma_A(T_m) - \sigma_B(T_m) \tag{70.5}$$

Note that these relationships apply whether or not the thermoelements are joined at a junction. The thermally paired segments often are remote in a circuit. It is this relative Seebeck coefficient that has been called by the anachronistic term "*thermopower*." (*Thermoelectric power* appropriately refers to electrical power generated using the Seebeck effect.) It is these *relative* voltage or coefficient values that are directly observable and are used in thermoelectric thermometry. These are the values that are presented in the tables of thermocouple emf vs. measuring junction temperature, T_m, referred to a designated reference temperature, T_r [5]. Standardized thermocouple characteristics, by convention, are referenced to $T_r = 0$ °C.

70.5 Basic Thermocouple Circuits

70.5.1 Thermoelectric Circuit Schematics

Conventional *electric* circuit schematics represent passive conductors that often all are of copper. Distinctively, in *thermoelectric* circuits, paired legs are of very different materials and the "*conductors*" are the emf sources. Much more is involved than circuit connectivity. True thermoelectric schematics must explicitly acknowledge each leg material, all junctions, essential relative junction temperatures, and any complementary voltage sources, as these all affect the Seebeck emf and net thermocouple voltage.

70.5.2 Seebeck's Circuit

Inappropriately, the thermocouple is sometimes defined essentially as a closed circuit of two dissimilar thermoelements joined at two junctions, a "hotter" junction and a "colder" junction, as studied by Seebeck. Based on this unbroken circuit, the Seebeck effect has been wrongly characterized as the occurrence of current in the closed loop, as was electromagnetically observed by Seebeck rather than properly as thermally caused emf. The net Seebeck source emf is indeed proportional to the temperature difference between the two junctions and to a relative coefficient for the paired materials, but that emf results in loop current and IR voltage loss in the typically high-resistance thermocouple circuit. That significantly reduces the net voltage. The Seebeck *emf* for accurate thermometry should be measured in "open-circuit" mode that suppresses current by monitoring with an instrument of relatively large input resistance (e.g., greater than 10 kΩ).

In practical thermometry, no applied thermoelectric circuit has only two dissimilar materials or only two junctions. Typical thermometry circuits have many materials and junctions. Several incidental junctions, if not temperature controlled, can result in significant unrecognized spurious Seebeck emf and error. In each of the two following basic thermocouple circuits, a particular electrical circuit connectivity, but with different materials and temperature distributions, results in very different thermoelectric characteristics and error sources.

70.5.3 Dual-Reference-Junction Circuit

A generalized form of the circuit that is most commonplace in practical thermocouple thermometry (Figure 70.2) involves a single *measuring junction* and one essential pair of isothermal temperature *reference junctions*. Most applied thermocouples also have one or more *optional* extensions. Two are shown. It is this circuit that is used in multimeters like Figure 70.1 and in most thermometry. The primary, *measuring*, thermocouple is a pair of thermoelements, **P** and **N**, that are joined at a single *measuring junction*, **a**. Legs **P** and **N** must have different Seebeck coefficients. The coefficients of either or both can be of either positive or negative polarity.

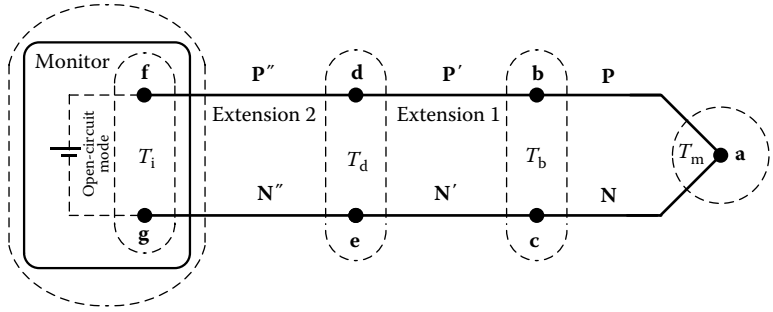


FIGURE 70.2 Dual-reference junction circuit. This most used circuit requires temperature reference to be applied at two junctions at locations in the circuit that depends on the type of optional extensions.

70.5.4 Reference Temperature Compensation

In thermometry, the reference junctions usually are not at the 0 °C reference temperature. Complementary compensation voltage to effect the 0 °C reference temperature is applied internally by the monitor or by an external compensator. (Note that the internal circuitry of the monitoring instrument—and of any compensator—unavoidably is itself a complex thermoelectric network that also contributes an indefinite varying Seebeck voltage unless, after warm-up, it remains stable at sufficiently uniform temperature.)

70.5.5 Types of Optional Extension

In Figure 70.2, pairs **P′N′**, **P″N″** and any additional thermoelement pairs are *optional* circuit extensions. A connector is an extension. Every thermoelement must be homogeneous. Only two junctions may function as the essential reference junctions. Which two depends on the type of extensions? If there are no extensions, **b** and **c** are the reference junctions and are directly connected to the instrument terminals.

In most practical applications, the **PN** measuring thermocouple must be remote from the monitor. In such cases, extension legs **P′** and **N′** (and any additional ones) might be either similar to or very different in Seebeck characteristic from continuous thermoelements **P** and **N** and from each other. Extension leads are of three types: (1) *passive*, (2) *corresponding*, and (3) *compensating*. These types serve different applications. All extension styles are commercially available. Some are proprietary.

70.5.5.1 Passive-Type Extensions

The positive and negative legs of *passive*-type extensions are alike in Seebeck characteristic. As they are alike, they are intended to contribute no significant emf. Usually both legs are of un-plated electrical grade copper, but any conductor could be used. If all extensions are of passive type, reference temperature compensation is applied only at isothermal reference junctions **b** and **c** of thermocouple, **PN**. Compensation must not additionally be applied at the monitor terminals.

As junctions **f** and **g** at the monitor are often remote from junctions **b** and **c** of the measuring thermocouple, they usually are at very different temperatures. With passive extensions, if compensation incorrectly were applied at instrument terminals **f** and **g**, essential Seebeck emf from the temperature interval between T_b and T_i would be missing causing substantial error.

Passive leads often are used with the dual-reference junction circuit in precision calibration. An accurate 0°C reference temperature can be applied to reference junctions **b** and **c** by a fixed-point physical temperature reference such as a precision ice bath. Inappropriate emf contributions of other thermoelements are eliminated. For passive extensions, the temperature value at incidental junctions, such as **d** and **e**, does not matter *if* they are isothermal.

Two other kinds of extension leads are *corresponding* and *compensating* types. They are used where compensating voltage is to be provided by the monitor. For these two styles, the paired extension thermoelements are essential contributing sources of emf. Commercial corresponding- or compensating-style leads, *as a pair*, are of the same nominal thermocouple characteristic as the principal thermocouple but usually they are uncalibrated.

70.5.5.2 Corresponding-Type Extensions

Each leg of *corresponding-type* extensions and connectors for base-metal thermocouples has an absolute Seebeck coefficient of the same nominal type as the corresponding leg of the thermocouple. Corresponding-type extensions contribute emf. They are used with base-metal thermocouples, particularly those many measuring thermocouple assemblies specialized for diverse applications. Extensions can be of thermocouple-grade material. If of extension grade, they might have a broader tolerance and match over a narrower temperature range than the **PN** thermocouple. Corresponding extensions usually are not calibrated. If all extensions are of the *corresponding* type, reference junctions must be at terminals **f** and **g**. They must be isothermal. Extension leads for refractory or precious metal thermocouples are of compensating style rather than corresponding type.

70.5.5.3 Compensating-Type Extensions

The third extension type also is intended to contribute emf. This *compensating* style is used for economy and convenience to extend costly precious or refractory metal thermocouples that are necessary for measuring very high temperatures and for some environments.

Compensating-type extensions can have positive legs that are of different Seebeck coefficient than positive leg **P** and also from other linked positive extension legs. Likewise, negative extension legs are all not like **N**. Rather, each paired *compensating extension* must have the relative coefficient of each extension, as a pair, approximate the σ_{PN} of the **PN** pair.

Usually, only a short length of costly precious or refractory metal measurement thermocouple material has to be exposed to the adverse environment. Extensions can be less costly, insulated differently from the thermocouple, can add bulky shielding and mechanical protection, and can extend a great distance to a recording facility. Often, compensating extensions are made for much higher temperatures than are corresponding-type extensions. However, accuracy is degraded if extensions are used to the higher allowed temperatures because more of the emf is from the compensating extensions than would be from corresponding extensions limited to lower temperature.

70.5.5.4 Extensions of Mixed Type

One or more passive extension pairs and connectors can be used in series with multiple extensions of the corresponding or of compensating types. All emf-supplying extensions and connectors must be nearer the measuring thermocouple than any passive extension. Mixed corresponding and compensating extension pairs can be connected. However, all passive extensions must follow all emf-supplying extensions and must connect to the monitor. The only reference junctions are of the final pair of the emf-contributing corresponding-type or compensating-type extensions. The monitor must not apply additional compensating voltage if the last extension is of passive style. It is essential that the incidental junctions at each end of every extension be isothermal.

70.5.6 Single-Reference-Junction Circuit

Another practical circuit form, Figure 70.3, combines two separate thermocouples and their connectors. The measuring thermocouple is **PN**. The reference thermocouple is **P′N′**. The reference thermocouple often is a commercially fabricated assembly that includes the several components ending with connector terminals, **d** and **e**. The connector of measuring thermocouple **PN** (or any emf-contributing extensions, as **P″N″**) plugs into the reference at **d** and **e**. For calibration, extensions should not be used between the

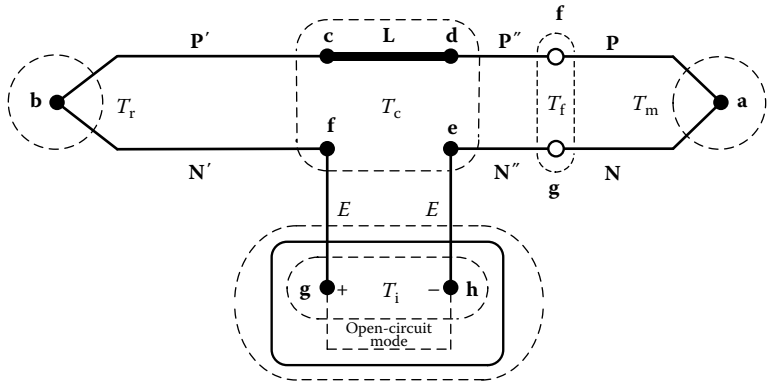


FIGURE 70.3 Single-reference junction circuit. This circuit is often used for convenience in calibration. A reference temperature is physically applied at a single junction.

measuring thermocouple assembly and the reference thermocouple assembly. If extensions are used, they must never be of *passive* type. This circuit is commonly used for convenience in routine calibration laboratories where it is desired to physically impose a controlled and precisely known reference temperature on junction **b**. (For accurate calibration, the circuit of Figure 70.2 with *passive*-style extension is less subject to hidden error, particularly near ambient temperature.)

Typically, in circuits like Figure 70.3, the measuring thermocouple and reference thermocouple are separately acquired thermocouple assemblies. The two thermocouples likely have significantly different Seebeck characteristics even if they are nominally of the same thermocouple type. The calibration of the reference thermocouple initially should be confirmed around room temperature but often it is not. For T_m near ambient temperature, much of the emf is from the reference thermocouple—not from the measuring thermocouple. Therefore, reported calibration of a measuring thermocouple, with contribution from the reference, is misleading at all temperatures. Only if connector link **L** is isothermal, it may be of any material, inhomogeneity, length, or temperature, T_c. The input terminals **g** and **h** of the monitoring instrument may be of any isothermal temperature, T_i. The reference temperature is physically applied and only at **b**. The monitoring instrument *must not* also apply compensation voltage for temperature T_i of the instrument junctions.

70.5.7 Differential Thermocouple Thermometer

The temperature difference between two points often is of interest. The difference emf of a pair of thermocouples of the same type connected in series opposition provides a direct *indication* of temperature difference. However, the difference emf depends nonlinearly on the temperatures. Examples: standard values of emf for the same 100 °C temperature difference of a differential Type *T* thermocouple are

$$T_1 = 20\ °C, \quad T_2 = 120\ °C, \quad \Delta T = 100\ °C, \quad \text{net emf} = 2849.2\ \mu V$$

$$T_1 = 120\ °C, \quad T_2 = 220\ °C, \quad \Delta T = 100\ °C, \quad \text{net emf} = 3204.5\ \mu V$$

$$T_1 = 220\ °C, \quad T_2 = 320\ °C, \quad \Delta T = 100\ °C, \quad \text{net emf} = 3665.3\ \mu V$$

The emfs, therefore the deduced difference temperatures, would significantly disagree. Accurate scaling of emf to temperature would require applying separate nonlinear scaling relations for a known temperature of the thermocouple type. For accuracy, some monitors instead provide dual sets of input

terminals for simultaneous monitoring of a pair of thermocouples of the same type. Either of the measured temperatures is displayed along with the accurate calculated temperature difference.

70.6 Functional Model of Thermoelectric Circuits

The commonplace obscure problems of thermoelectric circuits are difficult to visualize without a graphic tool. Conventional electrical circuit diagrams conceal the actual thermoelectric emf sources, unrecognized thermoelements, incorrect pairings, incidental junctions, and changing functions that often cause significant hidden thermometry error. The spurious "junction-source" model conceals the fact that all thermoelectric problems occur apart from the measuring junction. A general-purpose model of thermoelectric circuits shows what junction temperature structures and relationships must be maintained for reliable thermoelectric thermometry and why.

This nontraditional *Functional Model of Thermoelectric Circuits*, combines (1) a basic thermoelectric circuit element, (2) a single fundamental law that describes the sensitivity of that element, (3) a set of practical corollaries from the law that reveal practical implications, and (4) an informal **T/X** graphic tool for circuit visualization and to aid analysis [2,4,6].

70.6.1 Basic Thermoelectric Circuit Element

A *Seebeck emf cell* (Figure 70.4) is any thermoelectrically homogeneous nonisothermal segment of material **M** of any length within any thermoelement. (An overall inhomogeneous thermoelement is composed of such locally homogeneous segments.) Seebeck emf is dependent on cross section for different metals only if dimensions are less than 50–3500 Å. Each such *homogeneous* segment across which a net temperature difference, ΔT, exists is a *nonideal voltage source* with resistance $R(T)$. (Thermocouple thermometry is rarely of very fast transient temperature change so circuit capacitance and inductance usually are irrelevant.) Therefore, for accurate thermometry, the Seebeck *source emf* must be observed in an "open-circuit" mode (i.e., current effectively nulled by high input resistance.) Any *IR* voltage drop due to current through the resistive thermoelements allowed by too-low input resistance of the voltage monitor lowers the net voltage from the source emf and causes significant error unless compensated.

The emf from each Seebeck Cell was, from Equation 70.2, $\Delta E_M(T) = E_M(T_2) - E_M(T_1)$.

Simply, the net Seebeck emf of the cell can be determined as the difference between the tabulated emf value for the point at $\mathbf{T_2}$ and that value for the point at $\mathbf{T_1}$, both referenced to $\mathbf{T_r}$. Any segments

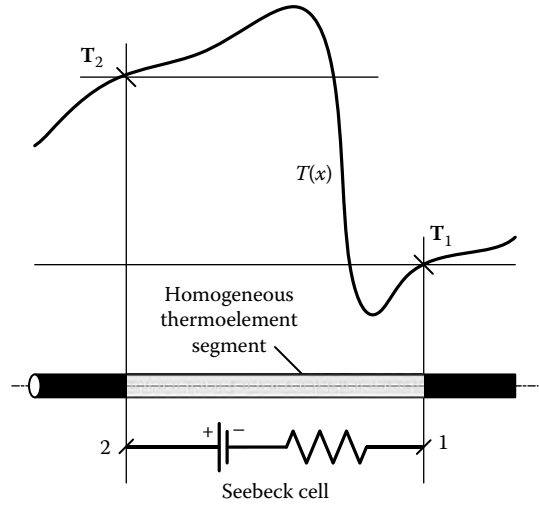


FIGURE 70.4 The Seebeck cell. All thermoelectric emf is produced along thermoelements.

of materials **A** and **B** that happen momentarily to share a pair of endpoint temperatures are *thermally paired*, regardless of their location, proximity, or connectivity in a circuit. The net relative emf for such a pair is $\Delta E_{AB}(T) = E_{AB}(T_2) - E_{AB}(T_1)$ from Equation 70.4. For calculation alone, they need not even be electrically joined.

Either *relative* properties of thermally paired thermoelement segments or *absolute* properties of individuals can be used in analysis. Seebeck properties of a pair, each relative to a third shared reference material, such as standardized Pt-67, also can be used.

Either the Seebeck coefficient, or the sense of the temperature increment across a segment, or both of them, can be either positive or negative. Therefore, unlike an electrochemical cell, the polarity of a Seebeck cell within a circuit depends *momentarily* both on the material and on the sense of temperature difference across the segment. Cell polarity—and even its function—change with temperature distribution.

70.6.2 Law of Seebeck emf

Equation 70.1, $dE = \sigma(T)dT$, is invoked by all quantitative descriptions of the Seebeck effect, whether the description attributes emf to thermoelements or wrongly to junctions. All, in effect, agree that it is the only functional law that governs the Seebeck emf. In this Functional Model, it is named the sole *Law of Seebeck emf*. The Functional Model recognizes that every thermoelectric aspect of circuit behavior follows from this one simple relation.

Note: If this simple relationship does not apply, then accurate thermometry is not possible.

70.6.3 Corollaries from the Seebeck Law

Corollaries express useful and practical truths and implications that may not be obvious from a concisely stated law. Here, they encourage a distinctive way to view real thermoelectric circuits. From the Law of Seebeck emf, five particularly instructive practical corollaries are recognized. The Functional Model corollaries are stated in Table 70.1. These are corollaries of (1) *functional roles*, (2) *functional determinacy*, (3) *temperature determinacy*, (4) *emf determinacy*, and (5) *Seebeck emf*. They are to be comprehended, not memorized. These corollaries relate more directly to practical thermoelectric thermometry than do the commonplace archaic statements of three thermocouple "laws" that actually also are only corollaries of the sole Seebeck Law, not "laws" themselves [2,4]. The significance of the five functional corollaries is illustrated by a nonconventional **T/X** sketch.

70.6.4 T/X Sketch

An informal **T/X** sketch, Figure 70.5, plots one present distribution of *temperatures*, **T**, of all junctions against the *sequence*, **X**, in which those junctions occur in the circuit. (**X** is not a space variable; slopes do not represent temperature gradients.) A **T/X** sketch is not rendered to scale. It is not for graphic calculation. It is only for visualization of momentary functions and consequences.

The **T/X** sketch clearly illustrates how thermoelements, not junctions, function as the physical Seebeck sources. This nontraditional **T/X** sketch is one proper form of thermoelectric schematic that necessarily acknowledges (1) electrical connectivity, (2) junction relative temperatures, (3) leg materials, and (4) any applied reference temperature compensation. The sketch forces attention to problem areas that are not apparent from a conventional electrical schematic nor from alternative traditional **E**(**T**), **E**(**x**), or **T**(**x**) plots. It directs attention to inherently hidden errors so that they can be recognized and avoided.

Note: The practical **T/X** sketch requires no more information than is essential for successful thermometry. If the relative temperatures of incidental junctions are not known and properly controlled, then thermoelectric thermometry is of indefinite accuracy regardless of any perceived calibration.

TABLE 70.1 Corollaries from the Law of Seebeck emf

In any circuit of electrically conducting dissimilar material segments that have absolute Seebeck coefficients $\sigma(T)$, that are each thermoelectrically homogeneous, and each of which follow the Seebeck Law,

$$dE(T) = \sigma(T)dT$$

1. *The corollary of functional roles*

   In a thermoelectric circuit, for any temperature distribution, there are three momentary thermoelectric functional roles: Seebeck emf sources, "conductors," and junctions:

   a. Real thermoelectric *junctions* are interfaces that ohmically connect dissimilar materials
   b. "*Conductors*" are thermoelement segments that, *in effect*, individually or in combination, contribute no net Seebeck emf
   c. All other segments are *sources* of Seebeck emf

2. *The corollary of functional determinacy*

   Thermoelectric roles of segments cannot be *predetermined* by physical construction, material choice, or circuit arrangement alone; they are *determined* by temperature distribution.

3. *The corollary of temperature determinacy*

   In a thermoelectric circuit, only a single junction temperature can be *determined* from the net emf and only if the individual or relative temperatures of all other real junctions are known.

4. *The corollary of emf determinacy*

   Seebeck emf occurs only in thermoelements, but the circuit net Seebeck emf is *determined* by the temperatures of only real junctions.

5. *The corollary of Seebeck emf*

   The Seebeck emf of any thermoelement segment of material, **M**, of any cross-section area,[a] with endpoint temperatures, $\mathbf{T}_1$ and $\mathbf{T}_2$, is independent of *temperature distribution* or of *temperature gradients*, dT/dx, between those endpoints because it is determined only by

$$E_M(T) = \int_{T_1(X_1)}^{T_2(X_2)} \sigma(T)dT = E_M(T_2(X_2)) - E_M(T_1(X_1))$$

[a] Seebeck characteristics of different metals depend on dimension below thresholds of, 30 to −3500 Å.

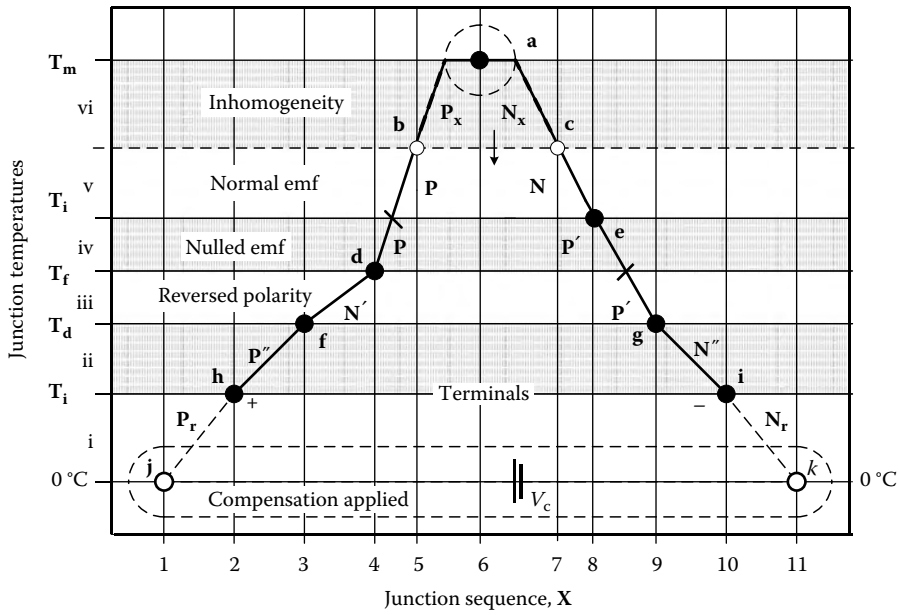


FIGURE 70.5 **T/X** sketch of dual-reference junction circuit. Illustration of sources of error that is due to improper temperature distribution at incidental junctions.

70.6.4.1 Junctions

Filled circles represent real physical junctions between dissimilar thermoelements. *Measuring junctions* are those of which an unknown temperature is to be determined. *Reference junctions* are any of which the temperature must be accurately known and controlled for thermometry. Other real junctions are *incidental junctions*. For incidental junctions, relative temperatures must be controlled to avoid error.

Open circles importantly represent connections only *presumed* to be *splice connections* between identical thermoelements. (Consequentially, these joins, unrecognized, might actually be real junctions between merely *similar* thermoelements or between thermoelements unintentionally misconnected in reversed polarity.) Open circles can also depict indefinite invisible boundaries between initially homogeneous segments and any local segments made inhomogeneous along thermoelements damaged in use. These might not be visible in the physical circuit though effectively they are real junctions.

70.6.4.2 Thermoelements

Thermoelements between the junctions may be shown simply as solid straight lines. Optionally, a conventional electrical cell symbol with indicated polarity and a resistor symbol for segments in each temperature zone may be shown on each leg for clarity, as in some electrical schematics. Usually, these distinctly thermoelectric electrical functions are just understood. Circuit analysis is the same as for regular electrical circuits except that, momentarily, both thermoelectric cell emfs and their polarities are temperature-determined. If numerical estimation of plausible magnitudes of errors is wanted, both the *actual* and *relative* temperatures of incidental junctions must be estimated.

70.6.4.3 Isotherms

Lines are drawn across the sketch at the temperature of every real junction. Isotherms through two adjacent junction temperatures bound a temperature zone that functionally pairs thermoelements. Intercepts of isotherms where they cross thermoelements apart from real junctions may be indicated by *tics* that identify segment ends that can be viewed as nonfunctional *virtual junctions* used only for visualization and calculation.

In a strictly series thermoelectric circuit, all thermoelement segments that span a temperature interval between adjacent isotherms occur in pairs. Each zone contributes to the net Seebeck emf. In such circuits, each temperature zone contributes a specific net *relative* Seebeck emf. Any temperature zone might include paired segments of several different materials. For a temperature zone defined by two temperature-adjacent junctions (e.g., **d** and **e** in Figure 70.5), both the sign and the magnitude of the relative emf as well as which materials are paired depend just on whether the temperature of **d** is greater than or less than that of **e**.

70.6.4.4 Reference Temperature Compensation

Complementary nonthermoelectric voltage sources, $\mathbf{V}_c$, are indicated as dashed equivalent *virtual thermoelements*. Thermocouple monitors or else separate modules might apply such reference junction temperature compensation voltage. It is important to show any compensation voltage source on the **T/X** sketch. Some circuits require that compensation *not* be applied. If applied redundantly or at the wrong junctions, such complementary voltage results in substantial error.

70.6.4.5 Evaluation of emf

Beyond visualization, optionally, the **T/X** sketch can support calculation of zone and circuit emfs. When available, the *relative* Seebeck coefficients of thermally paired materials can be used for direct evaluation of net emf, zone by zone. The values of $E_{AB}(T)$ for each of the thermally paired cells are obtained from the standard thermocouple polynomial defining equations, simpler representations of those equations, tables, or graphs. Alternatively, the emfs of individual standardized thermoelements relative to a particular Pt-67

platinum reference, also tabulated in the NIST defining document, can be used [7]. *Absolute* Seebeck properties for many nonstandardized materials also are known, but are less commonly reported [4].

The relative Seebeck emf from thermoelement segment pairs in all temperature zones is summed from the reference temperature, 0 °C. Like thermoelements that cross the same temperature zone in opposite temperature directions as the circuit is traversed contribute no *net* emf (Corollary 1). They cancel. Any such opposed pairs can immediately be eliminated from analysis by inspection as passive "conductors" that, momentarily, are contributing no net emf. Alternatively, the *absolute* Seebeck emfs could be summed as the **T/X** circuit is traversed on the sketch from one instrument terminal (e.g., the negative), to the other terminal. The two approaches can be intermixed.

70.7 Applications of the Functional Model

A representative **T/X** sketch, Figure 70.5, for a dual-reference junction thermocouple with two extensions, Figure 70.2, illustrates **T/X** application. It shows one deliberately inappropriate yet plausible temperature distribution. It reveals how particular segments of thermoelements, variably, are thermally paired in electrical effect, with different segments that may be remote in the circuit. It also illustrates the commonplace profound effect of unrecognized inhomogeneity in thermometry. The **T/X** sketch, applied here to the two basic thermocouple circuits, depicts the significant circuit elements (the junctions and thermoelements) in a way that focuses on their temperature-determined thermoelectric functions (Corollaries 1 and 2). The **T/X** sketch illustrates realistic, commonplace, yet usually unrecognized, problems of circuits used in thermometry. For illustration, circuits are here depicted with plausible defects or with improperly controlled incidental junction *relative* temperatures. Each kind of error source represented here has been recognized as significant in practical application and even in calibration performed by qualified calibration laboratories. The errors, if not detected, would have significantly degraded thermometry and invalidated costly calibrations. Viewed from the perspective of the spurious junction-source model, the thermal response of such circuits sometimes has been very puzzling.

Note: Though usually hidden, most such commonplace problems are easily recognized, diagnosed, and avoided if thermoelectric principles are understood and applied.

70.7.1 Dual-Reference-Junction Circuit

Figure 70.5 is a **T/X** sketch for an initially homogeneous *then use-degraded* dual-reference junction thermocouple, Figure 70.2, with measuring thermocouple **PN**, junctions **d** and **e**, and two extensions **P′N′** and **P″N″**. This single **T/X** sketch illustrates eight of the commonplace obscure sources of error additional to any calibration, instrument, or thermal coupling error. One or more of these, or other such errors, can occur in any installation. Temperature Zones **i** through **vi** each contribute essential Seebeck emf. Zone **ii** shows the *corresponding*-type extension in appropriate polarity, but Zone **iii** has the extension, **N′P′**, in reversed polarity. (Zone **ii** might have been improperly of a passive-type extension.) Zone **iv** thermally pairs very *similar* segments, **P** and **P′**, inappropriately contributing too little emf. Note that if $\mathbf{T_d}$ is greater than $\mathbf{T_e}$, the emf from Zone **iv** is from a very different thermoelement pair.

70.7.1.1 Inhomogeneous Segments

Troublesome inhomogeneity most often occurs around a severely heated measuring junction. Severe damage, experienced as drift, often is not visible on the thermoelements nor recognized as inhomogeneity in thermometry. Figure 70.5 shows measuring junction, **a**, with adjoining segments of thermoelements appropriately immersed within an isothermal region (as in calibration.) However, heat-damaged segments, $\mathbf{P_x}$ and $\mathbf{N_x}$, extend from junction **a** to phantom real junctions **b** and **c** that bound the damage. This portion of the inhomogeneity, as in calibration, lies partially in the isothermal zone at junction **a**, where it contributes no emf, but it extends to a location of much lower temperature, $\mathbf{T_x}$. (Realistically, the degradation would affect the dissimilar thermoelements differently so **b** and **c** would not likely be isothermal.)

In thermometry, this damage usually spans the highest temperature zone where the effect of its modified emf is greatest in thermometry. (When inhomogeneity is hidden isothermally, recalibration is invalidated.)

70.7.1.2 Electrical Shorting

Note that mechanical damage apart from the junction could incidentally result in a progressive contact short junction between legs, as directly between **b** and **c**. That would result in uninterrupted temperature indication from a different location along the thermoelements and probably at a different local temperature than the perceived measuring junction with no definite warning of serious thermocouple failure.

70.7.1.3 Compensation Scaling Error

This **T/X** sketch illustrates the contribution of reference junction temperature compensation, when needed, by virtual thermoelements $\mathbf{P}_r$ and $\mathbf{N}_r$. The compensation characteristic, which only approximates the standard Seebeck characteristic, is at least slightly different from the one used for scaling the thermocouple emf. Scaling error in this temperature zone can be substantial yet remain unnoticed. That reference affects scaling at all temperatures. Some approximations initially are, or else progressively become, inaccurate.

70.7.1.4 Inappropriate Reference Temperature Compensation

Compensation error is usually unrecognized even if it is significant and variable. It can occur (1) due to lack of needed compensation, (2) as redundant compensation at more than one set of reference junctions, (3) as compensation applied to junctions inappropriate for the extension style, or (4) simply of incorrect value. A similar magnitude of error occurs whether no compensation is applied or whether proper compensation is applied, but redundantly at two sets of reference junctions.

70.7.1.5 Compensation at the Monitor

Typical thermocouple monitors apply compensation voltage at input terminal junctions **h** and **i**. That reference temperature compensation feature should be selectable because compensation sometimes must *not* be applied there. Compensation must be at the monitor terminals for the basic measuring thermocouple **PN** alone or else with complete *corresponding*- or *compensating*-type extensions, **P′N′**, and so on. It must *not* be at the monitor if passive extensions are used.

70.7.1.6 Compensation Apart from the Monitor

If *all* extensions are of the *passive* style, reference temperature compensation must be applied only at terminal junctions **d** and **e** external to the monitor and not also or alternatively at the monitor. If $\mathbf{T}_i$ and $\mathbf{T}_d$ are similar and compensation is improperly applied at the monitor, the error might be small but significant and variable and remain unrecognized. However, if the transition from the measuring thermocouple to extensions is remote from the monitor, those junction temperatures could be very different and the error significant but inconspicuous relative to the indicated temperature.

70.7.1.7 Compensation with Multiple Extensions

Each extension interface risks introduction of thermoelements that might accidentally be of slightly or very different material types, of incorrect style, or else be connected in crossed polarity as by misinterpretation of color codes. Each extension connection introduces incidental junctions. In such instances, the compensation could be misplaced at the measuring thermocouple, at the monitor, between a *corresponding*-style extension and a *passive*-style extension, and so on.

70.7.1.8 Reversed Polarity

Error resulting from polarity reversal of extensions may not be recognized even if it is significant. Misunderstood or contradictory polarity color codes risk accidentally connecting extensions in reversed polarity at **d** and **e** or at **f** and **g** so that the extension, unnoticed, effectively becomes relatively an **N′P′**pairing in the circuit, Zone **iii**. If temperatures of incidental extension junctions are different,

polarity reversal at either end of extension **P′N′** introduces error. Also if an obvious polarity error is noticed in temperature indication and polarity then is merely reversed at the monitor or measuring thermocouple, a different error is introduced. If dissimilar thermoelements are spliced, hidden within insulators or connector shells, along the circuit, the hidden incidental junction can be found with a swept "temperature spike" test (see Section 70.10.2.3.2).

70.7.2 Single-Reference-Junction Circuit

This commonplace circuit, Figure 70.3, is widely used for convenience in routine calibration of a measuring thermocouple, **PN**, as well as for thermometry. Represented by a **T/X** sketch, Figure 70.6, plausible error sources are introduced for illustration. The monitor *must not* apply electronic compensation for this circuit inappropriately as is shown in the sketch.

70.7.2.1 Reference Thermocouple

The reference thermocouple, **P′N′**, of the same nominal type as the measuring thermocouple, usually is separately purchased as an assembly with thermocouple connectors and *passive*-type extension leads, **E**, that contribute no net emf in Zone **iv** as they are of the same type. Except for use in thermocouple type-characterization, they are of a type different from **N**. Usually they are of copper. The measuring thermocouple, **PN**, with optional extension **P″N″**, plugs into the jack of the reference thermocouple assembly at points **d** and **e**.

Reference junction, **b**, should be held accurately at reference temperature, 0°C but is shown at slightly higher temperature. However, as this reference thermocouple is separately fabricated and purchased, reference junction thermocouple, **P′N′**, has a Seebeck characteristic that, even if within tolerance, is different from the **PN** thermocouple. Typically, it is not individually calibrated. As $\mathbf{T_b}$ and $\mathbf{T_f}$ are very different, when the measuring junction temperature is near ambient, all or most of an observed "calibration" emf is from the uncalibrated reference thermocouple even with proper isothermal temperature control of link **L**. Calibration of **PN** is compromised at all temperatures. This systematic reference error source should be acknowledged in a calibration report, but it can neither be eliminated nor corrected.

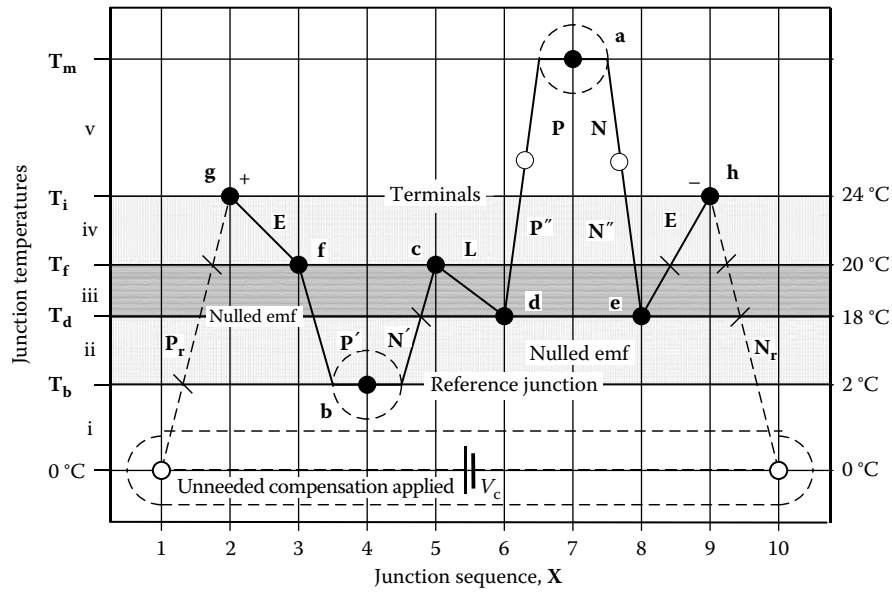


FIGURE 70.6 **T/X** sketch of single-reference junction circuit. Measuring junctions and reference junctions are usually sold as separate assemblies.

70.7.2.2 Incidental Junctions

The connector components, ideally isothermal in Zone **iii**, and the monitor input terminals separately are to be isothermal at temperatures that should both be near ambient temperature. Just as in the dual-reference junction circuit, the temperatures of incidental junctions must be controlled. For illustration of error source, they are shown uncontrolled in the **T/X** sketch (Figure 70.6). Such incidental errors often are not evident in calibration or in temperature indications, even if they are significant.

Fortunately, most of the hidden error sources illustrated in Figures 70.5 and 70.6 are very easily avoided, but only if recognized. Avoidance of most errors by isothermal control of incidental junctions using passive thermal lagging and insulation usually is very easy. A practical significance of the **T/X** schematic is that it forces the attention to specific incidental junctions of which temperatures must be controlled to avoid hidden error. It also guides estimate of plausible error.

70.7.3 Paralleled Thermoelements

In some situations, two or more thermoelements might accidentally be paralleled by electrical contact at widely separate points that are at different temperatures as by multiple shorting of a shield to thermoelements or between multiple thermoelements within a mineral-insulated, metal-sheathed (MIMS) thermocouple. Alternatively, by design, legs of dissimilar paralleled thermoelements have been used as a compound leg, intending to make a leg of customizable relative Seebeck coefficient.

Figure 70.7 shows a circuit with a negative leg paralleled to form a loop by **T/X** sketch. The overall circuit is monitored in the usual "open-circuit" mode, but it includes a closed loop that locally results in current that affects the net voltage. The net loop voltage is affected by temperature-dependent resistance, therefore, by wire diameter and thermoelement lengths. Thus, the effective Seebeck coefficient of the paralleled leg is not a function of junction temperature alone as it must be for thermometry. This is a source of error. The paralleled leg circuit is inappropriate for accurate thermometry but occurs as a defect.

Just for illustration, let thermoelement legs **A** and **B** both be of Type **TP** and leg **C** of Type **TN**. Values for these thermoelements relative to Pt-67 are tabulated by NIST. Resistances of 3 Ft., 22 AWG, legs **A** and **B** are 0.0483 Ω. Leg **C** resistance is, 1.728 Ω. Between junction temperatures 20 and 300 °C, relative to a shared Pt-67 reference, Seebeck emfs for the TP legs would be 3,024.0 μV. Leg **C** alone would contribute 11,048.4 μV. The Seebeck emf would result in a loop current of 4.517 mA reducing the net

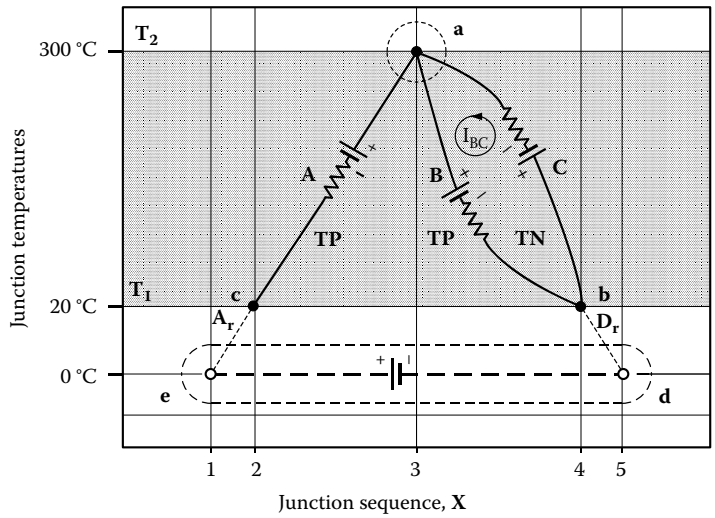


FIGURE 70.7 **T/X** sketch of a paralleled-leg circuit. Thermoelements, paralleled accidentally or by design, are inappropriate for thermometry.

loop voltage to 3242.1 μV. As a normal "open-circuit" **TP/TN** thermocouple, an **AB** pair alone would contribute 14,072.4 μV. Seebeck emf-induced current has reduced the net emf voltage by 10,830.3 μV.

Note: The BC loop is a closed-circuit "thermocouple" as employed by Seebeck.

70.8 Thermocouple Characteristics

70.8.1 Standardized Thermocouple Types

In the United States, nine material combinations presently are letter-designated by the American Society for Testing and Materials (ASTM) as thermocouple types: **B**, **C**, **E**, **J**, **K**, **N**, **R**, **S**, and **T** [7,8]. Representative properties of letter-designated thermocouples are summarized in Table 70.2. They are specialized for use in different temperature ranges and environments. For U.S. standards, Seebeck characteristics are experimentally determined by the National Institute of Standards and Technology (NIST) from samples of the same nominal thermoelement material type provided by several manufacturers. Then, the independent standardizing committee, ASTM, with volunteer members representing manufacturers, calibrators, NIST, and users, prepares a consensus standardizing document based on the NIST data. The standards committee adds commercial tolerances and application recommendations. The standards are only sanctioned by the American National Standards Institute (ANSI).

From measurements at intrinsic standard reference temperatures, NIST defines the Seebeck characteristics, emfs, and coefficients, by $E(T)$ polynomial functions of degrees between 8 and 14, rather than by the tabulated values. NIST then derives the corresponding inverse $T(E)$ values, determined from $E(T)$ by iteration. For application convenience, NIST also presents polynomials of lower degree that approximate the actual characteristics over narrower temperature ranges. Properties of most letter-designated types are published in NIST Monograph 175. The tables are also available at www.nist.gov.

Significantly, it is the relative Seebeck *characteristics* rather than the material *compositions* or alloy *names* that are standardized. The specific alloy composition and metallurgical treatments of actual thermoelements may be varied, as for improved stability, so long as the pair initially conforms to the standard within tolerance. For example, the nominal alloy composition used for **EN** and **TN** legs is slightly different from **JN** thermoelements that share the same alloy name ("Constantan"). Therefore, individual thermoelements should be used as selectively paired for thermocouples by the producer. Randomly paired thermoelements of the same letter-designated type from the same or different manufacturers or batches are not assured to conform to neither standardized paired characteristics nor tolerances. Any thermocouple exposed to excessive high temperature for durations longer than a few weeks might progressively exceed the initial tolerance. (Such thermocouple "drift," incorrectly viewed as global "de-calibration," is from progressing local inhomogeneity.)

The letter designation is complemented by suffixes. A first suffix **P** or **N**, as in thermoelement types **KP** or **KN** of a thermocouple Type **K** pair, distinguishes positive from negative thermoelements. An added final **X** suffix distinguishes *corresponding-type* extension-grade wire material, as in **KPX** for a positive Type **K** extension thermoelement or **KX** for an extension-grade pair. (Some manufacturers distinguish *compensating*-type extensions by a nonstandardized **C** final suffix.) Thermoelement extension-grade wire has the same nominal Seebeck coefficient as the nominal thermocouple type over some temperature range around ambient temperature but has a broader standardized tolerance than thermocouple-grade material.

70.8.2 Color Codes

Both thermocouple letter type and thermoelement polarity should be visually distinguishable. Seebeck *emfs*, *coefficients*, *letter designations*, and *tolerances* of the thermocouple types as standardized in the United States are essentially the same in many international standards [8]. Therefore, it is now functionally possible, *but very inadvisable*, to intermix in a circuit thermoelements of a particular type made to different standards and marked with multiple conflicting color codes. Unfortunately, the *color codes* used to identify these few types in different standards are inconsistent in every possible way. The unfortunate Babel of national color codes that presently exists is displayed in Table 70.3.

TABLE 70.2 Characteristics of the U.S. Letter-Designated Thermocouples

Type	Common Name	Resistivity (mΩ·cm)	MP (°C)	Recommended Range (°C)	emf at 400 °C (mV)	Initial Tolerance (Norm/Special)
B	—	34.4	1810	870–1700	0.787	0.25%
BX	—	—	—	—	—	0.50%
BP	Pt30Rh	18.6	1910	—	—	—
BN	Pt6Rh	15.8	1810	—	—	—
C	W5Re	—	—	0–2315	6.731	4.40%
	W26Re	—	—	—	—	0.11 mV
CPX	—	—	—	—	—	—
CNX	—	—	—	—	—	—
E	—	127	1270	−200 to 870	28.946	1.0 °C or 0.40%
EX	—	—	—	—	—	1.7 °C or 0.50%
EP	Chromel®	80	1430	—	—	—
EN	Constantan	46	1270	—	—	—
J	—	56	1270	0–760	21.848	1.1 °C or 0.40%
JX	—	—	—	—	—	2.2 °C or 0.75%
JP	Iron	10	1536	—	—	—
JN	Constantan	46	1270	—	—	—
K	—	112	1400	−200 to 1260	16.397	1.1 °C or 0.40%
KX	—	—	—	—	—	2.2 °C or 0.75%
KP	Chromel	80	1430	—	—	—
KN	Alumel®	31	1400	—	—	—
N	—	—	—	0–1260	12.974	1.1 °C or 0.40%
NX	—	—	—	—	—	2.2 °C or 0.75%
NP	Nisil	—	—	—	—	—
NN	Nicrosil	—	—	—	—	—
R	—	29	1769	0–1480	3.408	0.6 °C or 0.10%
RX	—	—	—	—	—	1.5 ° C or 0.25%
RP	Pt13Rh	19	1840	—	—	—
RN	Pt	10	1769	—	—	—
S	—	30	1769	0–1480	3.259	0.6 °C or 0.10%
SX	—	—	—	—	—	1.5 °C or 0.25%
SP	Pt10Rh	20	1830	—	—	—
SN	Pt	10	1769	—	—	—
T	—	48	1083	−200 to 370	20.81	0.5 °C or 0.40%
TX	—	—	—	—	—	1.0 °C or 0.75%
TP	Copper	2	1083	—	—	—
TN	Constantan	46	1270	—	—	—

Source: Park, R.M. (ed.), *Manual on the Use of Thermocouples in Temperature Measurement*, 4th edn. or later, MNL12, American Society for Testing and Materials, Philadelphia, PA, 1993; *Annual Book of ASTM Standards, Temperature Measurement*, Vol. 14.03, American Society for Testing and Materials, Philadelphia, PA, 2013.

Chromel and Alumel are trademarks of Hoskins Mfg. Co.

Initial tolerances are for material as manufactured and used within recommended temperature range protected in a benign environment for moderate durations.

TABLE 70.3 International Color Codes for Letter-Designated Thermocouples

Std. No.	ASTM E230-03	IEC 584-3	BS 1843	DIN 43714	JIS 1620	NMI	NFC42 42-323
Letter	United States	Euro	England	Germany	Japan	China	France
B	Brown	—	—	—	—	—	—
BX	Gray	—	—	Gray	Gray	—	—
BP	Gray	—	—	Red	Red	—	—
BN	Red	—	—	Gray	Gray	—	—
C	Brown	—	—	—	—	—	—
CX	Red	—	—	—	—	—	—
CP	Green	—	—	—	—	—	—
CN	Red	—	—	—	—	—	—
E	Brown	—	—	—	—	—	—
EX	Purple	Purple	Brown	Black	Purple		Purple
EP	Purple	Purple	Brown	Red	Red	Red	Yellow
EN	Red	White	Blue	Black	White	Brown	Purple
J	Brown	—	Red	—	—	—	—
JX	Black	Black	Black	Blue	Yellow		Black
JP	White	Black	Yellow	Red	Red	Red	Yellow
JN	Red	White	Blue	Blue	White	Purple	Black
K	Brown	—	—	—	—	—	—
KX	Yellow	Green	Red	Green	Blue	—	Yellow
KP	Yellow	Green	Brown	Red	Red	Red	Yellow
KN	Red	White	Blue	Green	White	Blue	Purple
N	Brown	—	—	—	—	—	—
NX	Orange	—	—	—	—	—	—
NP	Orange	—	—	—	—	—	—
NN	Red	—	—	—	—	—	—
R	Brown	—	—	—	—	—	—
RX	Black	Orange	Green	White	Black	—	Green
RP	Black	Orange	White	Red	Red	Red	Yellow
RN	Red	White	Blue	White	White	Green	Green
S	Brown	—	—	—	—	—	—
SX	Black	Orange	Green	White	Black	—	Green
SP	Black	Orange	White	Red	Red	Red	Yellow
SN	Red	White	Blue	White	White	Green	Green
T	Brown	—	—	—	—	—	—
TX	Blue	Brown	Blue	Brown	Brown	—	Blue
TP	Blue	Brown	White	Red	Red	Red	Yellow
TN	Red	White	Blue	Brown	White	White	Blue

Source: Park, R.M. (ed.), *Manual on the Use of Thermocouples in Temperature Measurement*, 4th edn. or later, MNL12, American Society for Testing and Materials, Philadelphia, PA, 1993; *Annual Book of ASTM Standards, Temperature Measurement*, Vol. 14.03, American Society for Testing and Materials, Philadelphia, PA, 2013.

American-standardized color codes used in the United States and Canada have remained fixed for many decades. Therefore, color code confusion of material types in the United States is mostly due to user error, carelessness, or, infrequently, to flawed coding or labeling in manufacture.

Now, from a single vendor, thermocouple materials are available made to several of the conflicting color codes. The disagreement between color codes could present a costly problem of misidentification, as of polarity. Circuits with multiple extension cables risk crossed-polarity error that may be hidden in thermometry. In the U.S. thermocouple standards, uniquely, the *negative* thermoelement is always coded as *red* (unfortunately, contrary to other common U.S. electrical and instrument practices). This is also contrary to historic national thermocouple color codes of China, Germany, and Japan, in which *red always* designated the *positive* thermoelement. The Japanese standard denotes thermocouple type only by overall jacket color with positive leg always red and negative always white. By U.K. standard, the *negative leg* of all letter-designated types is *blue*. In France, the *positive* thermoelement is always yellow, but both Types **E** and **K** have the same yellow/purple wire in the French code with type distinguished by overall jacket color. In English standards, yellow denotes only the positive Type **JP** material. In the U.S. standards, yellow marks only the **KP** thermoelement and Type **K** thermocouple material. The copper wire leads of some U.S. electric blasting caps use yellow insulated wire with a faint narrow parallel red tracer stripe that, in field use, has been confused with Type **KX** thermocouple extension wire.

The outer jacket of U.S. thermocouples is brown. Extension-grade material is distinguished by a jacket of the color of the positive thermoelement. An IEC standard provides for "intrinsic-safing" of instrumentation that must reside within a hazardous environment but be electrically isolated by electronic barriers from separated monitoring areas. (Intrinsic safety is not addressed in U.S. thermocouple standards.) For uniform distinction of the intrinsic-safe class of application, even the low-voltage thermocouple circuits uniformly must be distinguished by a blue outer jacket. Special materials and electrical parameters may also be specified. Nevertheless, the thermoelectric characteristics of intrinsically safe and regular thermocouples are functionally identical. The user should confirm the material of each leg.

70.8.3 Thermocouple Selection

Thermocouple type selection is based on temperature range, compatibility with environment, sensitivity, wire size, cost, and user access to compatible indicators.

70.8.3.1 Temperature Range

Base-metal thermocouples cover the range from –200 °C to 1260 °C (Table 70.2). Higher temperatures require more expensive precious or refractory materials that extend the available upper limit to 1700 °C.

70.8.3.2 Environmental Compatibility

Potentially harmful environments are classified as oxidizing, oxygen-poor, neutral, vacuum, reducing, radiation, sulfurous, and so on. Thermocouples and sheaths react adversely to some environments depending on temperature range. A thermocouple type is best suited for one or more of these environments but often is very unsuited to the others. An adverse environment locally introduces inhomogeneity into thermoelements and breaches protective sheaths. Merely conditional recommendations are summarized in Table 70.4. Unfortunately, some environments alternate between incompatible conditions or simultaneously combine them. Thermocouple suppliers should be consulted for material recommendations for especially difficult situations that involve harsh chemical or radiation environments or mixed environments. Trial may be required.

70.8.3.3 Sensitivity

Relatively low emf once was a major limitation of thermocouples. Now, commonplace modern thermocouple monitoring instruments, in "open-circuit" (nulled-current) mode, can routinely indicate temperatures to 0.1°C or even 0.01°C resolution and stability with excellent noise rejection for all

TABLE 70.4 Environmental Tolerance of Letter-Designated Thermocouples

Type	Oxygen Rich	Oxygen Poor	Neutral	Reducing	Hydrogen	Humid	Vacuum	Below 0 °C	Sulfurous	Neutrons
B	Good	Good	Fair	Poor	—	Good	Fair	Poor	Poor	Fair
C	Bad	—	Good	Good	Poor	—	Good	—	—	—
E	Good	Poor	Poor	Poor	—	Good	Poor	Good	Poor	Poor
J	Fair	Good	Good	Good	—	Poor	Good	Poor	Fair	Poor
K	Good	Poor	Poor	Poor	—	Good	Poor	Fair	Poor	Good
N	Good	Fair	Fair	Poor	—	Good	Poor	Good	Fair	Good
R, S	Good	Good	Fair	Poor	—	Good	Poor	Fair	Poor	Poor
T	Fair	Fair	Fair	Good	—	Good	Good	Good	Fair	Poor

Source: Park, R.M. (ed.), *Manual on the Use of Thermocouples in Temperature Measurement*, 4th edn. or later, MNL12, American Society for Testing and Materials, Philadelphia, PA, 1993; *Annual Book of ASTM Standards, Temperature Measurement*, Vol. 14.03, American Society for Testing and Materials, Philadelphia, PA, 2013.

letter-designated thermocouple types. Thermocouple tolerances are much larger. As seen in Table 70.2, the sensitivities of various types are very different. Precious and refractory metals have less than a quarter of the emf of the least sensitive Type **N** base-metal thermocouple.

70.8.3.4 Wire Size

Bare thermoelement material is readily available in diameters from 0.13 mm (0.0005 in.) to 3.3 mm (0.128 in.). Thermoelements of all diameters are sold separately or else matched as a thermocouple pair with standard tolerance. The finer wire sizes have less effect on the temperature being measured but degrade more quickly. Upper temperature limits vary with wire size (Table 70.5). Fine wires are favored for transient thermometry, where space is limited, or for fine spatial resolution. Thermoelement resistance is not a factor in nulled-current thermometry. Some special applications favor thin foil or film thermoelements.

70.8.3.5 Cost

Thermoelement and thermocouple wire is available bare or in a wide variety of insulants ranging from vinyl, Kapton®, and Teflon® to glass or ceramic fiber with metal over-braid. The price varies irregularly with material and diameter. Typically, the lowest price is for bare wire of mid-range diameters between 20 and 28 AWG. Price depends more on diameter than on material. The much larger and very fine diameter wires are more costly. Present unit price ranges widely between $0.24 and $5.28/ft over the range of available wire sizes.

TABLE 70.5 Temperature Upper Limits for Different Wire Diameters

Dia., mm	0.025	0.127	0.254	0.330	0.510	0.813	1.600	3.175
Dia., in.	0.001	0.005	0.010	0.013	0.020	0.032	0.063	0.125
AWG	50	35	30	28	24	20	14	8
Type	Temperature Limit, °C							
B	—	—	—	—	—	1,700	—	—
C	—	—	—	—	—	2,315	—	—
E	290	325	370	430	430	540	650	870
J	230	275	320	370	370	480	590	760
K, N	690	730	760	870	870	980	1090	1260
R, S	—	—	—	1480	1480	—	—	—
T	90	110	150	200	200	260	370	—

70.8.4 Standardized Tolerances

ASTM standardized tolerances for as-manufactured thermocouple materials are established by committee formal consensus judgment of many experienced producers, users, and calibration standards laboratories staff [8]. Manufacturers characterize the Seebeck property typical of a production batch by calibration of one or more surrogate samples. Sampled first-cycle calibration is taken to be representative of all other thermoelements of the same batch. No authentic inhomogeneity tests are performed. Inhomogeneity has been shown usually to be well within tolerance for new material. Once conformity to standardized tolerance is affirmed, rather than the specific calibration, the standardized type-characteristic is often attributed to the entire batch. Standardized tolerances include not only overall deviation from the standardized Seebeck properties, but also cover inevitable small uncertainty from tolerable inhomogeneity and, in application, even possible irregular deviations, over small temperature spans, of user-simplified scaling relations from the standardized characteristic.

Tolerances apply only initially and briefly to material as delivered, not exposed to excessive temperature or adverse environments. The initial tolerances for standard letter-designated thermocouple types are presumed to be reliable until the thermocouples are exposed to an aggressive environment or for a few weeks at excessive temperature. Thermocouple tolerances are additional to uncertainties of monitoring instruments and of thermal coupling and mass.

There are separate initial tolerances for *normal-* and for *special-grade* material. Initial percent tolerances of ordinary material may be double or more of that of special-grade material. Yet the different degree uncertainty between *commercial-* and *premium*-grade tolerances is small and the cost premium is large (but less than the cost of individual calibration). In severe use, special-grade materials can quickly degrade to exceed the initial tolerances of normal-grade material. Additionally, ASTM Standard E235-06 specifies special requirements for Types **K** or **N** thermocouples destined for nuclear or high-reliability application.

70.8.5 Other Thermocouple Types

70.8.5.1 Standardized Non-Letter-Designated Thermocouples

Some less used commercial thermocouple pairs are for specialized use in thermometry. Some alloys are proprietary. These have not been formally standardized by letter designation or color code. Though not letter-designated, Seebeck properties of some commercial alloys that have significant use are addressed in ASTM Standard E1751 that presents NIST-defined Seebeck properties. Most are relatively expensive high-temperature alloys that promise benefits of accuracy, stability, and environmental tolerance for special application. These include elemental pairs, Au vs. Pt, Pt vs. Pd, and Platinel®, a proprietary thermocouple alloy of Englehard Industries that has Seebeck characteristics very similar to Type **K** with improved stability and accuracy and with greater sensitivity than other precious metal materials. The very pure precious metal thermocouples initially can be very accurate but are particularly vulnerable to contamination and introduced inhomogeneity in use. Infrequently, as usage broadens, additional commercially popular materials are considered for standardization.

Producers and vendors apply proprietary trade names to letter-designated types. A few apply their own unofficial letter designations and color codes. Informal designations might be inconsistent between sources. Identifiers assigned by individual manufacturers have no universal meaning. Letter Type **U** is ambiguous. German standard DIN EN 60584-1 specified thermocouple Type **U** as Cu vs. CuNi with brown jacket, red positive, Brown negative. However, that letter also has ambiguously been used to denote a Cu/Cu pair used as a passive thermocouple extension or for RDT or thermistor wiring. Seebeck characteristics and informal color codes for nonstandardized materials should be confirmed by the vendor.

70.8.5.2 Unconventional Thermoelement Materials

Any thermoelectrically dissimilar pair of conducting materials can form a thermocouple. The *absolute Seebeck characteristic* of thermocouple platinum, Pt-67, a common thermoelectric reference material, was determined by the Center for Information and Numerical Data Analysis and Synthesis of Purdue University under contract to, but not published by, NIST. Others have published the E(T) absolute characteristic for platinum as well [5,9,10]. Limited information on Seebeck properties and application for a multitude of nonstandard materials used as thermoelements is available in papers or in one useful compendium [10]. Most such combinations have not been used widely and are not appropriate for general thermometry.

70.9 Thermocouple Hardware

70.9.1 Material Sources

Thermoelectric alloys and thermoelements are manufactured by a very few specialized sources. These manufacture bare thermoelement bar stock, wire, and foil, and selectively pair thermoelements for thermocouples. Fewer sources fabricate the MIMS (see Section 70.9.3.5) thermocouple materials that are incorporated by others into finished products. These producers of basic forms and very many other sources fabricate these starting forms into finished specialized thermocouple cables and a multitude of products. Few vendors of thermocouples make the thermoelements. Some vendors do not manufacture thermometry products they offer.

70.9.2 Quality Assurance

Pioneer thermocouple users worked with commercial-grade metals and alloys that were not controlled for thermoelectric sensitivity or homogeneity. Each thermocouple had to be individually calibrated and tested for uniformity. Now, thermoelectric quality assurance is first applied by the thermoelement manufacturer. However, properties might later be modified during additional fabrication. Additional thermoelectric testing could be performed by the provider of the final thermocouple product. Infrequently, through the production chain, materials might be modified or misidentified in type or color code. The prudent user confirms the materials.

70.9.3 Forms of Thermoelements

70.9.3.1 Bar Stock

Some thermoelectric products, such as links, connectors, foil, and so on, require bulk feedstock available in a form not usually seen by consumers. The processed materials may be re-characterized as later fabricated. Circuit components that are short and of high thermal conductivity tend to remain nearly isothermal in application; therefore, they introduce little Seebeck emf.

70.9.3.2 Foil and Film

The Seebeck coefficient of metals is independent of material dimension down to about 3500 Å, or even much less, thickness. Commercial thermoelement foils have thickness down to 0.013 mm (0.0005 in.), leaf to 0.01 in. Vapor-deposited alloy films, properly conditioned, can have Seebeck characteristics like thicker material. That allows fabrication of intricate thermoelectric circuits. Very thin thermoelements are more vulnerable to mechanical damage and degradation than are the wire sizes most used in thermometry. Rolled thin foil, ribbon, or vapor-deposited film thermoelements are used in thermocouples specialized for transient or much localized thermometry.

70.9.3.3 Wire

Thermoelements are most familiar to users as wire or cylindrical rod in diameters from 0.013 to 3.3 mm (0.0005–0.128 in.) or more. Individual or thermocouple-paired bare thermoelement wire is available

to consumers. Such bare materials are not indelibly identified by color code. Identification is by label. Labels are easily separated from the material. It is important to maintain identification of all thermoelectric materials by type, color code, source, and batch throughout their use. Individual thermoelement wire is also available with color-coded insulation. Insulating tubing in all the international coding colors is available for use on bare thermoelements. Casually applied by a user, the color code might be incorrect or misinterpreted.

70.9.3.4 Cable

A very wide variety of measurement and extension-grade thermoelements and thermocouple pairs are available cabled as insulated in thermally and chemically resistant plastics and other materials. Individual pairs in solid or stranded form may be covered with protective insulation and wire braid or shielding foil. For severe duty field and process industry use, they are assembled into cables of many separately color-coded pairs that have mechanically and chemically resistant jackets. Some have rugged steel reinforcement. Special connectors can connect extension cables of 24 or more thermocouple pairs.

70.9.3.5 Mineral-Insulated, Metal-Sheathed Thermocouple Material

Many fabricated thermocouple assemblies have thermoelements tightly drawn within a thin-walled tubular metal sheath. Such bulk material is available drawn to overall diameters of 0.075–6.35 mm (0.003–0.25 in.). Materials are controlled to conform to ASTM E585. The process results in highly compressed mineral insulation that allows bending to a radius as small as a few sheath diameters without harm. Many measuring thermocouples of special design use MIMS thermocouple construction. Needle-sized hypodermic probes must be rigid; lengthy small-diameter material is very flexible.

High isolation resistance between thermoelements is essential. The crushed insulant is very hygroscopic. The resistance is difficult to maintain in very small-diameter MIMS material. Immediate sealing is particularly important for MIMS material of small diameter. Ends opened for fabrication must be quickly sealed to avoid absorption of moisture that decimates resistive isolation of thermoelements. Special epoxy sealants are used for the lower temperature end. Sealants for the exposed measuring junctions of MIMS thermocouples require fused glass or other high-temperature material. Such seals may breach in service.

MIMS material is available in all standardized thermocouple types. Different sheath and insulation materials are adapted to a variety of application environments and to temperatures up to 2315°C. MIMS thermocouple material as seen by most users is as commercially fabricated into the variety of special measuring thermocouple assemblies. Like bare wire, the bulk MIMS material is not color-coded. It is identified by batch or product label.

The sheath of a MIMS thermocouple can contain several thermoelements of the same or different thermocouple types. Additionally, for concurrent longitudinal temperature profile measurement, several individual small-diameter MIMS thermocouples can be enclosed and swaged tightly within an overall metal sheath with measuring junctions longitudinally spaced at intervals.

The thin-walled MIMS thermocouple sheaths appear impervious to external contaminants but unrecognized pinholes and fine cracks can allow penetration of moisture or other contaminants. Some surrounding materials at very high temperature can actually diffuse through intact sheaths. Moreover, constituents of the sheath alloy or of its thermoelements can migrate through intact mineral insulation to locally affect the Seebeck properties of the thermoelements.

An externally similar less costly metal-sheathed form simply encloses insulated thermoelements loosely within a tubular sheath without drawing. This less expensive rigid form cannot be bent. It lacks most benefits of MIMS construction.

70.9.4 Insulation

Formerly, most measuring thermocouples were of bare wire insulated using simple ceramic beads. Beads of nesting fish-spine and many other forms still provide flexible high temperature insulation

for bare thermoelements. The insulators and exposed thermoelement are vulnerable to contamination. Specialized insulators now include lengthy rigid cylindrical forms with one to six holes for thermoelements. Lengths may be up to 5 ft. Some are usable to temperatures up to 1600 °C. Typical materials are alumina or mullite. Diverse applications require a variety of insulating materials of different forms and costs adapted to particular temperature ranges.

Individual thermoelements and thermocouple pairs usually are insulated by flexible insulation. Typical materials are Kapton, Teflon, and glass or ceramic braid. Flexible insulation temperature upper limits range from 230 °C to 1200 °C. Extension pairs are normally exposed to temperatures near room temperature. Insulations of connectors or extensions may be rated to 220 °C or even to 650 °C. However, if used to such high temperatures, a significant part of the emf will be from the uncalibrated extension rather than from the measuring thermocouple.

Special thermally conducting and electrically isolating silicone pastes and epoxies are used for coupling or temporarily bonding thermocouple sensors to the test subject. Typical temperature limits are between −40 °C and 200 °C. Special adhesives for thermal coupling, electrical isolation, and bonding are for use to 260 °C.

70.9.5 Connectors

Quick-connect plug and jack connectors using thermoelement materials as contacts are of full, miniature, and sub-miniature sizes with either two or three pins. For MIMS dual element thermocouples, four-pin styles are used. The largest pin marks negative polarity. Pins can be of cylindrical, flat, or banana-plug form. Other styles include specialty connectors of circuit board or "Sub-D" form and multipin connectors for 25 or more pairs. For thermometry in electrically noisy environments, connectors for shielded cable add shield continuity straps. Some have built-in ferrite core noise filters. Most connector shells are of plastic with temperature limits up to 220 °C or of ceramic rated to 650 °C. Connector shells are color coded like the associated thermocouples. For large installations of multipair cable, special barrier terminals or panel-mounted connectors, maintained isothermal, can be used.

Some thermocouple measurements must be made within high-pressure or vacuum chambers. The circuit must cross an impervious wall. The temperature difference across thermoelectric feedthroughs is likely to be much greater than for connectors used near ambient temperature. Therefore, they are more subject to error than are quick-connects. An alternative penetration method uses a compression-style feedthrough that allows several uninterrupted insulated thermoelements to pass sealed through a wall while maintaining pressure or vacuum integrity. That avoids adding incidental thermoelements.

70.9.6 Thermocouple Junction Styles

Junction designs vary widely for specialized application. Most junctions are very localized. Some, to average surface temperature along a line, have widely extended a junction between foil thermoelements. Measuring and reference junctions are conspicuous. However, the many incidental junctions are inconspicuous and may not be recognized yet, uncontrolled; their temperatures can significantly degrade accuracy.

70.9.6.1 Contact Junctions

The area of contact between thermoelements has no influence on the Seebeck emf *if* thermometry is performed under proper nulled-current conditions. An electrical point contact is as effective as contact over a broad area. A pair of pointed probe tips contacting a conducting subject closely spaced can serve as the measuring junctions to measure local temperature. The characteristic emf difference between a pair of well-spaced thermocouples with contact measuring junctions heated at two different controlled temperatures can be used to identify metals. Temporary measuring junctions can be formed merely by holding or taping together crossed thermoelements. Thermoelements accidentally shorted by contact

apart from the measuring junction monitor temperature where shorted, not at the intended measuring junction. In calibration, ends of thermoelements often are electrically connected to extensions by immersing bare tips into a bridging isothermal pool of mercury in an ice bath.

70.9.6.2 Intrinsic Thermocouples

If the subject of thermometry is an electrically conductive material, tips of the two thermoelements can be fused directly to the test subject as by spot welding. These form junctions of an "intrinsic" thermocouple. Closely spaced fine wires or thin foils spot-welded to the subject allow the fastest possible electro-thermal response and reduce the influence of the thermocouple on the temperature being measured. For fastest response, the two thermoelements should be fused to the subject nearly in contact side-by-side or end-to-end (but not one over the other). As the fused spots are very close together, they tend to be nearly isothermal even in transient thermometry so any bridging conductive subject material has negligible effect on the Seebeck emf. However, if the points of fusion are widely separated, the nonisothermal bridging conductor of indefinite thermoelectric property can contribute spurious emf to the measurement while the two weld spots are at different transient temperatures.

For transient thermometry, the thermoelements at the intrinsic junctions should be short and as fine as feasible to reduce thermal loading. A strain-relieved transition from the fine filaments to more substantial *corresponding-type* extension wires should be made. The effect of circuit capacitance or inductance on electrical response is significant only in transient thermometry. As the intrinsic junction is electrically common to the subject, circuit grounding and isolation must be appropriate.

70.9.6.3 Twisted Junctions

A temporary junction contact can be formed merely by twisting together ends of dissimilar wire thermoelements. Twisted junctions may also be soldered or tip-welded for strength. Such robust elongated connections do not affect the Seebeck emf as the sensing junctions are the dual interfaces where the adjacent thermoelements first contact the twisted section. The mass of the extended twisted section only slows thermal response and thereby can affect the temperatures of the junctions and the subject. Such junctions are not used for transient thermometry.

70.9.6.4 Fused Junctions

The familiar "bead-junction" style of thermocouple is formed by fusion of paired thermoelements. Specialized thermocouple welders perform the melt in a localized inert atmosphere to reduce local oxidation of the thermoelements. The spheroidal bead is neither the "junction" nor an emf source. Its purity is irrelevant. The bead is an indeterminate thermoelement alloyed from the joined thermoelements. It is the *interfaces* between the bead and thermoelements that are junctions. They function as one if the bead is isothermal.

Compact junctions are formed by spot welding of crossed thermoelements. The free ends can be trimmed to leave a small robust junction. A very compact junction can also be formed as a smooth "butt-weld" for thermoelement wires as small as 0.003 in. diameter fused end-to-end.

70.9.6.5 Enclosed Junctions

The metal sheath of a mineral-insulated (MIMS) style thermocouple usually encloses two or more thermoelements. Thermoelements can be of different thermocouple types. Thermoelements may be paired with separate junctions or several legs may share a common junction as a "tip-branched" thermocouple. Junctions usually are enclosed within the sheath and are either isolated from or made electrically common with the closed sheath tip. Isolated "junctions" are fused beads insulated within the enclosed end of the sheath. A form with junctions electrically common to the sheath tip inappropriately is termed a "grounded junction" even though it may not be electrically common with the important circuit "electrical ground" point (i.e., "grounded" junctions might be electrically "isolated" from electrical power reference ground). Alternatively, bead-style junctions, bare or insulated, can extend outside the sheath

for faster response. The sheath end must be sealed immediately and remain sealed. A short span of the MIMS material tip that includes enclosed junctions can be drawn to reduced diameter. Such thermoelements are reduced proportionately in diameter. The reduced tip diameter speeds transient response and localizes temperature measurement.

70.9.7 Specialized Thermocouple Assemblies

70.9.7.1 Measuring Junction Assemblies

The simple Seebeck phenomenon invites more diverse specialty thermometry designs than any other thermal sensing technology. Seebeck would be astonished! Process, industrial, and scientific applications for thermocouples now are specialized to diverse applications. Manufacturer's catalogs present dozens of variations. These fabricated assemblies may conceal the functional measuring junction and adjacent thermoelements and circuit components within housings that are designed for convenient handling and for protection. Thermoelements of these specialty forms might be subjected to degrading mechanical treatment in fabrication. The irregular geometries and size, and the fact that sensitivity is distributed rather than localized to the measuring junction, complicate calibration and homogeneity verification.

Most specialized thermocouple assemblies are of the dual-reference-junction circuit style of Figure 70.2, except that they contain many more inconspicuous incidental junctions and thermoelements. For ease of connection, each packaged measuring junction assembly is fitted with a thermocouple plug connector that mates through a connector jack to extension cables or instruments.

The concealed measuring junction, incidental junctions, thermoelements, and extensions are subject internally to all the problems illustrated by the Functional Model for dual-reference junction circuits. The distributed nature of the Seebeck emf requires an authentic understanding of thermoelectric circuits to recognize sources of significant thermoelectric error concealed within assemblies.

70.9.7.2 Reference Junction Assemblies

Most specialized thermocouple assemblies are for use as measuring junctions. As a convenience, a few thermocouple assemblies are specialized to physically apply the 0 °C temperature references as in Figure 70.3. Such reference junction assemblies form most of the thermoelectric circuit as they combine the reference junction with extension leads and connectors. The separate measuring thermocouple simply plugs into the assembly. Though convenient, that reference system is not best suited for accurate calibration.

Resembling a regular color-coded connector, an optional battery-powered Reference Temperature Compensator can bridge the measuring thermocouple and its passive extension with regular thermocouple plug and jack connectors. Inserted near the measuring thermocouple assembly, the compensator electronically applies reference voltage to place the reference junctions effectively at the conventional 0 °C reference temperature. This component, used only with passive-style extensions, is usually joined directly to the measuring thermocouple or its "active" extensions. If this local reference is used, the remote monitor must *not* redundantly apply additional compensation. The connector-like reference is small so it allows the reference compensation to be individually oven-calibrated over the temperature range near ambient temperature.

70.9.7.3 Wells

Process thermometry may require the measuring junction to be isolated from a chemically damaging or mechanically abrasive environment, sealed from pressure or vacuum, or to isolate a sanitary process from the thermocouple. Thermometry, as of incinerators, thermal reactors, and coal gasifiers, is a severe challenge because vibration and the chemical environment are particularly destructive of refractory thermocouples required for the high temperatures. Thermocouple-specialized metal, ceramic, or plastic

wells extending sealed into a process allow replacement of the sensor without interruption. These rugged thick-walled metal or ceramic wells separate the junction from direct contact with the process so they are used for thermometry of steady or slowly varying temperatures. Special wells can flow protective gases along the thermoelements adjacent to the junction. The wells usually are used with MIMS thermocouples that must be seated and thermally coupled in good thermal contact with the well. These massive protections with lengthy immersion depths also beneficially tend to place the measuring junction in a nearly isothermal region where locally degraded segments near the junction contribute less spurious emf.

70.10 Quality Assurance by Users

Manufacturers routinely apply quality control to their thermocouple products, yet users are responsible locally to assure that no occasional defect, error, or change has occurred in manufacture or installation. Several simple and routine test procedures are available to the user.

70.10.1 Thermoelement Identification

Color codes apply to extension-grade wires, to insulated thermocouple-grade wire, and in principle to bare thermocouple wire. Bare wires are not indelibly color-coded. An installer very skilled in electronics might not be familiar with the distinctive thermocouple codes or their consequence. Thermocouples of the same letter-designated type but of different international standards have different international color codes for both type and polarity. The prudent user, before installation or application, will confirm material-type identification of each thermoelement independent of the color code. Definite type identification must be by a combination of methods. A single identification method may be indefinite. No method fits all materials or circumstances.

70.10.1.1 Visual Identification

Type **TP** thermoelements are of copper. Bare, they are visually distinguished by their distinctive reddish color. **JP** thermoelements are of iron and have a distinctive matte gray cast. Other base-metal alloys and platinum and its alloys, if bare, are not easily distinguished by appearance. Usually they all have a similar bright silvery appearance unless, if heat treatment or if bared from compressed mineral insulation, fabrication has produced a roughened gray matte surface appearance.

70.10.1.2 Magnetic Identification

A type **JP** (iron) leg is strongly magnetic. Type **KN** is only weakly magnetic. All other standard thermoelement materials are nonmagnetic. **JP** and **KN** thermoelements can be distinguished from the others by testing the relative attraction to them of a small magnet.

70.10.1.3 Resistive Identification

Thermocouple assemblies or cables usually have paired thermoelements that are of the same length and wire size. If the temperatures of the ends are about the same, the resistance of each thermoelement and the ratio between leg resistances are diagnostic for identification of the materials. Room temperature resistivities of thermoelements are given in Table 70.2. If the identity of one thermocouple or a thermoelement of a batch of like material, wire size, and length is determined, the others can be confirmed by comparison of loop resistances. Resistance measurement of isolated individual thermoelements requires that both ends be accessible at the same location and temperature. For assembled thermocouples with inaccessible measuring junctions, only the isothermal loop resistance can be measured and compared with calculated loop resistances. This quantitative loop resistance measurement is complicated by Seebeck emf that can affect apparent resistance if the circuit ends are at a substantially different temperature.

70.10.1.4 Thermal Identification

Less conveniently, a pair can be identified by the terminal voltage or indicated temperature for a known temperature of measuring junction and the reference junction. Definite identification does not require the accuracy of formal thermometry. Reference junction temperature may be applied by simple ice bath or compensation by the monitor. An independently known temperature, at least 200 °C, applied to the measuring junction for identification is required. Otherwise, uncertainty of the thermocouple calibration and of the imposed temperature would make the emf distinction of thermocouple pairs unreliable, such as **E** from **J** and **K** from **N** or **T** by using a 100 °C boiling water bath. Very similar types **R** and **S** (that have identical color codes) can be reliably distinguished only at much higher temperatures or by formal calibration.

70.10.2 Active Tests

70.10.2.1 Symptom and Occurrence of Inhomogeneity

Thermoelectric inhomogeneity, though now widely misunderstood, has long been recognized as a very real and commonplace problem of thermocouples applied in damaging circumstances. Modern production of thermoelements now results in only slight, but measurable, irregular longitudinal variation of Seebeck coefficient, $\Delta(T, x)$, distributed along the full length of a thermoelement. Maximum deviations of $\Delta(T)$ typically are a very small fraction of the standard tolerance. The risk of significant inhomogeneity in as-received new wire material is slight.

Inhomogeneity that is significant in thermocouple thermometry occurs as a broad localized distinctive deviation from the initial uniform Seebeck coefficient. Recognizable drift during measurement of a temperature expected to be constant is always a symptom of progressing inhomogeneity. Inhomogeneity rarely is authentically tested in commercial practice or by calibration laboratories.

Profound inhomogeneity very often results from prolonged use at high temperature, in a harsh environment, or during fabrication. Most often, significant inhomogeneity is caused or aggravated by excessive temperature so it usually occurs broadly distributed adjacent to the measuring junction precisely where it has the greatest adverse effect in thermometry when junction temperature varies. Less often, it results from defective insulation, metallurgical interaction between thermoelements or with their insulant, evaporation of alloy constituents, or from exposure to higher temperature regions or chemical attack remote from the measuring junction.

Conventional NIST-traceable precise and costly calibrations might sincerely but incorrectly be certified even if a degraded thermocouple can accurately be measured to be locally *severely* inhomogeneous and of indefinite uncertainty as actually applied in thermometry. Most troublesome inhomogeneity is unrecognized even by experienced thermocouple users. Therefore, inhomogeneity remains a commonplace, very real, but usually discounted problem. Display of the distribution of such inhomogeneity requires a special test.

70.10.2.2 Authentic Inhomogeneity Test Methods

A consequentially inhomogeneous segment has, for a test temperature step, ΔT at $\mathbf{T_a}$, a deviant Seebeck emf, $\Delta E(x)$, which is significant relative to initial tolerance. The objective of inhomogeneity testing is to identify locations, longitudinal distributions, and the relative amount of deviations from the initial Seebeck coefficient along thermoelements. The inhomogeneity test is used only to disqualify for calibration or thermometry. It is not a recalibration. Inhomogeneity identified by scanning a 100°C step above ambient temperature is adequate to disqualify a thermocouple for use or recalibration. While the relative magnitude of excessive inhomogeneity at moderate temperature might be greater or less at higher temperature, inhomogeneity significant at a temperature of use is also apparent at moderate temperature. It is not necessary to test inhomogeneity at the temperature of intended use.

Several authentic thermoelectric inhomogeneity tests, simple to sophisticated, have been developed and extensively demonstrated [9,11,12]. Effective primitive methods were practiced by pioneers.

Valid tests range from methods of many-cm spatial resolution to advanced methods that have accurately resolved inhomogeneity with spatial resolution of a few millimeters [11,12]. Their practical need has been very clearly proved, yet few authentic inhomogeneity tests are now performed.

Consequential inhomogeneity is spatially distributed; therefore, all authentic inhomogeneity tests of thermocouples require moving a very narrow temperature step between two temperature plateaus scanned progressively along the paired thermoelements while spatially monitoring the voltage. This appropriate step temperature pattern is best applied by progressive immersion of a thermocouple from ambient temperature into an isothermal-heated deep liquid bath of high thermal conductivity while recording the voltage as a function of position. This step compares the Seebeck emfs of widely separated segments of nearly equal length. The Seebeck coefficients of the two segments are likely most dissimilar if the segments are well separated. The finest resolution is by stepped-progressive immersion of the thermocouple into a deep isothermal liquid bath of high thermal conductivity, such as of a eutectic alloy of gallium, indium, and tin. That alloy is liquid at room temperature. Heated to near 100°C, the nondamaging bath imposes a narrow step of temperature above ambient temperature at the liquid interface. The bath test is well adapted to MIMS thermocouples. It is applied to other thermocouple forms by enclosing them temporarily in a thin close-fitting insulating plastic sheath. The spatial resolution depends on immersion speed, step duration, and thermal properties of the liquid. Even water-based liquids below their boiling point can be used but with lower spatial resolution. Such advanced tests well-demonstrated by several laboratories have used specialized equipment so they have not been accepted by the thermometry community. However, the Australian National Measurement Laboratory has routinely preceded calibration by one *authentic* inhomogeneity test [9]. The test uses progressive entry of the thermocouple into a lengthy isothermal oven while monitoring the emf. That produces an authentic inhomogeneity test with spatial resolution of several cm. The vertical immersion tests restrict the length of the test subject.

A similar authentic test is performed on individual thermoelements of any length. A thermoelement is run through a heated liquid bath or furnace with widely separated entry and exit points. This imposes two spaced temperature steps of equal width and temperature different from ambient temperature. The effect of such progressive immersions is well visualized by the **T/X** sketch of the Functional Model.

70.10.2.3 False Inhomogeneity Tests

70.10.2.3.1 Comparison Test

Some procedures are misunderstood to be inhomogeneity tests. For example, two or more samples taken from well-separated locations along a batch of thermocouple wire are calibrated. Calibrations are of indefinite local segments of thermoelements. Agreement between two calibrations only suggests that homogeneity is within thermocouple tolerance, but it effectively misunderstands troublesome inhomogeneity to be global rather than localized. It is *not* an authentic inhomogeneity test.

70.10.2.3.2 Temperature Spike Test

Another commonplace quality procedure also is misunderstood to be an inhomogeneity test. This "spike test" is the very opposite of a true inhomogeneity test. The thermocouple pair, of any length, must be electrically joined at the end opposite the thermocouple monitor. That junction is held at any fixed temperature. While monitoring the temperature indication, a symmetric very narrow swept spike of temperature is applied by a narrow-tipped soldering iron, flame, or a transverse heated wire that is slowly scanned along the thermocouple. This commonplace practice inappropriately compares the difference between Seebeck coefficients of two very closely spaced segments that are most likely to be similar even in severely inhomogeneous thermoelements. The narrower the temperature spike, the worse the test as an indicator of consequential inhomogeneity. Its widespread misunderstanding as an inhomogeneity test has misled many to discount inhomogeneity as a practical significant problem in accurate thermoelectric thermometry.

70.10.2.4 Hidden Junction Test

The temperature spike test does serve a very different practical purpose. As the very narrow temperature source scans past a location, a localized spike of temperature indication reveals and locates a real junction. This quick and simple procedure actually is an important test for hidden real junctions. Such defects do infrequently occur where different thermoelements, spliced or connected hidden within a cable assembly, are different though of the same nominal type or even of a different type. Otherwise, concealed, they might be connected in crossed polarity. Such consequential defects, unless discovered by this important test, are very difficult to recognize during setup and thermometry.

70.10.2.5 Thermocouple Instability Test

The most obvious occurrence of thermocouple inhomogeneity is that which, unrecognized, is produced in the traditional thermocouple test of long-term drift of sensitivity. The measuring junction, and adjacent lengths of thermoelements, is fixed deeply immersed (more than 20 probe diameters), within an essentially isothermal zone of a furnace. Oven temperature and environment remain constant. Thermoelectric instability is measured as a sustained progressive change of thermocouple voltage that, with exposure time, drifts toward the tolerance limit and beyond. Costly long-duration thermocouple instability tests, conducted by suppliers rather than by users, are usually performed over durations from weeks to months.

The traditional instability test result is misperceived as if a test of overall rather than localized thermocouple sensitivity. Peculiarly, the emf drift observed is not from those segments that are uniformly degrading within the isothermal furnace. Actually, as understood from the Functional Model and T/X sketch, the material degrading isothermally within the furnace contributes little of the observed voltage. Rather, the observed drift of emf is of thermoelement segments at the oven entry that span the temperature difference between furnace and ambient temperature remote from the measuring junction. Therefore, the uniform change of Seebeck emf of the segments within the furnace actually is much greater than is reported. Likewise, tolerance is exceeded earlier than observed in this test.

70.10.2.6 Inhomogeneity Test Incidental to Instability Test

A simple in-place inhomogeneity test, complementing the instability test, can reveal the actual inhomogeneity pattern and improves the significance of the drift test at little additional cost. Beginning the instability test, voltages $E_1(x)$ are recorded as the measuring junction, and thermoelements are progressively advanced from ambient temperature to dwell at the drift test position. That data document, for new material, the initial temperature profile of the entry temperature step and of the oven. Eventually, at the end of the drift test, the exit profile, $E_2(x)$, is recorded as the measuring junction is first advanced a distance greater than the width of the entry step into the furnace. That documents the final profile of the entry step with undamaged material. Then, the thermocouple is slowly withdrawn until the measuring junction dwells at ambient temperature. The difference between the two profiles, $E_1(x)$ and $E_2(x)$, from each test subject documents the actual test-induced pattern of inhomogeneity and the relative magnitude of the actual change of the degraded material. Unlike the formal inhomogeneity test conducted near room temperature, this complementary test is performed at the elevated test temperature that produced the inhomogeneity that appeared as drift.

A special automated calibration furnace, Model FUR-1200, offered by the Electronic Development Labs, Inc., www.edl.com, is well adapted to such complementary programmed in-place inhomogeneity test as it allows programmed computer-controlled immersion and withdrawal of the test subject. It also is used to document the effect on stability of cyclic rather than fixed temperature. MIMS assemblies are degraded differently by varying rather than constant temperatures.

70.10.3 Thermocouple Calibration

Thermocouple calibration is intended to improve accuracy [9,13]. Conventional calibration might not. Calibration of a sample from a uniform set of as-delivered thermocouples to confirm conformity to

standard tolerance is appropriate. However, despite individual calibration, the approximated standard scaling relation, rather than the actual calibration, usually is used in thermometry regardless of calibration result. Individual calibration of each new thermocouple can reduce uncertainty only if the specific calibrations are used in data analysis.

Recalibration of thermocouples recognized or suspected as "drifting" in sensitivity is rarely beneficial. Some quality assurance programs mandate periodic recalibration of unused thermocouples. Initial calibration applies generally so long as thermometry is restricted always to temperatures well below the upper temperature limit, in benign environments and for moderate duration. Sensitivity of unused thermocouples does not "age." Recalibration of severely used but possibly reusable thermocouples that have been suspected of error from observed drift is appropriate only after the subjects are proved acceptably homogeneous by *authentic* inhomogeneity test. Otherwise, costly recalibration may be counterproductive in certifying a thermocouple that is not recognized to be seriously inhomogeneous by conventional calibration.

70.10.3.1 Option to Calibrate

70.10.3.1.1 Base-Metal Thermocouples

Two ways to address calibration problems presented by possible hidden inhomogeneity are the following: *First*, for inexpensive base-metal thermocouples, it is acceptable to presume that unused thermoelements are homogeneous within standardized tolerances. First-cycle calibration of unused thermocouples with only brief exposure, and to temperatures, not exceeding a damage threshold, can usefully confirm the actual as-received state and allow more precise data analysis. *Second*, used base-metal wire thermocouples that have shown significant "drift" might be damaged only near the measuring junction. From the known circumstance of exposure, a length of segments believed damaged adjacent to the measuring junction could be discarded and the measuring junction re-formed. For relatively inexpensive thermocouples, economy may suggest that they be replaced if drift is suspected.

In-place *recalibration* of degraded thermocouples by comparison with reference thermocouple temporarily placed adjacent is often suggested. That common practice assumes the spurious junction-source model. Calibration is not of the "junction." Such comparison presumes that the temperature distribution remains constant in thermometry. However, the Functional Model clearly illustrates why in-place recalibration of a drifted inhomogeneous thermocouple can be deceptive if temperature distribution along the thermocouple varies in use.

70.10.3.1.2 Precious-Metal Thermocouples

The accuracy of precious-metal alloy thermocouples relies on their purity. They are vulnerable to contamination and inhomogeneity. Damage from the local migration, absorption, or evaporation of alloy constituents or other chemical contamination is not reversible. However, unlike base-metal thermoelements, costly precious metal thermoelement materials sometimes can be reconditioned but only if they are degraded by plastic strain alone. Platinum-based thermocouple wire might be annealed to an original state by electrically heating full length in air. Such annealed thermoelements can be presumed to be free only of reversible *strain-induced* inhomogeneity. Confirming recalibration is justified after annealing, but homogeneity must be verified. Precious metal thermocouples have a significant material salvage value.

70.10.3.2 Calibration Process

Thermocouple calibration of effective $E(T)$ requires immersing the thermoelements at the measuring junction into an isothermal calibration zone to a depth sufficient to assure that longitudinal heat conduction along the thermoelements does not affect the junction temperature [9,13]. A junction immersion depth of 10–20 times the thermocouple probe diameter is usually recommended. Emphasis on avoiding the effect of longitudinal heat conduction on measuring junction temperature is appropriate.

However, the spurious junction-source model has encouraged the false notion that it is the junction that is being calibrated. As illustrated by the Functional Model, the calibration emf is from unidentified segments of thermoelements, remote from the measuring junction where they pass through the temperature transition from calibration to room temperature. Those unidentified segments then are presumed to be representative of the entire thermocouple. This conventional calibration obscures a practical fact. It is those thermoelement segments adjacent to the measuring junction that are most likely to become inhomogeneous in fabrication and application. Significantly, it is those segments made isothermal that are the most likely to be inhomogeneous so they are eliminated from a recalibration.

70.10.3.2.1 Routine Calibration

Commercial benchtop dry-block calibrators simplify routine calibration. Electronic refrigeration or heating applies the selected isothermal calibration temperature. Different units are required to cover a range from –30 °C to 1292 °C. Thermocouples are presumed to be homogeneous. One or more probes of different diameters simultaneously can reside at sufficient depth in isothermal wells to assure that heat conduction along each probe does not affect measuring junction or reference temperature.

70.10.3.2.2 Certified Calibration

It is inefficient for incidental users of thermocouples to set up for formal thermocouple calibration. Calibration is often performed by specialized in-house or commercial calibration laboratories that offer calibration certified to be directly traceable to NIST. In the most accurate calibration, fixed-point temperature cells realize intrinsic melting point temperatures of pure metals or the melting or triple point of water ice. These sophisticated sources are more accurate and reproducible than dry-well calibrators or monitored baths. A set of cells, each of one particular metal and standardized freezing point temperature, imposes one of the few specific standardized temperatures. They are used in the establishment of the temperature scale, for thermocouple type-characterization, and in the most accurate calibration. The sources are extremely accurate but are very costly and inconvenient to operate. They are used in the highest-level calibration and standards laboratories for calibration of homogeneous thermocouples.

The melting point of pure water ice is most widely used to accurately impose the standard 0 °C reference temperature. A Dewar thermal flask is filled with a slush of finely crushed ice just immersed in pure chilled deionized water. The flask is surrounded by crushed ice so the melting point temperature can be maintained for a very long duration. The flask must be deep enough to accommodate electrically insulated junctions to sufficient depth. Such baths, properly established, can maintain the standard 0 °C melting point within 0.01 °C–0.02 °C. Standard procedures for 0 °C reference baths are defined in ASTM standard E563-08. A poorly formed and carelessly maintained bath of sizeable ice chunks in tap water in an un-insulated glass beaker might deviate from 0 °C by several degrees.

70.10.3.3 Calibration of Measuring-Junction Thermocouple Assemblies

Measuring junction thermocouple assemblies are of diverse specialized designs. The complicated geometry of some styles makes meaningful calibration difficult or impossible. The Functional Model forces recognition that immersion of all or most of the thermocouple assembly into an isothermal oven actually eliminates it from the intended calibration.

70.10.3.4 Calibration of Reference-Junction Thermocouple Assemblies

Reference junction thermocouple assemblies are of the single-reference junction circuit style (Figure 70.3). The reference thermocouple is intended never to be exposed to either excessive temperature or to a degrading environment; therefore, its initial sensitivity is not expected to change. However, initial calibration should assure that the thermocouple is well within standard tolerance. Usually, they are not calibrated as sold and often not as used in calibration of measuring thermocouples. Unfavorably, reference-junction-thermocouple assemblies always contribute some of the emf when used in a calibration procedure. With the measuring junction at ambient calibration temperature, all the emf is from

the reference thermocouple; none from the test subject. For calibration temperatures to about 40 °C, up to half the emf is from the reference thermocouple. It provides the portion of net emf from 0 °C to the ambient temperature at all higher calibration temperatures. If the reference is calibrated, the data cannot be used to correct calibrations of measuring junctions for which they were used.

70.10.3.5 Calibration of Monitor Scaling Functions

Usually, the scaling relations within monitors are not directly accessible to the user. Separate electronic calibrators for monitors allow scaling between emf and temperature to be confirmed. Scaling within the calibrator is performed by internal scaling relations that approximate with the standard tables. The user specifies a temperature and a thermocouple type. The calibrator presents to the monitor a corresponding emf compensated for the temperature of its own reference terminals. The calibrator and monitor must be connected by an extension of the *passive* type. Such calibration is jointly of thermocouple and compensation scaling. Stand-alone reference temperature compensators can be calibrated by the same kind of electronic calibrator.

70.10.3.6 Calibration of Reference Temperature Compensators

Over the small span around normal ambient temperature, monitor reference temperature compensation uncertainties often range from 0.25 °C to 0.75 °C. Few monitors have reference temperature compensators that are separable for calibration. Nevertheless, infrequently, both internal and external references have been observed to drift by many degrees. Such drift is usually unnoticed as it is indistinguishable from variation of measuring junction temperature.

Compensation should independently be confirmed. By temporarily shorting the input terminals, the monitor indicates the scaled temperature compensation it actually is applying. The monitor should indicate a plausible ambient temperature of the input terminals. (Note: In Figure 70.1, with terminals shorted at an actual ambient temperature of 32 °C, the meter incorrectly, by design for economy, is applying a fixed compensation for a constant 21 °C, a systematic error of 11 °C at all indicated temperatures.)

70.11 Thermocouple Instrumentation Systems

70.11.1 Signal Transmission

The few elementary thermocouples used in a compact laboratory or classroom setting often have simple circuits without extensions. The measuring junction is near the monitor. However, in commercial application, as in field and process thermometry, indication and recording often are separated from the measuring junction by hundreds of meters. There are many ways to transmit thermocouple information over great distances.

70.11.1.1 Wired Transmission

70.11.1.1.1 Extension Cables

Passive-, *corresponding*-, and *compensating*-type thermoelectric extension cables are available in a wide variety of constructions and insulations. They can be a single, short, twisted, insulated thermoelectric pair, a retractile cord, or a very lengthy ruggedized multipair cable designed for suppression of both electrostatic (ES) and electromagnetic (EM) noise and for environmental protection.

Twisted-pair unplated copper instrumentation cable can serve as *passive-type* extension grade leads. Thermoelectric *corresponding-type* or *compensating-style* extensions are more expensive than ordinary copper instrumentation cable but must be used where reference junctions must be at a remote monitor.

70.11.1.1.2 Thermocouple Transmitters

A less familiar form of thermocouple signal conditioning, the *thermocouple transmitter* also conveys signals to distant locations by wire, *not* by wireless transmission. The cable is twisted shielded pair

instrumentation cable. Instrumentation transmitters convert the Seebeck emf data to an Instrument Society of America (ISA) Standard modulated 4–20 A current signal protocol or else to an alternative 1–5 V dc protocol. The transmitter is powered through the signal pair from the receiving end. It is a small module that, additional to the protocol, local to the thermocouple compensates for reference temperature, scales the thermocouple emf to a temperature, related voltage, or current, and provides circuit isolation. The transmitter may also locally display temperature on alternative scales. Some fit within the conventional connection head of a thermocouple well. A few are integral to thermocouple connectors that mate with the measuring thermocouple. Others are rack mounted or are adapted for use with the DIN rail mounting system. Thermocouple transmitters are suited to monitoring slowly changing temperatures at monitoring sites distant from the measuring junction.

70.11.1.2 Wireless Transmission

70.11.1.2.1 Radio Frequency

Cabled wire systems may be simpler for general technicians to install and to maintain than are more elaborate electronic systems. However, hardware, cable, and setup cost may favor alternative wireless transmission systems rather than extension wiring. In some field or process situations, widely distributed measuring junctions must be very distant from a central monitoring location. Wireless data transmission can be particularly economical where the distance between the measuring junction and recorder is great, when many thermocouples must be recorded, when little setup time is available for field cabling, and where the measuring setup must be moved occasionally.

Several inexpensive radio frequency wireless data acquisition systems are adapted to combine thermocouple signal conditioning, data acquisition, recording, and wireless signal transmission. Inexpensive connector- or palm-sized radio transmitter systems communicate wirelessly over distances up to 300 m under favorable conditions. Radio modems can communicate data from a measuring location to recording sites as far as 5 km away. In the United States, they do not require a communications license.

70.11.1.2.2 Internet

Where transmission distance is even greater, real-time temperature monitoring via the Internet can provide, local or worldwide, control and alarms for remote process monitoring or administration of field experiments.

70.11.2 Thermocouple Instruments

70.11.2.1 Indicators and Monitors

A wide variety of nonrecording instruments is available for merely observing temperature by thermocouple. Some basic thermocouple indicators are handheld and multipurpose like the multimeter in Figure 70.1 with 1 °C resolution and marginal accuracy. Advanced handheld meters can be very accurate and certified with calibration NIST-traceable and resolution to 0.1 °C. A few provide dual thermocouple inputs and allow direct differential temperature measurement. Some accommodate only a single thermocouple and of only a single type. Other benchtop or rack-mounted units accommodate many thermocouple inputs of the same or of intermixed thermocouple and other sensor types.

70.11.2.2 Simple Recorders

Temperature might simply be monitored visually in setup or for control, but most temperatures now are digitally recorded for documentation and detailed analysis. Local portable recording can be on inexpensive SD, USB, or other flash memory. Some specialized multichannel thermocouple data recorders are now battery powered for stand-alone field use. The range of recorders now extends from tiny handheld devices to specialized fieldable data loggers for untended local recording in isolated locations.

70.11.2.3 Data Acquisition Systems

Stand-alone systems specialized to thermocouple thermometry allow a user to focus on the measurement rather than on details of its indication or recording. Larger digital data loggers with multiplex switching that sequentially samples the output from individual thermocouples in fixed order can accommodate up to several hundred inputs and can record a mix of both thermocouple and other signals at sample intervals programmable from a few milliseconds to hours or days without special programming. Microvolt-level signal resolution, stability, noise reduction, nonlinear scaling, reference temperature compensation, and high input resistance are commonplace. Most provide desirable three-wire and balanced switching between thermocouples. Despite the sophistication of the advanced systems, the manufacturer cannot shelter the user from most thermoelectric problems.

70.11.2.4 Computer-Based Systems

Even tiny specialized recording thermocouple thermometers now perform signal processing digitally. Desktop personal computers accommodate inexpensive internal digital data acquisition boards that can add high-resolution temperature recording to complement analysis capability. Laptop computers can accept special palm-sized PCMCIA or USB plug-ins for multiple thermocouple input for field use.

Computer-based systems offer great flexibility. Digital computer-based data acquisition systems with software integrated for data acquisition, analysis, and presentation allow the user to apply special reference temperature compensation, custom linearization of thermocouple scaling functions for either standard or individual calibrations and of special thermocouple types and sampling to arbitrary sequence and intervals. Computer-based thermocouple thermometry systems with suitable commercial data acquisition and other software provide for experiment design, prediction, data acquisition, alarming, information management, analysis, reporting, archiving, and communication within a single compact fieldable laptop computer. Computer-based data acquisition also allows complementary communication via the Internet.

70.11.2.5 Modular Components

The many diverse and economical packaged data acquisition systems satisfy the thermocouple thermometry needs of most users. However, many major electronic component manufacturers offer thermocouple-focused modular integrated circuit components that simplify special application. Small integrated circuit modules, www.analog.com, for thermocouples provide reference junction compensation, linearization, isolation, open input indication, amplification, and set-point control for incorporation of thermoelectric components into special-purpose systems. Outputs may be of temperature-linearized current or voltage or of the ISA protocol.

Individually, the components are capable. They can be integrated into systems that have other principal functions. Some customers, hoping for economy of hardware, have built customized thermocouple thermometry systems using modular components. Some have proved more costly and less accurate because the designer did not understand critical distinctions between thermoelectric circuits and ordinary electric circuits. Related components, apart from modules, like switches, relays, connectors, and printed circuits can introduce troublesome thermoelectric noise emf. When applying modular components into a system, it is particularly important to understand thermoelectric circuit pitfalls and their avoidance.

70.11.2.6 Nonconventional Components

Occasionally, useful special functions do justify the knowledgeable design and application of unconventional system components. One example is a nonconventional *Combinatorial Multiplexing Switch*, a "CoMux" [14]. To efficiently share a single metering device, reference, and scaling, ordinary computer-controlled switches successively multiplex-sample thermoelements by conventional "positive–negative" thermocouple pairs. Instead, the CoMux samples nonconventional multi-thermoelement "Tip-branched" or "Leg-branched" thermocouples. These allow unique complementary validation measures of continuing

thermoelectric circuit integrity. Tip-branched thermocouples have three or more dissimilar thermoelements that share a single "juncture." Such multi-thermoelement circuits have long been commercially available and used but are sampled only as conventional pairs. Uncommon leg-branched thermocouples have three or more thermoelements with junctions along different legs. The programmable CoMux samples arbitrary combinatorial nonconventional pairing of the several circuit-connected thermoelements. These samples allow such critically important independent complementary diagnostic measures as loop resistance, noise voltage, redundant measure of temperature, and with standard and nonstandard thermoelement pairings of thermoelements of these compound thermocouple circuits. These unconventional diagnostic measures unambiguously reveal and document otherwise unrecognizable flawed thermocouple readings as where thermocouple failure in harsh environments is expected [14].

70.12 Thermocouple Thermometry Practice

70.12.1 Temperature Ranges

70.12.1.1 Moderate and High Temperatures

Temperature and duration of exposure are unavoidable environmental variables in thermometry. Other environments can sometimes be controlled. Standard thermocouple tables extend only to the greatest temperature of *recommended* use for benign protected environments, short durations, and for wire materials of 3 mm or greater diameter. Reduced temperature limits apply for materials of smaller cross section (Table 70.5). Degradation of the Seebeck property can occur due to excessive temperature alone, well below the melting point, eventually even well within the range recommended for use. Environmental degradation is accelerated by heat. Frequent cyclic heating and cooling accelerate damage of MIMS thermocouple. Within the recommended temperature range, damage progresses gradually with time at temperature over periods of weeks to months. Calibration should be performed quickly, allowing time only for equilibration, and only to the maximum temperature of intended use. Although thermoelements may not be visibly affected, the Seebeck coefficient of thermoelement segments adjacent to the measuring junction might be substantially degraded.

70.12.1.2 Very High Temperatures

Type **K** thermocouples are limited to temperatures up to 1372 °C. Proprietary Platinell® has a Seebeck characteristic similar to Type K with improved stability and recommended upper limit of 1392 °C. Types **R** and **S** extend the range to 1768 °C, and Type **B** to 1820 °C. Special commercial gold vs. platinum thermocouples of very special design and construction are used for stable precision thermometry to 1000 °C. Damage by chemical exposure is not reversible, but the precious materials can be recycled. Refractory metal thermocouples are of specially treated materials such as tungsten, rhodium, or rhenium alloys.

70.12.1.3 Low Temperatures

The scaling functions for use of normal thermocouple material below 0 °**C** are different than for elevated temperature use of the same thermocouple type. Standardized thermocouple characteristics below 0 °C are based on less data than for higher temperatures. Also materials manufactured for thermometry at elevated temperatures conform less well to the standard characteristics for cryogenic thermometry than do special modified alloys of the same thermocouple letter type especially made for such use. This quality issue should be discussed with the thermocouple supplier.

70.12.1.4 Cryogenic Temperatures

Cryogenic thermometry has been loosely defined as measurement of temperatures below 280 K (7 °C). That range broadly overlaps the measure of atmospheric temperatures below the ice point (down to –50 °C) that, along with higher ambient temperatures, often are measured by a single thermocouple system. A more restrictive definition limits cryogenic thermometry to temperatures below 90 K (–183 °C)

(the boiling point temperature of liquid oxygen at 1 atm). The latter definition characterizes as cryogenic only the extremely low-temperature regime over which thermometry involves distinctive severe problems and very different techniques. The Seebeck properties are degraded at extremely low as well as high temperatures.

Common thermoelement materials progressively decrease in thermoelectric sensitivity below the ice point. The Seebeck coefficient of all conductors is insignificant at 0 K (–273.15 °C). Superconductive materials experience an abrupt drop in Seebeck coefficient to near zero at a characteristic superconducting transition threshold. That threshold is below 10 K for most unalloyed metal superconductors. Special alloys have been developed to raise the atmospheric pressure superconducting threshold to as much as 138 K—well above the strict cryogenic range. Only between 0 K and a superconductive threshold is it possible directly to observe the absolute Seebeck characteristic of a normal thermoelement by pairing it with a superconductive thermoelement.

Standard Seebeck characteristics are defined for Types **E**, **J**, **K**, **N**, and **T** down to –270 °C. The characteristics for Types **R** and **S** extend down only to –50 °C. Type **B** is not characterized below 0 °C. It is not recommended for thermometry below 400 °C because its sensitivity is very small at lower temperatures. Often, simple *passive-type* extensions are used for Type B thermocouples near ambient temperature where the thermocouple emf is very small.

For very low temperatures, the letter-designated thermocouple material, Type **E**, is preferred for use down to –233 °C (40 K) because of its higher relative Seebeck coefficient in that range. The less-sensitive Type **K** and Type **T** materials are also used in this range. Below 40 K, special nonstandardized alloy combinations, such as Type **KP** vs. Au/0.07 Fe, are preferred.

Some thermoelement alloys experience grain growth and incur serious inhomogeneity under prolonged exposure to deep cryogenic temperatures. More sensitively at cryogenic temperatures than at elevated temperatures, the Seebeck coefficient of most thermocouple alloys is very strongly dependent on magnetic field. Strong magnetic fields are often involved in cryogenic experiments, so thermomagnetoelectric effects become significant in studies of superconductivity.

Peltier heating at junctions and Thomson heating along thermoelements are current-dependent thermoelectric effects that rarely affect thermometry. Fortunately, neither effect is significant if thermometry is conducted properly by "open-circuit" measurement. There is no significant thermocouple self-heating as with resistance thermometers.

70.12.2 Thermometry Process

70.12.2.1 Coupling

As any thermal sensor, the temperature of the measuring junction of a thermocouple must approach and track the temperature of the subject [9,13]. Close thermal coupling of the measuring junction to the point of measurement is enhanced by a very thin layer of thermally conducting coupling grease or adhesive, avoiding even thin air gaps. Electrically conducting coupling between junction and subject may be appropriate for noise control. Thermal coupling and loading problems are aggravated in the measurement of subjects of low thermal mass. Heat transfer is more delayed for surface, liquid, and gas measurements than for internal solid measurements. Centering the measurement point in an isothermal region is preferred. Even for steady-state measurement, temperature gradients can cause the measuring junction to be at a significantly different temperature from the intended point of measurement.

70.12.2.2 Transient Thermometry

The very small dimensions possible for the junction and low thermal mass of its adjoining thermoelements well adapt thermocouples to transient thermometry. The Seebeck effect occurs at electronic speed, independent of thermoelement size. Seebeck emf immediately corresponds to the actual temperatures of all junctions in the circuit. An isolated thermocouple alone has no characteristic response.

Thermal response is necessarily a joint property of the thermoelements at the measuring junction, of its subject, and of the monitoring system. Traditionally, thermocouple transient behavior is represented simply as a first-order exponential step rise time with a time-constant reflecting rise time to respond to 63.2% of an instantaneous temperature step. The actual forms of response for different geometries are similar yet more complex. Risetimes are often stated for beaded-junction thermocouples of different wire sizes plunged into an air or water stream. The response of actual thermocouples depends on many factors such as junction dimension, material, and style of the junction and of adjacent thermoelements, the thermal properties of the test subject, and thermal coupling, as well as circuitry and the recording electronics. Butt-welded fine wire and foil thermocouples might respond within 3 ms. Small-diameter MIMS thermocouples with exposed or enclosed junctions in air have response times in the range from fractional seconds to several seconds. Surface contact thermocouples have time constants around 2 or 3 s. Unlike for static thermometry, in fast transient thermometry, the electrical parameters of the thermocouple and associated circuitry and the dynamic response of the recording system might further slow response time.

70.12.3 Distinctive Thermocouple Noise Problems

Thermocouple emfs are small. Fine temperature resolution requires substantial amplification. For quasi-static thermometry, modern noise suppression techniques available to designers of data loggers overcome ordinary coupled electrical noise to routinely achieve as small as 0.1 °C temperature resolution and long-term stability (much smaller than required by thermocouple tolerance). Thermocouple measurement of transient temperatures requires more careful attention to classical noise reduction techniques because thermoelectric circuits present distinctive problems.

70.12.3.1 Electromagnetic Noise

In thermocouple probes of MIMS construction, ceramic-bead-insulated thermocouples, and some paired insulated thermocouple wires, the paired thermoelements are not twisted and are spaced uniformly apart in a plane. Thermocouples are unbalanced lines because the paired legs have very different resistance. Lengthy and larger diameter thermocouple probes present a significant circuit loop area to couple magnetic noise fields. Even though MIMS probes are entirely metal-enclosed, the sheaths are very thin and of low magnetic permeability. They are scarcely effective for EM shielding. These features prevent the use of some classic techniques for the rejection of EM noise. To reduce troublesome EM noise, the measuring thermocouple probe should be of the smallest feasible diameter and of minimum length. In instances where the EM source is localized and identified, the orientation of the probe relative to the source might be arranged to minimize EM coupling. For rejection of EM noise, extension leads should always be of twisted-pair construction, with a pitch small enough to reject high-frequency noise. Some special thermocouple connectors incorporate a ferrite core intended to suppress transient noise.

70.12.3.2 Electrostatic Noise

ES noise is more easily reduced than is EM noise. Shields of low resistance, though thin, can be effective if properly connected. Shields of plated copper braid are commonly used and are effective for noise of moderate frequency. Continuous shields, such as the MIMS sheaths and aluminized Mylar® film, are effective for low-frequency ES noise and are more effective than braid shields for very-high-frequency noise. ES shields must be continuous (without gaps and holes as in braid) for maximum effectiveness. Shield continuity must be maintained across connectors. Optimum benefits from shielding require use with thermocouple monitors that provide three-wire input and multiplexers that, individually for each thermocouple, switch both the signal pair and their separate shield lead.

70.12.3.3 Grounding and Shielding

Dynamic thermocouple thermometry can require monitoring with rapid frequency response in electrically noisy environments. The technically well-founded principles of noise control for electric circuits

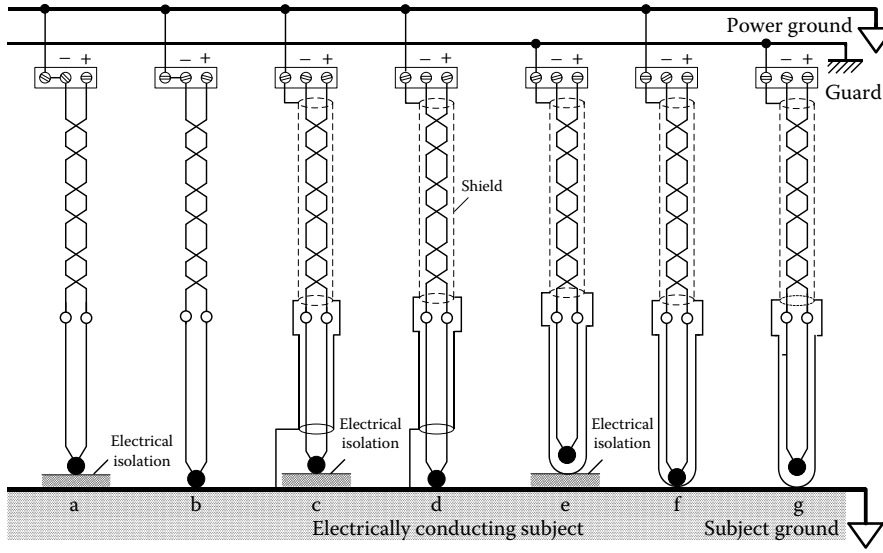


FIGURE 70.8 Preferred thermocouple grounding arrangements. Minimization of electrical noise requires careful attention ion to grounding and shielding.

apply equally to thermoelectric circuits. Well-designed thermocouple instruments carefully isolate the circuitry from other components and control ground reference points. A secondary overall shield, isolated from inner shields and separately grounded, can improve the rejection of EM noise. The circuit grounding of thermocouple is complicated by intimate thermal contact of the measuring junction with an earth-grounded conducting test subject, as for an intrinsic thermocouple, is needed for rapid transient thermometry. Ordinary *coaxial* instrumentation cable must not be used as *passive-type* extension leads because the outer braid and center conductors—as thermoelements—have very different Seebeck coefficients. If not at uniform temperature, such cable can be a significant source of Seebeck dc noise emf.

Figure 70.8 illustrates appropriate grounding for several of the many possible situations. Because the measuring junction is not a localized site of source emf, the point of thermocouple grounding must be carefully considered. Circuit grounding should be at one point only and never at any point of the thermocouple circuit other than at measuring or reference junctions. Electric contact, and particularly shunting or shorting, to a thermoelement at any points at different temperatures between measuring and reference junctions can introduce spurious Seebeck emf from incidental unrecognized thermoelements or ground potential differences.

Feasibility of a grounding method is usually dictated by the internal design of monitors and data loggers. Instrument designers address noise suppression in different ways. Simple and inexpensive line-powered thermocouple thermometers often have only single-ended inputs with the negative lead internally connected to the power ground. This references the sensor grounding also to power ground. In conflict, a "grounded-junction"-style thermocouple might be required for fast transient thermometry. In use, initially isolated thermocouples might short to the protective sheath or test subject. Although this does change the grounding configuration, resulting errors due to thermoelectric effects then usually are predominant.

70.12.3.4 Filtering

Most common thermometry is of static or low-frequency temperature changes. This allows designers to complement the user-provided passive rejection of EM and ES noise by proper grounding and shielding. Internal filtering provided by manufacturers of data acquisition systems can effectively reject high-frequency noise. Some systems provide very effective averaging or "double-slope integration"

techniques for low-frequency thermometry. Routine resolution and stability to 0.1 μV variation over a period of days are achieved by many modern thermocouple data systems. However, these noise rejection methods restrict sampling to intervals no shorter than 10–20 ms that are unsuited to very fast transient thermometry.

70.12.4 Thermocouple Modes of Failure

70.12.4.1 Short or Open Circuit Failure

Thermocouple thermometry has *failed* when its indications are beyond uncertainty limits required for a measurement [14]. Thermocouples sometimes do fail "open" as junctions separate or thermoelements corrode, yield, or melt. Special "open-circuit" indication is a promoted feature of many modern thermocouple indicators. If failure is for other reasons, such indicators encourage unwarranted confidence.

More often thermocouples fail before the circuit opens. The more common but less apparent circuit failures are by insidious shunting, shorting, or progressing inhomogeneity of thermoelements. These common faults are not detected by the "open-circuit" indication [14].

70.12.4.2 Environmental Damage

Corrosive or reactive environments breach protective sheaths and disrupt circuits (Table 70.4). Inhomogeneity is introduced locally in thermoelements. Plastic, ceramic, or fiber insulators are vulnerable to contamination. The compressed granular insulation of MIMS thermocouples, if exposed, rapidly absorbs moisture that can seriously degrade resistive isolation. Electrically conductive fluid or solid shunts can displace the location of measurement without interrupting temperature indication.

70.12.4.3 Metallurgical Change

Alloy thermoelements can locally change composition by evaporation of constituents when exposed to vacuum. Appropriate MIMS construction and material selection usually extend life and increase temperature limits, but damage can occur even within apparently fully sheathed assemblies. Alloy constituents evaporated internally from a sheath or from one thermoelement can migrate and coat adjacent thermoelements even through an insulant. Traces of contaminants can penetrate through pinholes and hairline cracks in metal sheaths and can actually diffuse through intact protective sheaths and affect apparently isolated thermoelements. Minute trace impurities in sheaths or insulants can interact with thermoelements. Alloy components can migrate from one thermoelement to the other through the measuring junction, but the effect is negligible if the junction region is properly maintained isothermal during measurement. Excessive distributed strain can locally modify the Seebeck coefficient of one or both thermoelements resulting in temporary or persistent inhomogeneity. However, very narrowly localized plastic strain as might be introduced by very sharp bends has negligible effect on practical thermometry as is evident from the Functional Model.

70.13 Information Sources for Thermoelectric Theory and Thermometry

70.13.1 Internet

The Internet is a readily accessible and ever-changing source for information on the multitude of thermocouple products. One website, www.temperatures.com, links to more than 300 thermometry and thermocouple-related sites. (These linked sites are not individually screened for accuracy.) Notable is www.omega.com that displays a huge variety of specialty products. Many vendors also provide current tables of Seebeck characteristics, color codes, and properties for standardized as well as for less common materials. Suppliers of nonstandardized specialty thermocouple materials post properties of their products and welcome inquiries.

The definitive NIST tables for seven of the letter-designated types are available on the Internet without charge at www.nist.gov. Site www.hartscientific.com provides values and online conversion between emf and temperature well as Seebeck coefficient, and tolerances for the eight letter-designated types. Another site, www.capco.com provides conversions for the eight U.S. letter-designated and four added non-letter-designated types. A convenient free Apple iPod® thermocouple application presents the full tables and direct conversion between individual temperatures and emfs for the eight U.S. letter-designated types as well as the conflicting color-codes of several international standards.

Unfortunately, very many websites that address thermocouple thermometry contain well-intentioned, but misleading, simplistic or even incorrect information. Most consequential in thermometry is perpetuation of the false thermoelectric junction-source model that obscures most causes of real problems in practical thermocouple thermometry.

70.13.2 Standards, Reports, and Books

A comprehensive, detailed, and authoritative thermocouple application manual is published by the ASTM [4]. That detailed manual addresses applied theory, application, and properties. Closely related standards prepared by the ASTM and sanctioned by ANSI adopt the values and complement the definitive and detailed NIST report 175 [7]. That report is the primary source for Seebeck properties and physical characteristics of most U.S. letter-designated thermocouples. ASTM standards are current as they are reviewed and updated by committee E20 at intervals of 5 years or less. The ASTM also publishes as "gray pages" interim tables of Seebeck emf for a few popular commercial materials that have not been letter-designated. All thermometry-related ASTM standards are collected in a single, annually revised volume [8]. The hardcopy thermocouple tables from these volumes are now available also as functional computer programs that calculate $E(T)$ and $T(E)$ of the NIST 175 document over ranges, at intervals, and in units selected by the user. Many other national and international consensus standards organizations publish thermoelectric test methods and material characterizations.

70.13.3 Technical Papers

Thermocouple thermometry is a mature technology. Progressively, fewer papers now are devoted specifically to thermocouples. Some are distributed across trade magazines and journals of professional societies. Critical details of thermocouple measurements supporting experimental data are most often described only incidentally and inadequately in reports of experimental studies.

The most concentrated and broadest sources of refereed technical papers addressing thermocouple thermometry are the serial proceedings of decennial international symposia on temperature [15]. The seven volumes published to date are a rich, refereed, and mostly reliable source of application data and theory. Other proceedings of annual symposia of the ISA and the National Conference of Standards Laboratories also include papers describing current developments in thermocouple thermometry. Reference 10 has thermoelectric data on the widest variety of nonstandard and ordinary materials.

Fundamental papers addressing the basic science of thermoelectricity are now mostly in the historic literature. Many of the very convincing treatments are directly contradictory. Persons who require historic information must rely on a few specialty technical libraries such as the Linda Hall Technical Library in Kansas City, MO. The online catalog is at www.lindahall.com.

70.13.4 Trade Publications

Abundant, free, extensive commercial catalog literature describes currently available thermocouple hardware and data acquisition products. Catalogs include tables of characteristics. Many manufacturers publish elaborate catalogs of thermocouple-related hardware, software, and books. Notably, the "Omega

Temperature Measurement Handbook," updated and enlarged periodically, presents a supermarket of diverse specialized thermocouple and other components and instruments for thermometry.

70.13.5 Workshops and Symposia

There are occasional workshops and short courses of a few hours' or days' duration devoted to measurement, thermometry, and occasionally even to thermocouple thermometry. Professional societies, such as the ISA, present tutorial workshops in association with special symposia or annual meetings.

70.14 Caveat

At every level, from the promotional to the most esoteric mathematical and physical thermoelectric theory, historic and current, convincing, innocently propagated misconceptions are commonplace concerning thermoelectric effects. More than any other measurement technology, the extensive literature of thermoelectric thermometry, refereed papers, advice from "experts," standards, catalogs, websites, and tutorial articles (*even this one*) all must be approached cautiously with an *authentic* thermoelectric circuit model in mind.

70.15 Summary

The very simple Seebeck thermoelectric effect, the occurrence of an emf in response to a temperature difference along a thermoelement, can be used to measure temperature at modest cost, with finer spatial and time resolution, over a broader temperature range, and in more diverse sensor geometries, than any other class of electric temperature sensors. Properly understood and used within its limitations, the thermocouple is capable of accurate and reliable thermometry. Misunderstood, the thermocouple is subject to inconspicuous, very significant, and sometimes costly errors in thermometry and calibration that too often remain unrecognized. Thermoelectric circuits are distinctly different from ordinary electrical circuits. The practical and useful Functional Model of Thermoelectric Circuits encourages recognition and avoidance of commonplace problems unique to thermocouple thermometry. The principles apply to the most complex generalized thermoelectric circuits, to diverse applications in thermometry, and in the design of low-level electronic circuits.

References

1. D. D. Pollock, *Physics of Engineering Materials*, Englewood Cliffs, NJ, Prentice Hall, 1990.
2. D. D. Pollock, *Thermocouples—Theory and Properties*, Boca Raton, FL, CRC Press, 1991.
3. R. P. Reed, Thermal effects in industrial electronics circuits, in J. D. Irwin (ed.), *The Industrial Electronics Handbook*, Boca Raton, FL, CRC/IEEE Press, pp. 57–70, 1996.
4. R. M. Park (ed.), *Manual on the Use of Thermocouples in Temperature Measurement*, 4th edn. or later, MNL12, Philadelphia, PA, American Society for Testing and Materials, 1993.
5. R. P. Reed, Absolute Seebeck thermoelectric characteristics—Principles, significance, and applications, in J. F. Schooley (ed.), *Temperature, Its Measurement and Control in Science and Industry*, Vol. 6, Part 1, New York, American Institute of Physics, pp. 503–508, 1992.
6. R. P. Reed, Thermoelectric thermometry—A functional model, *Temperature, Its Measurement and Control in Science and Industry*, Vol. 5, Part 2, New York, American Institute of Physics, pp. 915–922, 1982.
7. G. W. Burns et al., Temperature-electromotive force reference functions and tables for the letter-designated thermocouple types based on the ITS-90. Gaithersburg, MD: U.S. Department of Commerce, National Institute of Standards and Technology, Washington, Supt. of Docs., U.S. G.P.O., 1993.

8. ASTM International, *Annual Book of ASTM Standards, Temperature Measurement*, Vol. 14.03, Philadelphia, PA, American Society for Testing and Materials, 2011.
9. R. E. Bentley (ed.), *Handbook of Temperature Measurement, Vol. 3, Theory and Practice of Thermoelectric Thermometry*, Singapore, Springer-Verlag, 1998.
10. P. A. Kinzie, *Thermocouple Temperature Measurement*, New York, John Wiley & Sons, 1973.
11. R. P. Reed, Thermoelectric inhomogeneity testing. Part I: Principles; Part II: Advanced methods, in J. R. Schooley (ed.), *Temperature, Its Measurement and Control in Science and Industry*, Vol. 6, Part 1, New York, American Institute of Physics, pp. 519–530, 1992.
12. R. P. Reed, The effect of interrogating temperature profile in the Seebeck inhomogeneity method of test (SIMOT), in D. C. Ripple (ed.), *Temperature, Its Measurement and Control in Science and Industry*, Vol. 7, Part 1, New York, American Institute of Physics, pp. 491–496, 915–922, 2003.
13. J. V. Nicholas and D. R. White, *Traceable Temperatures—An Introduction to Temperature Measurement and Calibration*, 2nd edn., New York, John Wiley & Sons, 2001.
14. R. P. Reed, Possibilities and limitations of self-validation of thermoelectric thermometry, in D. C. Ripple (ed.), *Temperature, Its Measurement and Control in Science and Industry*, Vol. 7, Part 1, New York, American Institute of Physics, pp. 491–496, 507–512, 2003.
15. *Temperature, Its Measurement and Control in Science and Industry*, Vols. 2–8 (Serial), New York, American Institute of Physics, 1942–2013.

71

Semiconductor Junction Thermometers

71.1 Transistor as a Temperature Sensor **71**-2
71.2 Thermal Properties of Semiconductors: Defining Equations **71**-3
71.3 Integrated Temperature Sensors **71**-4
Integrated Digital Temperature Sensor • Analog-Output Integrated Temperature Sensor • Digital-Output Temperature Sensor • External Circuitry Eliminates Calibration
71.4 Common Applications of Semiconductor Sensing **71**-7
Other Applications of Semiconductor Sensing Techniques
71.5 Temperature Sensing in Power ICs for Fault Protection and Diagnostics **71**-10
71.6 Reliability Implications of Temperature to Electronic Components **71**-11
Junction Temperature Measurement
71.7 Semiconductor Temperature Sensor Packaging **71**-13
Defining Terms **71**-14
Vendors and Suppliers **71**-14
References **71**-15

Randy Frank
Randy Frank & Associates, Ltd and Motorola, Inc.

Temperature sensors can be easily produced with semiconductor processing technology by using the temperature characteristics of the *pn junction*. The batch processing and well-defined manufacturing processes associated with semiconductor technology can provide low-cost and consistent-quality temperature sensors. The temperature sensitivity of the *pn junction* is part of the transistor's defining equations and is quite predictable over the typical semiconductor operating range of –55 °C to + 150 °C.

Most semiconductor junction temperature sensors use a diode-connected bipolar transistor (short-circuited collector-base junction) [1]. A constant current passed through the base–emitter junction produces a junction voltage between the base and emitter (V_{be}), which is a linear function of the absolute temperature (Figure 71.1). The overall forward voltage drop has a temperature coefficient of approximately 2 mV °C^{-1}.

When compared to a thermocouple or a resistive temperature device (RTD), the temperature coefficient of a semiconductor sensor is larger but still quite small. Also the semiconductor sensor's forward voltage has an offset that varies significantly from unit to unit. However, the semiconductor junction voltage versus temperature is much more linear than that of a thermocouple or RTD. In addition to the temperature-sensing element, circuitry is integrated to produce a monolithic temperature sensor with an output that can be easily interfaced to a microcontroller and to provide features that are useful in specific applications. For example, by using an *embedded temperature sensor* with additional circuitry, protection features can be added to integrated circuits (ICs). A temperature sensor becomes an embedded item in a semiconductor product when it has a secondary or supplemental purpose instead of the primary function.

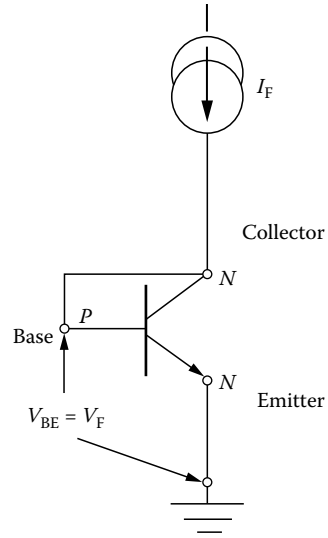


FIGURE 71.1 Bipolar transistor configured as a temperature sensor. The base of the transistor is shorted to the collector. A constant current flowing in the remaining *pn* (base to emitter) junction produces a forward voltage drop V_F proportional to temperature.

71.1 Transistor as a Temperature Sensor

A common semiconductor product for temperature sensing is a small-signal transistor such as 2N2222 or 2N3904. By selecting a narrow portion of the overall distribution of the V_{be} for these devices, a temperature sensor with a lower variation in characteristics can be obtained. The lower variation can provide a part-for-part replacement when a tolerance of only a few percent is acceptable. This device (formerly offered as MTS102 but no longer in production) demonstrates the performance characteristics of the transistor used as a temperature sensor [2].

As shown in Figure 71.2 [2], a silicon temperature sensor has a nominal output of 730 mV at −40 °C and an output of 300 mV at 150 °C. The narrowly specified V_{be} ranges between 580 and 620 mV at 25 °C. The linearity error or variation from a straight line of this device is shown in Figure 71.3 [2]. The total accuracy is within ±3.0 mV including nonlinearity, which is typically within ± 1 °C in the range of −40 °C to 150 °C.

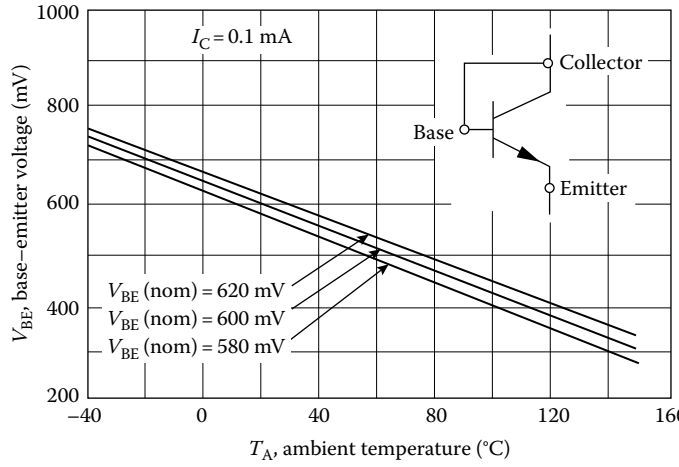


FIGURE 71.2 Base–emitter voltage versus ambient temperature for a silicon temperature sensor.

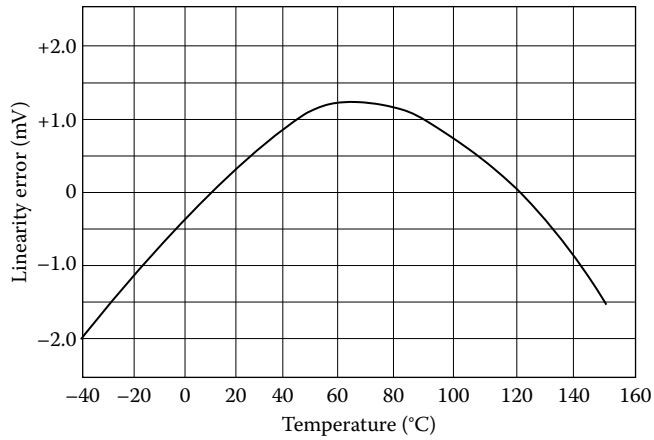


FIGURE 71.3 Linearity error (in mV) versus temperature for a silicon temperature sensor.

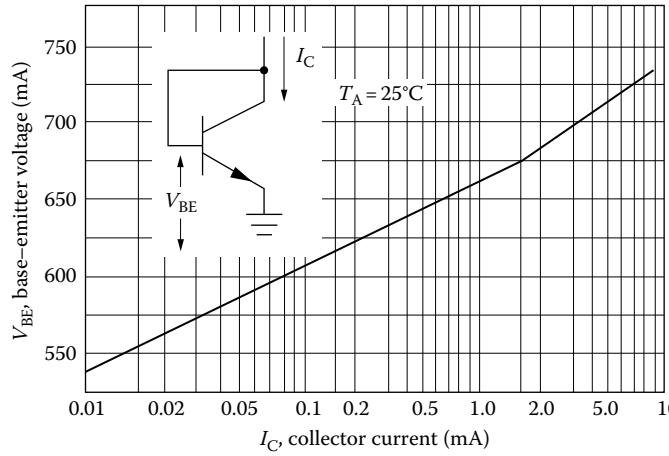


FIGURE 71.4 Base–emitter voltage versus collector–emitter current.

These readings are made with a constant (collector) current of 0.1 mA, passing through the device to minimize the effect of self-heating of the junction. When the constant current applied is larger than 0.1 mA, the effect of self-heating in the device must be taken into account. The variation of the V_{be} with current is shown in Figure 71.4 [2].

71.2 Thermal Properties of Semiconductors: Defining Equations

A constant forward current supplied through an ideal silicon *pn* junction produces a forward voltage drop, V_F [3,4]:

$$V_F = V_{be} = \left(\frac{kT}{q}\right)\ln\left(\frac{I_F}{I_s}\right) \tag{71.1}$$

where

k is the Boltzmann constant (1.38×10^{-23} J K^{-1})
T is the temperature (K)
q is the charge of electron (1.6×10^{-19} C)
I_F is the forward current (A)
I_s is the junction's reverse saturation current (A)

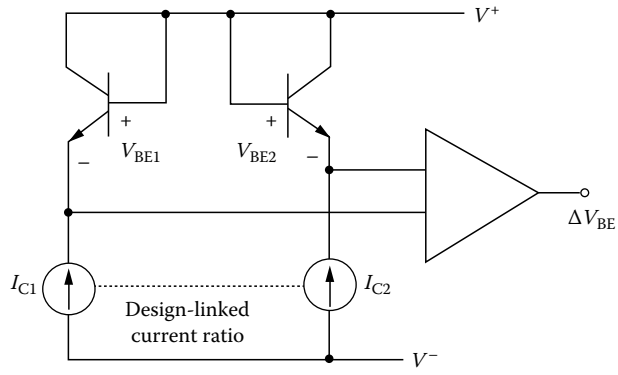


FIGURE 71.5 Differential pair formed by two *pn* junctions. The transistors are diode connected to form a temperature sensor independent of variations in source current.

For constant I_s, the junction voltage (V_{be}) would be directly proportional to absolute temperature. Unfortunately, I_s is temperature dependent and varies with the cube of absolute temperature. As a result, V_F has an overall temperature coefficient of approximately –2 mV °C^{-1}.

To reduce the temperature variation, a *bandgap reference* is formed based on two adjacent and essentially identical-behavior transistors with proportional emitter area designed in an IC process. The two base–emitter junctions are biased with different current densities (I/A), but the ratio of current densities is essentially constant over the operating temperature range (–55 °C to +150 °C). The following equation shows how the differential voltage (ΔV_{be}) is related to the current (I) and emitter area (A) of the respective transistors:

$$V_{be1} - V_{be2} = \Delta V_{be} = \left(\frac{kT}{q}\right)\ln\left(\frac{I_1/A_1}{I_2/A_2}\right) \tag{71.2}$$

The differential voltage appearing at the output can be amplified as shown in Figure 71.5 and used as a direct indication of absolute temperature. Additional circuitry can eliminate the offset voltage at 0 °C and provide an output in degrees Celsius or degrees Fahrenheit.

The ability to obtain temperature sensing using semiconductor processing techniques has two significant consequences: (1) semiconductor processing and IC design can be used to improve the temperature sensor's performance for specific applications, and (2) temperature sensors can be integrated within other ICs to obtain additional features. The next two sections explain these approaches.

71.3 Integrated Temperature Sensors

Once a temperature sensor can be manufactured using semiconductor processing techniques, a number of shortcomings of the sensor can be corrected by additional circuitry integrated into the sensor or by using circuit techniques external to the sensor. The linearity improvement, addition of precision voltage references, precision voltage amplifiers, and digital output for direct interface to a microcontroller (MCU) are among the enhancements possible. Furthermore, resistance-measuring circuitry (i.e., RTD sensors) or cold junction compensation (i.e., thermocouple sensors) are not required. Three ICs and one external circuit example are discussed. With any of these temperature-sensing techniques, many system-level circuit designs are produced that use temperature sensing for system monitoring or control. These ICs include fan temperature, battery temperature, computer temperature controls, and more.

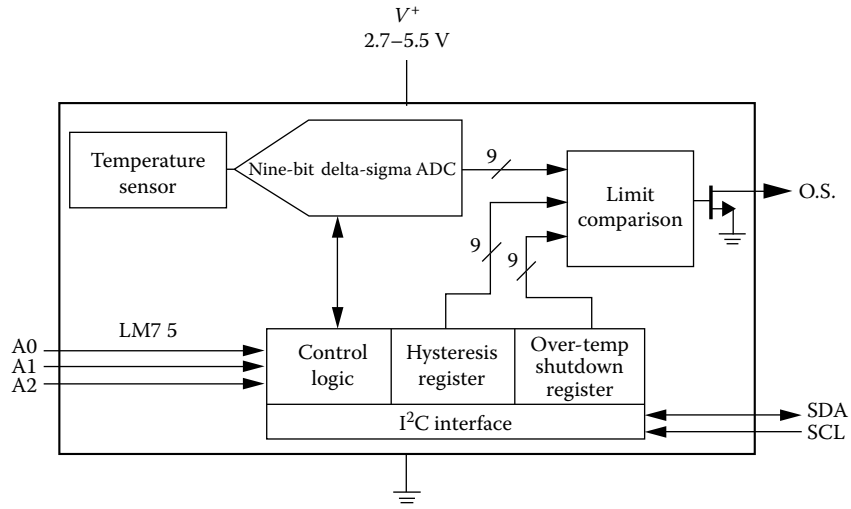


FIGURE 71.6 Block diagram of a monolithic digital-output temperature sensor. (Courtesy of National Semiconductor Corp., Santa Clara, CA.)

71.3.1 Integrated Digital Temperature Sensor

A *monolithic* (one piece of silicon) semiconductor junction temperature sensor (LM75 from National Semiconductor, now part of Texas Instruments) that incorporates several features, including a digital output, is shown in Figure 71.6 [3]. The analog signal of the temperature sensor is converted to digital format by an on-board sigma-delta converter [3]. Digital communication is provided directly to a host microcontroller through a serial two-wire interface. The sensor has a software-programmable set point that can be used to terminate the operation of the controller or implementing protection [5]. To avoid false triggering, a user-programmable number of comparisons (up to six successive over-temperature occurrences) can be implemented.

Eight different sensors can be operated on the bus. Resolution is ±1/2 °C and the accuracy is ±2% from –25 °C to +100 °C. The sensor consumes 250 μA during operation and only 10 μA in sleep mode.

More recent versions can be found at the suppliers' websites as indicated near the end of this section.

71.3.2 Analog-Output Integrated Temperature Sensor

Another approach to integrated temperature sensing is shown in Figure 71.7 [4]. The TMP-1 resistor-programmable temperature controller features a 5 mV $°C^{-1}$ output high and low set points and over- and under-temperature output. A low-drift voltage reference is also included in the 70 mil × 78 mil (2.76 mm × 3.07 mm) design. Figure 71.8 shows a photomicrograph of the silicon die. The TMP-1 is specified for operation between –55 °C and +125 °C, with ±1 °C accuracy over the entire range.

More recently, the DS600 is a ±0.5 °C accuracy analog-output temperature sensor. Other more recent versions can be found at the suppliers' websites as indicated near the end of this section.

71.3.3 Digital-Output Temperature Sensor

A direct-to-digital temperature sensor has been designed for multi-drop temperature-sensing applications [6]. A unique serial number is etched onto each device. The 64-bit read-only memory identifies the temperature of a particular sensor in a measurement system, with several sensors providing readings from different locations. The signal can be transmitted for distances up to 300 m.

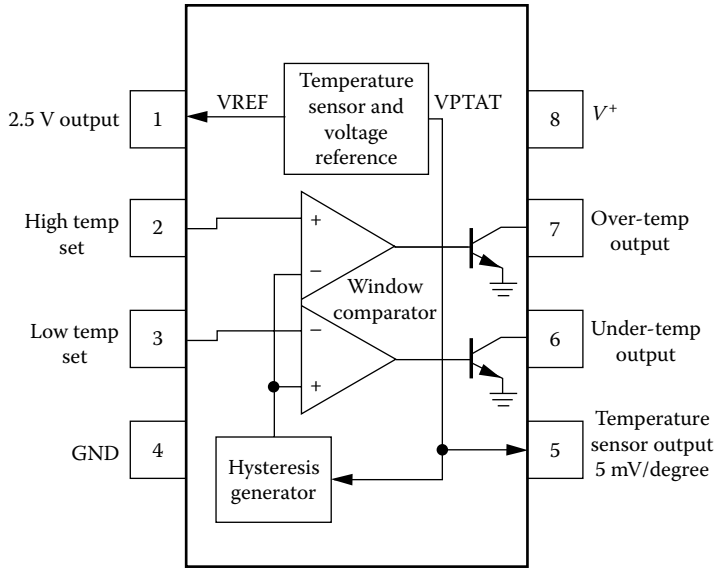


FIGURE 71.7 Block diagram and pinout of TMP-1 monolithic, programmable temperature controller. (Courtesy of Analog Devices, Inc., Norwood, MA.)

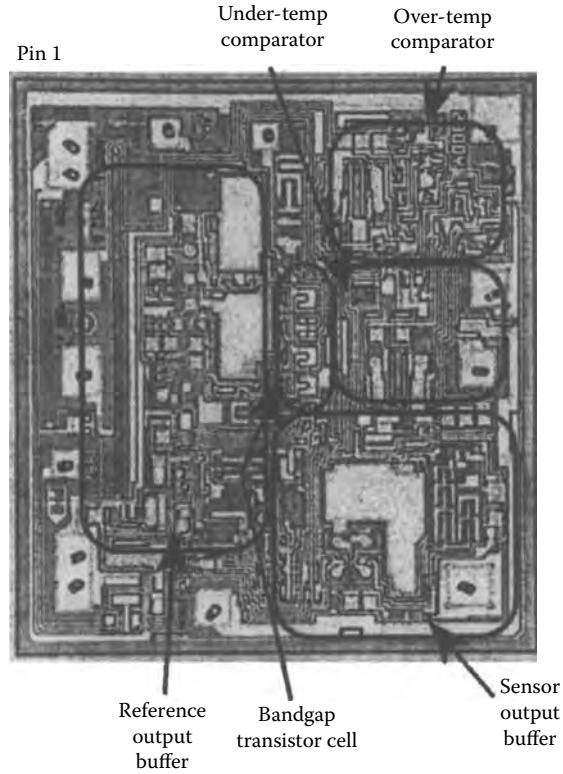


FIGURE 71.8 Detail of TMP-1 die. Note the relative size of the bandgap transistor cell to the other circuitry included in the monolithic device. (Courtesy of Analog Devices, Inc., Norwood, MA.)

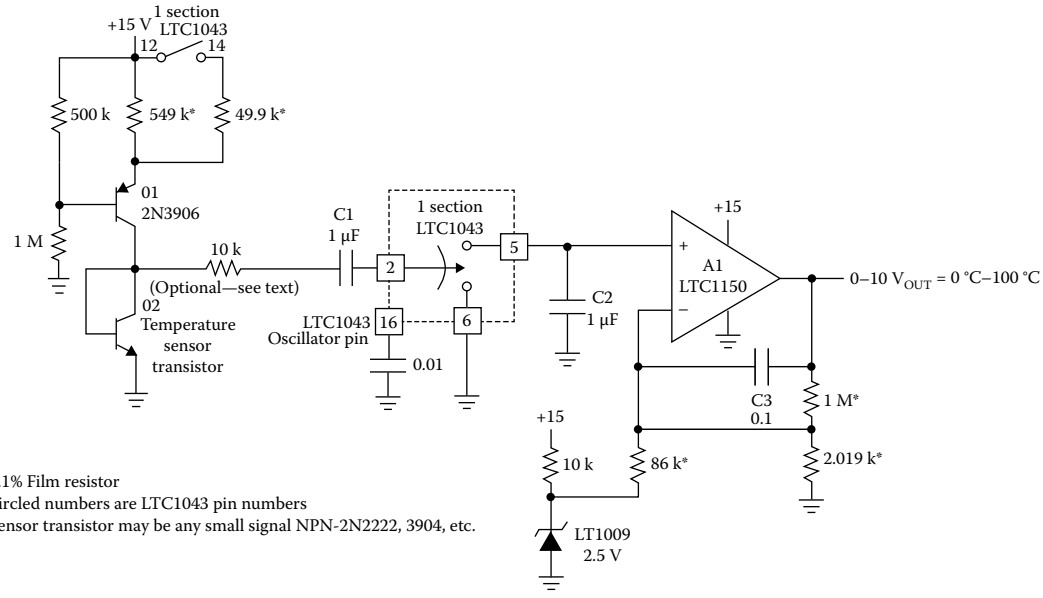


FIGURE 71.9 External circuitry provides ΔV_{be}-based thermometer that does not require calibration. (Courtesy of Linear Technologies Corporation, Milpitas, CA.)

The temperature sensor operates from −55 °C to 125 °C with the power supplied from the data line. The measurement is resolved in 0.5 °C increments as a 9-bit digital value with the conversion occurring within 200 ms. User-definable alarm settings are included in the device.

Digital-output sensors use different communication techniques, including

- 1-Wire
- 2-Wire, I²C, SMBus™
- 3-Wire, SPI™
- PWM
- SMAART wire interface

71.3.4 External Circuitry Eliminates Calibration

External circuitry can be designed to utilize the inherent sensing capability of low-cost transistors without incurring the additional cost of integrated circuitry or the need to calibrate each device. The circuit shown in Figure 71.9 uses an IC (LTC1043 from Linear Technology Corporation) to provide a 0–10 V output from 0 °C to 100 °C, with an accuracy of ±1 °C using any common small signal transistor as the temperature sensor [7]. The circuitry establishes a ΔV_{be} versus current relationship that is constant regardless of the V_{be} diode's absolute value. Substituting different transistors from multiple sources showed a variation of less than 0.4 °C.

71.4 Common Applications of Semiconductor Sensing

Common applications for semiconductor junction thermometers include thus:

- Medical equipment and instrumentation
- Handset and notebook personal computers
- Portable and battery-powered applications

- Power-supply temperature monitoring
- Computer peripheral thermal protection
- Battery management
- Office machines
- Electromechanical device temperatures
- Industrial controls
- Test equipment
- Food processing
- HVAC systems
- System temperature compensation
- Board-level temperature sensing
- Electronic thermostats

71.4.1 Other Applications of Semiconductor Sensing Techniques

Several semiconductor parameters vary linearly over the operating temperature range. Power MOSFETs used to switch high levels of current (typically several amperes) at voltages that can exceed 500 V provide an example of these characteristics. As shown in Figure 71.10 [8], the gate threshold voltage of a power MOSFET changes from 1.17 to 0.65 times its 25 °C value when the temperature increases from −40 °C to 150 °C. Also the breakdown voltage of the power MOSFET varies from 0.9 to 1.18 times its value at 25 °C over the same temperature range (Figure 71.11) [8]. These relationships are frequently used to determine the junction temperature of a semiconductor component in actual circuit operation during the design phase (see Section 71.6). External package level temperature measurements can be many degrees lower than the junction temperature, especially during rapid, high-energy switching events. The actual junction temperature and the resulting effect on semiconductor parameters must be taken into account for the proper application of semiconductor devices.

Polysilicon diodes (and resistors) that are isolated from the power MOSFET can be produced as part of the semiconductor manufacturing process with minor process modifications. The diodes can be used as temperature-sensing elements in an actual application [9]. The thermal sensing that is performed by the polysilicon elements is a significant improvement over power device temperature sensing that is performed by an external temperature-sensing element. By sensing with polysilicon diodes, the sensor can be located close to the center of the power device near the source bond pads where the current density

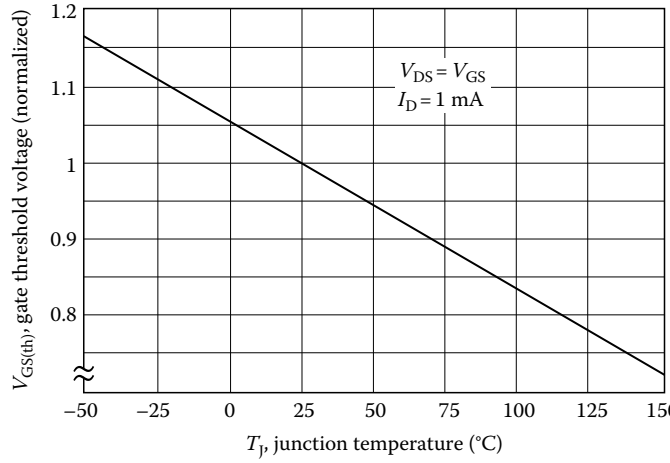


FIGURE 71.10 Power MOSFET's gate threshold variation versus temperature.

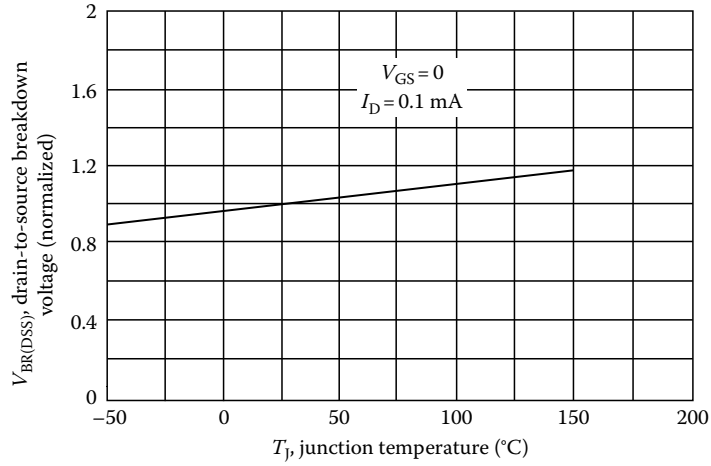


FIGURE 71.11 Power MOSFET's breakdown voltage variation versus temperature.

is the highest, and consequently, the highest die temperature occurs. The thermal conductivity of the oxide that separates the polysilicon diodes from the power device is two orders of magnitude less than that of silicon. However, because the layer is thin, the polysilicon element offers an accurate indication of the actual peak junction temperature.

A power FET that incorporates temperature-sensing diodes is shown in Figure 71.12 [9]. By monitoring the output voltage when a constant current is passed through the integrated polysilicon diode(s), an accurate indication of the maximum die temperature is obtained. A number of diodes are actually provided in the design. A single diode in this design has a temperature coefficient of 1.90 mV °C^{-1}. Two or more can be placed in series if a larger output is desired. For greater accuracy, the diodes can be

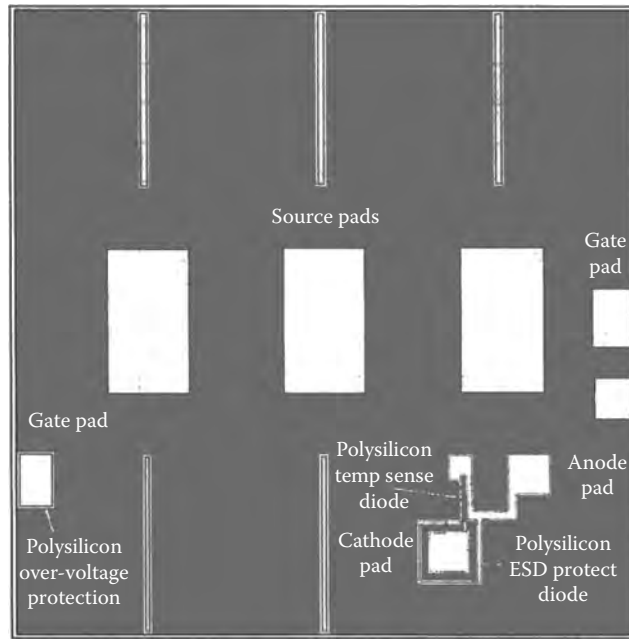


FIGURE 71.12 Photomicrograph of temperature sensor integrated in power MOSFET. Note the relative size of the temperature sensor compared to the total area of the power MOSFET and the source pads that allow attachment of 15 mil (0.60 mm) aluminum wire. (Courtesy of Motorola, Inc., Schaumburg, IL.)

trimmed during wafer-level testing by blowing fusible links made from polysilicon. The response time of the diodes is less than 100 μs, which has allowed the device to withstand a direct connection across an automobile battery with external circuitry providing shutdown prior to device failure. The sensing capability also allows the output device to provide an indication (with additional external circuitry) if the heatsinking is not proper when the unit is installed in a module or if a change occurs in the application that would ultimately cause a failure.

71.5 Temperature Sensing in Power ICs for Fault Protection and Diagnostics

Sensing for fault conditions, such as a short circuit, is an integral part of many smart power (or power) ICs. The ability to obtain temperature sensors in the semiconductor process provides protection and diagnostics as part of the features of these devices. The primary function of the power IC is to provide a microcontroller-to-load interface for a variety of loads. In multiple output devices, sensing the junction temperature of each device allows the status of each device to be provided to the microcontroller (MCU), and, if necessary, the MCU can shut down a particular unit that has a fault condition.

A *smart power IC* can have multiple power drivers integrated on a single monolithic piece of silicon [10]. Each of these drivers can have a temperature sensor integrated to determine the proper operating status and shut off only a specific driver if a fault occurs. Figure 71.13 shows an eight-output driver that independently shuts down the output of a particular driver if its temperature is excessive (i.e., between 155 °C and 185 °C) [10].

The octal serial switch adds independent thermal sensing through over-temperature detection circuitry to the protection features. Faults can be detected for each output device, and individual

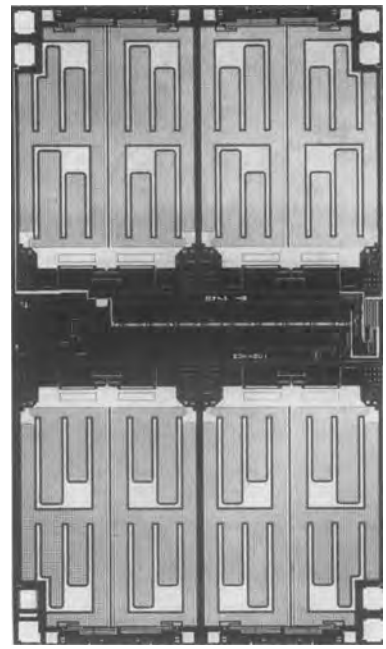


FIGURE 71.13 Photomicrograph of eight-output power IC. Note that the area of the eight-output devices (two located at each corner of the die) are considerably larger than the circuitry in the center, and top, and bottom that provides the temperature sensing, signal conditioning, and other control features. (Courtesy of Motorola, Inc., Schaumburg, IL.)

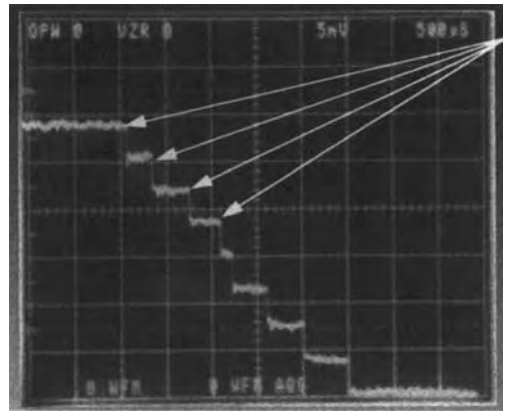


FIGURE 71.14 Independent thermal shutdown of an eight-output power IC. (Courtesy of Motorola, Inc., Schaumburg, IL.)

shutdown can be implemented. In a multiple output power IC, it is highly desirable to shut down only the device that is experiencing a fault condition and not all of the devices that are integrated on the power IC. With outputs in various physical locations on the chip, it is difficult to predict the thermal gradients that could occur in a fault situation. Local temperature sensing at each output, instead of a single global temperature sensor, is required.

As shown in Figure 71.14, the eight outputs of the device with individual temperature sensors can be independently shut down when the thermal limit of 170 °C is exceeded [10]. All of the outputs were connected to a 16 V supply at a room temperature ambient. A total current of almost 30 A initially flowed through the device. Note that each device turns off independently. The hottest device turns off first. Variations can result from differences in current level and thermal resistance. As each device turns off, the total power dissipation in the chip decreases, and the devices that are still on can dissipate heat more effectively.

Connecting directly to the battery is a hard short that could have been detected by current limit circuitry. However, a soft short is below the current limit, but exceeds the power-dissipating capability of the chip, and can be an extremely difficult condition to detect. Soft shorts require over-temperature sensing to protect the IC from destructive temperature levels.

The over-temperature condition sensed by the power IC could mean that the device turns itself off to prevent failure in one case, and in another situation, a fault signal provides a warning to the MCU but no action is taken, depending on the fault circuit design. The remaining portion of the system is allowed to function normally. With the fault conditions supplied to the MCU, an orderly system shutdown can be implemented. Integrated temperature sensing is essential to provide this type of protection in a multiple-output power IC.

71.6 Reliability Implications of Temperature to Electronic Components

The effect of temperature on electronic components and their successful application in electronic systems is one of the issues that must be addressed during the design of the system. Temperature affects the performance and expected life of semiconductor components. Mechanical stress created by different coefficients of thermal expansion can cause failures in thermal cycling tests (air-to-air) or during thermal shock (water-to-water) transitions.

The typical failure rate for semiconductor component can be expressed by the Arrhenius equation [9]:

$$\lambda = Ae^{-\text{ø}/KT} \tag{71.3}$$

where

λ is the failure rate
A is the constant
ø is the activation energy (eV)
K is the Boltzmann constant (8.62×10^{-5} eV K^{-1})
T is the junction temperature (K)

The failure rate of semiconductor components is typically stated to double for every 10°C–15°C increase in operating (i.e., junction) temperature. However, increased testing and design improvements have minimized the failures due to specific failure mechanisms.

One of the temperature-related parameters that must be taken into account during the design phase of a power switch is the transient thermal response, which is designated as $r(t)\ R\text{ø}_{JC}$, where $r(t)$ represents the normalized transient thermal resistance. The value of $r(t)$ is determined from the semiconductor manufacturer's data sheet using duty cycle and pulse duration used in the application. This reduced level of the thermal resistance (see Section 71.6.1), based on the transistor operating in a switching mode and being off for a period of time, approaches the dc level within a second. Excessive temperatures can be generated quickly and must be detected within milliseconds to prevent failure.

71.6.1 Junction Temperature Measurement

In a semiconductor, the change in temperature is directly related to the power dissipated through the thermal resistance. The steady-state dc thermal resistance junction-to-case, $R\text{ø}_{JC}$, is defined as the temperature rise per unit power above an external reference point (typically the case). The relationship is shown in Equation 71.4 [8].

$$R\text{ø}_{JC} = \frac{\Delta T}{P_D} \tag{71.4}$$

where

ΔT is the junction temperature minus the case temperature (°C)
P_D is the power dissipated in the junction (W)

The semiconductor device or silicon die is typically enclosed in a package that prevents a direct measurement of the junction temperature. The junction temperature is measured indirectly by measuring the case temperature, T_C; the heatsink temperature, T_S; for those higher-power applications that require a heatsink; the ambient temperature, T_A; and a temperature-sensitive electrical parameter of the device.

The first step of the process requires calibrating the temperature-sensitive parameter. Using a bipolar power transistor as an example, the base–emitter forward voltage is measured and recorded with a low calibration current (I_M) flowing through the device that is low enough to avoid self-heating (typically between 2 and 10 mA) and yet sufficiently high to be in the linear range of the forward voltage curve. The procedure is performed at room and elevated temperatures, typically 100 °C.

After calibration, a power switching fixture (such as Figure 71.15) is used to alternately apply and interrupt the power to the device under test (DUT) [8]. The on portion is long (typically several milliseconds) and the off portion is short (only a few 100 μs), so the temperature of the case is stabilized and junction cooling is minimal. The transistor is operated in its active region, and the power dissipation is varied by adjusting the I_E and/or V_{CE} until the junction is at the calibration temperature. This point is known by measuring V_{BE} during the time that I_M is the only current flowing.

When the V_{BE} value equals the value on the calibration curve, the junction temperature is at the calibration temperature. Measurements of V_{BE}, T_C, and I_E allow the thermal resistance for the device to

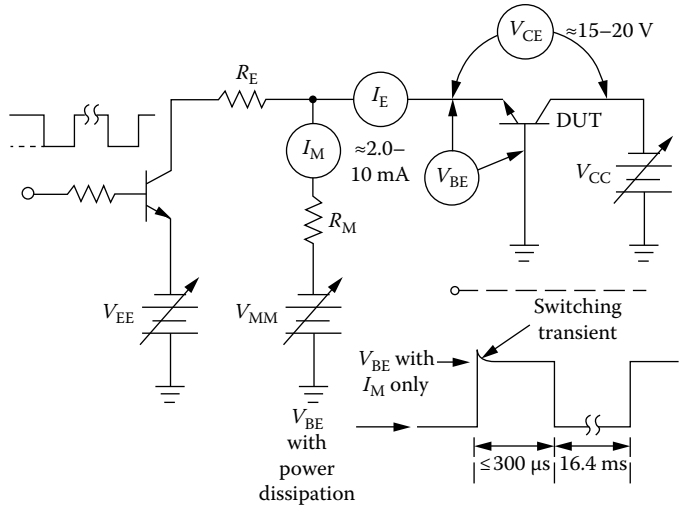


FIGURE 71.15 Example of steady-state thermal resistance test circuit for a bipolar power transistor. Power is applied for 16.4 ms and interrupted for ≤300 μs to measure the V_{BE}.

be calculated using Equation 71.4. Since $R\text{ø}_{JC}$ is a constant, subsequent measurements of T_C, V_{CE}, and I_E under different operating conditions can be used to calculate the junction temperature to keep the device within its safe operating range in the actual application. For devices with different electrical characteristics, such as the power MOSFETs discussed earlier, other parameters that have a linear relationship to temperature are used for calibration and measurement.

71.7 Semiconductor Temperature Sensor Packaging

Temperature sensors that are manufactured using semiconductor technology are typically packaged in packages common to the semiconductor industry. These include metal can (TO-99), ceramic, and more commonly available plastic (SOT-23, 8-lead DIP, TO-92, 8-lead SOIC, etc.) packages. These packages are designed for circuit board solder attachment that can be either through-hole or surface-mount technology. As a result, package form factors can be considerably different from packages for temperature sensors manufactured using other technologies. Figure 71.16 shows examples of five available silicon temperature sensor packages.

More recent packages include SC70, four-ball wafer chip-scale package, micro small outline package, and dual, flat no-lead packages.

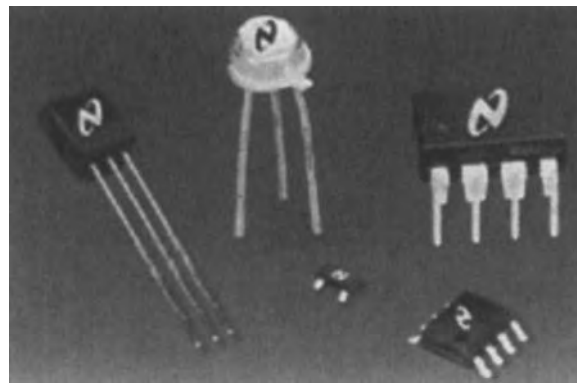


FIGURE 71.16 Plastic TO-92, TO-99 metal can, 8-lead DIP, 8-lead SOIC, and TinyPak™ SOT-23 plastic packages. (Courtesy of National Semiconductor Corp., Santa Clara, CA.)

Defining Terms

Bandgap reference: Forward-biased emitter junction characteristics of adjacent transistors used to provide an output voltage with zero temperature coefficient.

Die: An unpackaged semiconductor chip separated from the wafer.

Embedded temperature sensor: A temperature sensor included within an integrated circuit.

Integrated circuit: A multiplicity of transistors, as well as diodes, resistors, capacitors, and so on, on the same silicon die.

Junction: The interface at which the conductivity type of a material changes from *p* type to *n* type.

Junction voltage: The voltage drop across a forward-biased *pn* interface in a transistor (V_{be}) or diode.

Junction temperature: The temperature of the *pn* interface in a transistor or diode.

Monolithic (integrated circuit): Constructed from a single piece of silicon.

Power IC or smart power IC: Hybrid or monolithic (semiconductor) device that is capable of being conduction-cooled, performs signal conditioning, and includes a power control function such as fault management and/or diagnostics.

Self-heating: Temperature rise within a (semiconductor) device caused by current flowing in the device.

Soft short: An excessive load condition that causes excessive temperature but is below the current limit of a device.

Thermal resistance: The steady-state dc thermal resistance junction-to-case, $R\text{ø}_{JC}$, is the temperature rise per unit power above an external reference point (typically the case).

TinyPak: A trademark of Texas Instruments.

Wafer: A thin slice of semiconductor material, commonly silicon in 400 mm, 300 mm, used to manufacture or fabricate integrated circuits.

Vendors and Suppliers

Analog Devices, Inc.
http://www.analog.com
AD590
AD22100

Maxim Integrated Products
http://www.maxim-ic.com
DS600
DS18S20
MAX31826

Microchip Technology, Inc.
http://www.microchip.com
MCP9808
MCP9701
TC620
TC74A0
TC07
TC6501

Rohm Semiconductor
http://www.rohm.com/products/lsi/sensor/temperature_sensor/174sel2/

STMicroelectronics
http://www.st.com/internet/analog/class/1577.jsp

Texas Instruments, Inc.
http://www.ti.com
http://www.ti.com/lsds/ti/analog/temperature_sensor.page?DCMP=hpa_bestbets&HQS=hpa-sc-bb-temphome
MSP430
TMP275
TMP100

SMSC
http://www.smsc.com/index.php?tid=157

References

1. J. Carr, *Sensors and Circuits*, Englewood Cliffs, NJ, PTR Prentice-Hall, 1993.
2. Pressure sensor device data DL200/D Rev. 1, Phoenix, AZ, Motorola, 1994.
3. K. Lacanette, Using IC temperature sensors to protect electronic systems, Sensors, Peterborough, NH, 14(1), pp. 28–34, January 1997.
4. R. Wegner and H. Hulsemann, New family of monolithic temperature sensor and controller circuits present challenges in maintaining temperature measurement accuracy, *Proceedings of Sensors Expo West*, Anaheim, CA, February 8–10, 1994.
5. W. Schweber, Temperature sensors fill different needs, *EDN*, 41(20), 1996.
6. R. Downs, Digital thermometer IC simplifies distributed sensing, *Electronic Products*, 56, 1995, available at http://www.electronicproducts.com/PrintArticle.aspx?ArticleID=9763 (accessed on December 23, 2012).
7. J. Williams, High performance signal conditioning for transducers, *Proceedings of Sensors Expo West*, San Jose, CA, March 2–4, 1993.
8. TMOS power MOSFET transistor data DL135/D Rev. 4, Phoenix, AZ, Motorola Semiconductor.
9. R. K. Jurgen (ed.), *Automotive Electronics Handbook,* New York, McGraw-Hill, 1999.
10. R. Frank, *Understanding Smart Sensors*, Boston, MA, Artech House, 2013.
11. J. R. Gyorki, Designing with semiconductor temperature sensors, Sensor Tips website: http://www.sensortips.com/temperature/designing-with-semiconductor-temperature-sensors (accessed on March 14, 2013).



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

72

Noncontact Thermometers

72.1 Thermal Radiation (Physical Laws) **72**-1
Planck's Law • Wien's Law • Stefan–Boltzmann Law • Kirchhoff's Law
72.2 Emissivity **72**-4
72.3 Blackbody **72**-7
Cavity Effect • Practical Blackbodies
72.4 Detectors for Thermal Radiation **72**-9
Classification • Quantum Detectors • Thermal Detectors
72.5 Pyrometers **72**-14
Disappearing Filament Pyrometer • Two-Color Pyrometer
72.6 IR Thermometry **72**-17
Errors
72.7 Special Applications **72**-18
Measuring Semiconductor Materials • Medical Thermometers
References **72**-20

Jacob Fraden
Fraden Corporation

There are two ways of measuring temperature—contact and noncontact. In the former case, a temperature probe is placed into an intimate physical connection with the object, while in the latter case, a temperature probe is nearly disconnected from the object, except one link—the optical. In this chapter, we examine noncontact ways of measuring temperature by exchange of electromagnetic radiation from the visible to far infrared spectral ranges.

72.1 Thermal Radiation (Physical Laws)

In any object, every atom and every molecule exist in a perpetual motion. When an atom moves, it collides with other atoms and transfers to them part of its kinetic energy, thus losing some of its own energy in this perpetual bouncing. On the other hand, an atom that had a smaller kinetic energy, after a collision, gains some. Afterward, the material body consisting of such agitated atoms reaches the energetic equilibrium where all atoms, while not vibrating with exactly the same intensity and frequency, still may be described by an average kinetic energy. Such an average kinetic energy of the agitated particles is represented by the *absolute temperature*, which is measured in degrees kelvin. In other words, what we commonly call temperature is a measure on a macroscale of the microscopic atomic motion.

According to laws of electrodynamics, a moving electric charge (all atoms are made of electric charges) is associated with a variable electric field. An electric field around a charged particle changes while the particle moves. A variable electric field, in turn, produces an alternating magnetic field.

When the magnetic field changes, it results in a coupling with its variable electric field, and so on. Thus, a thermally agitated particle becomes a source of the electromagnetic field that propagates outwardly with the speed of light and is called *thermal radiation*. It is called thermal because it directly depends on the kinetic energy of vibrating particles. Since this radiation is predominantly situated in the optical region, it is governed by the laws of optics—it can be reflected, filtered, focused, and so on. And also, it can be used as a measure of the object's temperature.

Electromagnetic waves originated from the mechanical movement of particles may be characterized by their magnitude and wavelengths (or frequencies). Both of these characteristics relate to temperature, that is, the hotter the object, the shorter the wavelength and the stronger the magnitude. Very hot objects radiate in the visible portion of the spectrum of wavelengths between 0.4 μm (blue) and 0.7 μm (red). For example, a filament in an incandescent lamp is so hot (around 3000 K) that it radiates bright visible light. If the lamp is controlled by a dimmer, the light intensity may be reduced by turning the knob, and one may observe that the dimmed light becomes more yellowish and reddish and finally visibly disappears. This means that the filament cools down and its heat intensity drops and the radiated spectrum shifts toward longer wavelengths. Near the end of the dimmer control, the filament is still quite hot, yet we cannot see it because it emanates light in the invisible infrared spectral range, the wavelengths somewhere longer than 0.8 μm. Cooler objects radiate light in the near, mid, and far infrared spectral ranges that we cannot see. For instance, thermal radiation that is emanated from a human skin primarily is situated at wavelengths between 5 and 15 μm in the mid and far infrared ranges and is not visible to human eyes; otherwise, we all would glow in dark (sick people with fever would look a bit brighter).

If we imagine that all atomic vibrations for some mysterious reason would stop, no electromagnetic radiation will be emanated from the object. Such an imaginable but impossible event is characterized by the infinitely cold temperature, which is called the *absolute zero*.

Since temperature is a measure of the average atomic kinetic energy, it is logical to assume that one may determine the object's temperature by measuring intensity of the emanated thermal radiation or its spectral characteristics. This presumption is the basis for noncontact temperature measurements, which is known by the various names depending on the application: infrared thermometry, optical pyrometry, radiation thermometry, and so on. *Pyrometry* is derived from the Greek word *pyr* (*πυρ*), which means *fire*, and thus is more appropriate for measuring hotter temperatures. For the lower temperatures, the name *infrared thermometry* is used interchangeably with *radiation thermometry*.

72.1.1 Planck's Law

A relationship between the magnitude of radiation at a particular wavelength λ and absolute temperature T is rather complex and governed by Planck's law, which was discovered in 1901. It establishes a radiant flux density W_λ as a power of electromagnetic radiation per unit of wavelength:

$$W_\lambda = \frac{\varepsilon(\lambda) C_1}{\pi \lambda^5 (e^{C_2/\lambda T} - 1)} \tag{72.1}$$

where

$\varepsilon(\lambda)$ is the emissivity of an object
$C_1 = 3.74 \times 10^{-12}$ W cm^2 and $C_2 = 1.44$ cm K are the constants
e is the base of natural logarithms

Spectral densities have clearly pronounced, albeit not sharp maxima at different temperatures (Figure 72.1).

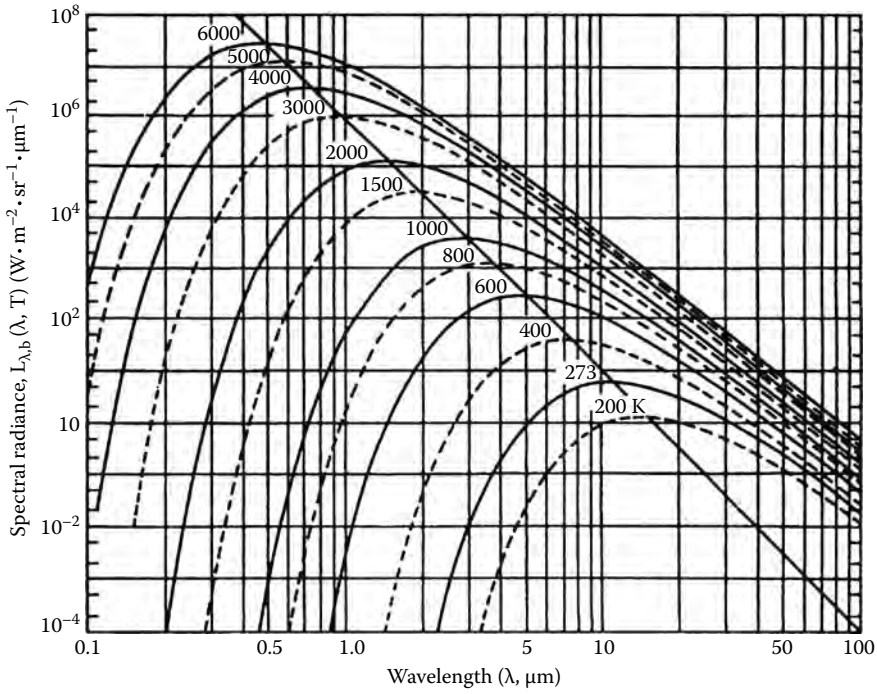


FIGURE 72.1 Spectral densities calculated within solid angle of 1 sr for radiation of a blackbody source.

72.1.2 Wien's Law

Equation 72.1 does not lend itself to a simple mathematical analysis and thus was approximated by a simplified version that is known as Wien's law:

$$W_\lambda = \frac{C_1}{\pi}\varepsilon(\lambda)\lambda^{-5}e^{-\frac{C_2}{\lambda T}} \tag{72.2}$$

Since temperature is a statistical representation of an average kinetic energy, it determines the highest probability for the particles to vibrate with a specific frequency and to have a specific wavelength. This most probable wavelength can be found from Wien's law by equating a first derivative of Equation 72.2 to zero. The result of the calculation is a peak wavelength near which most of the radiant power is concentrated:

$$\lambda_m \approx \frac{2898}{T}, \tag{72.3}$$

where

λ_m is in μm
T in kelvin

Wien's law states that the higher the temperature, the shorter wavelength. This formula also defines the midpoint of the spectral response of a pyrometer or infrared thermometer.

72.1.3 Stefan–Boltzmann Law

Theoretically, a thermal radiation bandwidth is infinitely wide. Yet most of the emanated power is situated within quite a limited bandwidth. Also filtering properties of the real-world windows used in the instruments allow passing only limited wavelengths and thus must be accounted for. In order to

determine the total radiated power (light flux) that is limited within a particular bandwidth, Equation 72.1 or (72.2) has to be integrated within the limits from λ_1 to λ_2:

$$\Phi_b = \frac{1}{\pi}\int_{\lambda_1}^{\lambda_2} \frac{\varepsilon(\lambda)C_1\lambda^{-5}}{e^{C_2/\lambda T}-1}. \tag{72.4}$$

This integral can be resolved only numerically or by approximation. For a narrow bandwidth (λ_1 and λ_2 are close to one another), the solution may be approximated by a high-order parabola:

$$\lambda_b = gT^x, \tag{72.5}$$

where

g is constant
$x \approx (12/\lambda_2)(1200/T)$

For example, in the visible portion of the spectrum at $\lambda_2 \approx 0.7$ μm and for temperatures near 2000 K, the approximation is a 10th order parabola.

For a very broad bandwidth where $\lambda_2 \to \infty$ or practically, when the range between λ_1 and λ_2 embrace well over 50% of the total radiated power, an approximation is a fourth-order parabola, which is known as the *Stefan–Boltzmann law*

$$\Phi_b = A\varepsilon\sigma T^4 \tag{72.6}$$

where

$\sigma = 5.67 \times 10^{-8}$ W/m^{2}K^4 (Stefan–Boltzmann constant)
A is the area from which the radiation is emanated
emissivity, ε, is assumed to be wavelength independent

It is seen that with the increase in temperature, intensity of the thermal radiation Φ_b grows very fast due to the fourth power of T. To improve accuracy and reduce noise, practically all radiation thermometers employ the magnitude of thermal radiation power within a relatively wide bandwidth and seldom, if ever, consider the wavelength shift resulted from changing temperatures.

72.1.4 Kirchhoff's Law

As it follows from Equations 72.5 and 72.6, besides a temperature, the magnitude of the radiation also is a function of the surface property called *emissivity*, ε, that is measured on a scale from 0 to 1. Emissivity is a ratio of the radiant flux that is emanated from a surface to the maximum theoretically possible flux that would be emanated from an ideal emitter (blackbody) having the same temperature. *Reflectivity*, ρ, and *transparency*, γ, also on a scale from 0 to 1, show what portion of the incident light is reflected and passed through, respectively. There is a fundamental equation that connects these three characteristics:

$$\varepsilon + \gamma + \rho = 1 \tag{72.7}$$

All these factors are wavelength dependent. Equation 72.7 that is known as Kirchhoff's Law indicates that any one of the three properties of a material may be changed only on the expense of the others. As a result, if at a particular wavelength the object is opaque ($\gamma = 0$), reflectivity, ρ, and emissivity, ε, are connected by a simple relationship: $\rho = 1 - \varepsilon$. This makes a mirror a good reflector but a poor emitter. That is, an ability to absorb light (not to reflect or transmit) is equal to an ability of emanating thermal radiation.

72.2 Emissivity

The emissivity of a material is a function of its dielectric constant and, subsequently, refractive index n. As it was mentioned earlier, emissivity is generally wavelength dependent (see Figure 72.2). For example, a white sheet of paper is very much reflective in the visible spectral range and emits no visible light. In the mid and

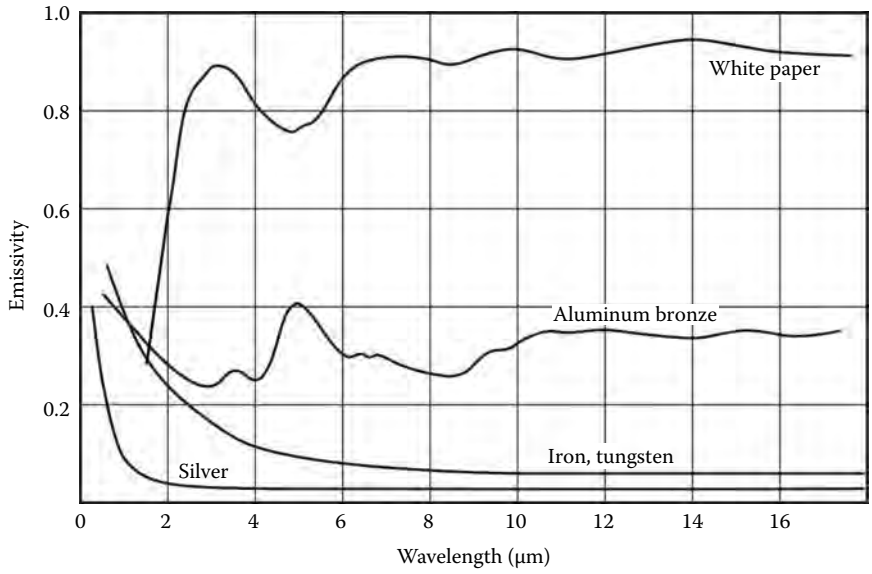


FIGURE 72.2 Wavelength dependence of emissivities of some materials. (Used with permission from Springer Science+Business Media: *Handbook of Modern Sensors*, 4th edn., 2010, Fraden, J., 2010.)

far infrared spectral ranges, its reflectivity is low and emissivity is high (about 0.92), thus making a paper a good emitter of thermal radiation at room temperatures. For many practical purposes, in the infrared thermometry, emissivity may be considered constant but seldom is accurately known. Since emissivity is an ability to emanate electromagnetic radiation, knowing its value for a particular object of measurement is important for the accuracy of a radiation thermometer, unless a special design negates the effects of emissivity on accuracy, as described later.

As a rule, emissivities of dielectrics are high and of bare nonoxidized metals, low. Due to a high emissivity of dielectrics, a dielectric emanates strong thermal radiation and thus lends itself for easy and accurate noncontact temperature measurement. On the other hand, a radiation thermometry of nonoxidized metals is difficult, due to a low level of the emanated thermal radiation flux. Table 72.1 gives typical emissivities of some opaque materials in a temperature range between 0 °C and 100 °C.

Unlike most solid bodies, gases in many cases are transparent to thermal radiation. When they absorb and emit radiation, they usually do so only in certain narrow spectral bands. Some gases, such as N_2, O_2, and others of the nonpolar symmetrical molecular structure, are essentially transparent at low temperatures, while CO_2, H_2O, and various hydrocarbon gases radiate and absorb to an appreciable extent. When thermal radiation light enters a layer of gas, its absorption has an exponential decay profile over the layer's thickness, governed by *Beer's law*

$$\frac{\Phi_x}{\Phi_0} = e^{-\alpha_\lambda x}, \tag{72.8}$$

where

Φ_0 is the incident thermal flux
Φ_x is the flux at thickness x
α_λ is the spectral coefficient of absorption

The ratio given earlier is called a monochromatic transmissivity at a specific wavelength λ. If gas is nonreflecting (usually it is), then from Kirchhoff's law, its emissivity is defined as

$$\varepsilon_\lambda = 1 - \gamma_\lambda = 1 - e^{-\alpha_\lambda x}. \tag{72.9}$$

TABLE 72.1 Typical Emissivities of Different Materials (from 0°C to 100°C)

Material	Emissivity	Material	Emissivity
Blackbody (ideal)	1	Green leaves	0.88
Cavity radiator	0.990–0.999	Ice	0.96
Aluminum (anodized)	0.70	Iron or steel (rusted)	0.70
Aluminum (oxidized)	0.11	Nickel (oxidized)	0.40
Aluminum (polished)	0.05	Nickel (unoxidized)	0.04
Aluminum (rough surface)	0.06–0.07	Nichrome (80Ni–20Cr) (oxidized)	0.97
Asbestos	0.96	Nichrome (80Ni–20Cr) (polished)	0.87
Brass (dull tarnished)	0.61	Oil	0.80
Brass (polished)	0.05	Silicon	0.64
Brick	0.90	Silicone rubber	0.94
Bronze (polished)	0.10	Silver (polished)	0.02
Carbon-filled latex paint	0.96	Skin (human)	0.93–0.96
Carbon lamp black	0.96	Snow	0.85
Chromium (polished)	0.10	Soil	0.90
Copper (oxidized)	0.6–0.7	Stainless steel (buffed)	0.20
Copper (polished)	0.02	Steel (flat rough surface)	0.95–0.98
Cotton cloth	0.80	Steel (ground)	0.56
Epoxy resin	0.95	Tin plate	0.10
Glass	0.95	Water	0.96
Gold	0.01–0.02	White paper	0.92
Gold-black	0.98–0.99	Wood	0.93
Graphite	0.7–0.8	Zinc (polished)	0.04

It should be emphasized that since gases absorb only in narrow bands, emissivity and transmissivity (transparency) must be specified separately for any particular wavelength. For instance, water vapor is highly absorptive (emissive) at wavelengths of 1.4, 1.8, and 2.7 μm while very transparent at 1.6, 2.2, and 4 μm.

All nongaseous nonmetals are good diffusive emitters of thermal radiation with a remarkably constant emissivity within a solid angle of about ±70°. Beyond that angle, emissivity begins to decrease rapidly to zero with the angle approaching 90°. Near angles of 90°, emissivity is very low. A typical calculated graph of the directional emissivity of nonmetals into air is shown in Figure 72.3a.

Metals behave quite differently. Their emissivities greatly depend on surface finish. Generally, polished metals are poor emitters within the solid angle of ±70° while their emissivity increases at larger angles (Figure 72.3b). Oxidized metals start behaving more like dielectrics with the increased thickness of oxides. To increase emissivity of a metal surface for a noncontact temperature measurement, it may be treated, for example, painted, oxidized, or anodized.

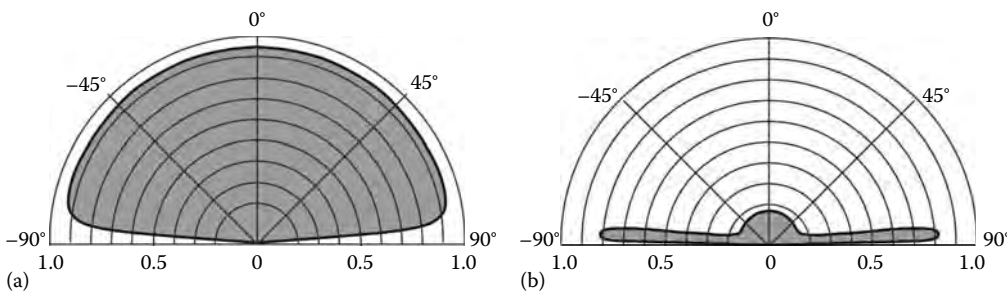


FIGURE 72.3 Spatial emissivities for (a) a nonmetal and (b) a polished metal. (Used with permission from Springer Science+Business Media: *Handbook of Modern Sensors*, 4th edn., 2010, Fraden, J., 2010.)

72.3 Blackbody

By definition, the highest possible emissivity is unity. It is attributed to the so-called *blackbody*—an ideal emitter of electromagnetic radiation. If an object is opaque ($\gamma = 0$) and totally nonreflective ($\rho = 0$), according to Equation 72.7, it becomes an ideal emitter and absorber of electromagnetic radiation. The name blackbody implies its appearance at normal room temperatures—indeed, since absolutely no light is reflected back, it does look very black. In reality, an ideal blackbody does not exist: it is just a fantasy, and any real object with a less-than-unity emissivity sometimes is called a *graybody*.

A practical blackbody (ε is about 0.99 or higher) is an approximation to an ideal blackbody and is an essential tool for calibrating and verifying accuracy of infrared thermometers.

72.3.1 Cavity Effect

To design a high-accuracy blackbody, a "cavity effect" is put to work. This effect appears when electromagnetic radiation escapes via an opening of a cavity [1]. For this purpose, a cavity means a void of a generally irregular shape whose inner wall temperature is uniform over the entire surface. Emissivity of the cavity opening dramatically increases and approaches unity at any wavelength, as compared with a flat surface that typically is much less than unity and is wavelength dependent. The cavity effect is especially pronounced when the cavity inner walls have higher emissivity. To explain the cavity effect, let us consider a cavity with a nonmetal inner surface. All nonmetals are diffuse emitters. Also they are diffuse reflectors. These are the key properties of the cavity wall surface for the cavity effect. We assume that temperature and surface emissivity of the cavity inner surface are homogeneous over an entire inner area. Figure 72.4 illustrates a simple conical cavity. Let us select area a on the cavity wall. In accordance with the Stefan–Boltzmann Law, the ideal emitter (blackbody) would emanate from a, the photon flux being $\Phi_0 = a\sigma T_b^4$. However, the surface is not an ideal emitter and has the actual emissivity ε_b.

The flux that is emitted by other parts of the cavity and incident on area a is also equal to Φ_b (the object is thermally homogeneous and assumes a uniform radiation within a hemisphere). A substantial portion of that incident flux Φ_b is absorbed by the surface of area a, without changing its temperature. The remaining part of the incident flux is diffusely reflected:

$$\Phi_\rho = \rho\Phi_b = (1 - \varepsilon_b)\varepsilon_b\Phi_0 \tag{72.10}$$

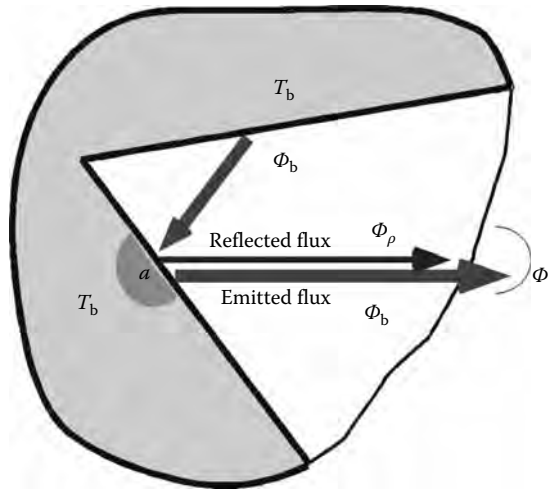


FIGURE 72.4 Cavity effect enhances emissivity by combining the emanated and reflected fluxes.

and a combined flux effusing from the area *a* is the sum of two fluxes:

$$\Phi = \Phi_b + \Phi_\rho = \varepsilon_b \Phi_0 + (1 - \varepsilon_b)\varepsilon_b \Phi_0 = (2 - \varepsilon_b)\varepsilon_b \Phi_0 \tag{72.11}$$

As a result, the *effective emissivity* may be expressed as

$$\varepsilon_e = \frac{\Phi}{\Phi_0} = (2 - \varepsilon_b)\varepsilon_b \tag{72.12}$$

To illustrate the effective emissivity, we assume that the cavity walls are painted with organic paint (a paint color in the visible spectrum makes no difference) having a surface emissivity $\varepsilon_b = 0.92$ in the mid and far infrared ranges. Then according to Equation 72.12, the effective emissivity at the cavity opening is $\varepsilon_e = 0.9936$. It follows from the earlier text that due to the cavity effect, a perceived (effective) emissivity of a cavity is equal to the wall emissivity magnified by a factor of $(2 - \varepsilon_b)$. For the cavity effect to work, the effective emissivity must be attributed only to the cavity opening (aperture) from which the radiation escapes. If the probe of a radiation thermometer is inserted inside the cavity facing the walls directly (blocking the flux from other walls of the cavity), the cavity effect is substantially diminished and emissivity will drop to that of a wall surface.

72.3.2 Practical Blackbodies

To achieve emissivity well over 0.990, there are four fundamental requirements to a blackbody design:

1. The inner wall temperatures must be uniform within ±0.01 °C all over the cavity surface. Thus, in most practical cases, the blackbody cavity should be fabricated of a metal of high thermal conductivity.
2. Temperature accuracy of the inner walls should be measured with uncertainty no greater than 0.02 °C.
3. A ratio of the inner wall area to the aperture area should be at least 100:1 and preferably much higher.
4. The cavity should have no inner surfaces that are parallel to the aperture. A blackbody serves as a light trap for any light incident on the aperture, thus there must be multiple inner reflections of the light beam arriving from the aperture until it is nearly extinguished. This allows to closely satisfying Kirchhoff's law for a blackbody, that is, no light is reflected back or passing through.

A practical blackbody can be fabricated in several ways. Copper and aluminum are the best choices for the cavity body material—thanks to their high thermal conductivity that helps equalizing temperatures of the cavity walls. Figure 72.5 shows simplified cross-sectional views of two practical blackbodies. One (A) is a solid-state blackbody having thermoelectric elements (heat pumps) that provide either heating or cooling to the cavity body. This solid-state blackbody has a useful temperature range from about –20 °C to +120 °C. To control temperature, the imbedded precision temperature sensor is connected to the heat pump control circuit (not shown). The inner shape of the cavity is partly conical to increase the number of reflections and prevent a direct reflection of light that may come to the aperture from outside, so the light will be trapped inside and not reflected back via the aperture. The entire inner surface is treated for a relatively high surface emissivity, typically over 0.9.

The aperture potentially may act as an entry for undesirable ambient air. Cooler or warmer air may disturb the thermal uniformity inside the cavity resulting in excessively high uncertainty of the radiated flux. To reduce that problem, the cavity has a slightly positive air pressure where air is continuously pumped inside through the air ducts. Before entering the cavity, air passes through the narrow channels inside the cavity walls and acquires temperature near that of the blackbody.

There is another problem with the cavity opening—a temperature of the front aperture plate. Since for the convenience of use the plate is relatively thin, it is very difficult to maintain it at exactly the same

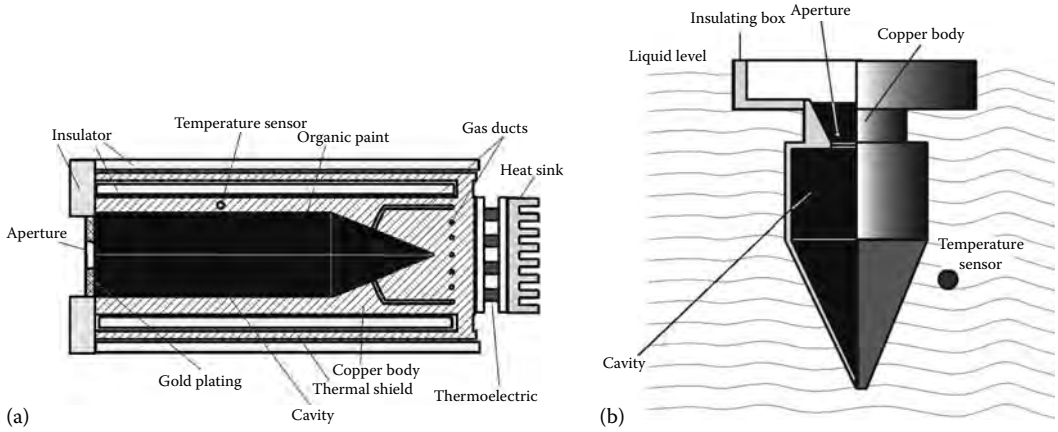


FIGURE 72.5 Solid-state blackbody with (a) thermoelectric elements and (b) water-immersed blackbody. A gold mirror creates a virtual cavity surface.

temperature as the cavity walls, thus violating the fundamental requirement (#1 given earlier)—the cavity temperature uniformity. To resolve this issue, the wall of the aperture plate that faces the cavity is mirror polished and coated with a thin layer of gold—the best reflector of infrared energy [2]. Since gold is an extremely poor emitter, the plate temperature becomes irrelevant. On the other hand, a very high reflectivity ($\sigma \approx 0.99$ for polished gold) bounces back the cavity thermal radiation. Thus, the gold mirror creates an extended virtual cavity that nearly doubles the size of the real cavity (see requirement #3 given earlier), improving the walls/aperture ratio.

To achieve a high temperature uniformity of the cavity walls, an active thermal shield is placed outside of the cavity body (see Figure 72.5a). This blackbody was tested to have an emissivity of at least 0.998.

Another example is an immersed blackbody (Figure 72.5b). The cavity body has relatively thin walls (few millimeters) and is immersed into a stirred liquid bath (often water) whose temperature is precisely controlled by the heating/cooling devices. The liquid assures uniform temperature distribution around the cavity with a typical thermal instability on the order of ±0.02 °C. The inner surface of the cavity is coated with organic paint. The aperture of the cavity is relatively small, just to allow the tested probe insertion, and can be also covered with a gold-coated plate as described earlier.

72.4 Detectors for Thermal Radiation

72.4.1 Classification

Generally speaking, there are two types of sensors (detectors) known for their capabilities of responding to thermal radiation within the spectral range from the visible to near infrared to far infrared, that is, approximately from about 0.4 to 40 μm of wavelengths. The first type is the *quantum* detectors and the second—*thermal* detectors.

72.4.2 Quantum Detectors

Quantum detectors (photovoltaic and photoconductive devices) rely on the interaction of individual photons with a crystalline lattice of semiconductor materials. Photovoltaic devices are the photodiodes and phototransistors, and photoconductive devices are the photoresistors. Their operations are based on the photo effect that was discovered by A. Einstein and brought him the Nobel Prize. In 1905, he made a remarkable assumption about the nature of light that at least under

certain circumstances, its energy was concentrated into localized bundles, later named photons. The energy of a single photon is given by

$$E = h\nu \tag{72.13}$$

where

ν is the frequency of light

$h = 6.63 \cdot 10^{-34}$ J · s (or $4.13 \cdot 10^{-15}$ eV) is Planck's constant derived on the basis of the wave theory of light

When a photon strikes a surface of a conductor, it may result in the generation of a free electron—meaning that the conductor becomes light sensitive.

The periodic lattice of crystalline materials establishes allowed energy bands for electrons that exist within that solid. The energy of any electron within a pure material must be confined to one of these energy bands that may be separated by gaps or ranges of forbidden energies.

In isolators and semiconductors, the electron must first cross the energy band gap in order to reach the conduction band. Hence, conductivities of these materials are therefore many orders of magnitude lower than that in conductors. For isolators, the band gap is usually 5 eV or more, whereas for semiconductors, the gap is considerably less.

When the photon of frequency ν_1 strikes a semiconductive crystal, its energy shall be high enough to separate the electron from its site in the valence band and push it through the band gap into a conduction band at a higher energy level. In that band, the electron is free to serve as a current carrier—that is what is required for the light sensing. The deficiency of an electron in the valence band creates a hole that also serves as a current carrier. This is manifested in the reduction of specific resistivity of the material.

The energy gap serves as a photon energy threshold, below which the material is not light sensitive.

Examples of the quantum detectors operating in the visible and near infrared ranges are the popular and inexpensive photodiodes, phototransistors, and photoresistors. They are quite sensitive because in that spectral range, energies of the photons (2 eV and more) are sufficiently high to cross a relatively wide energy gap. These detectors are used for noncontact measurement of hot objects that emanate photons of higher energies.

For cooler temperatures, that is—for the smaller photon energies (longer wavelengths)—the narrower band gap semiconductors are required, otherwise the "cooler" photons are just too weak to cross the gap. However, even if a special quantum detector has a sufficiently narrow energy band gap, at room temperatures, its own intrinsic noise is much higher than a signal originated by incoming photons. In other words, the detector will generate its own spurious "signal" (noise) and respond to its own temperature that will wash out signals from the external objects. This is the reason why for operation in the mid and far infrared spectral ranges, a quantum detector not only must have a sufficiently narrow energy gap, but its temperature also has to be lowered to the level where its own intrinsic noise is reduced to an acceptable level, that is, it must be cryogenically cooled down. Depending on the required sensitivity and operating wavelength, the following crystals are typically used for the cryogenically cooled photoresistors (see in Figure 72.6 the curves just below the dotted lines): lead sulfide (PbS), indium arsenide (InAs), germanium (G), lead selenide (PbSe), and mercury-cadmium-telluride (HgCdTe).

The sensor cooling allows it to respond to longer wavelengths. However, the response speeds of PbS and PbSe become less with cooling. Methods of cooling include a Dewar cooling using dry ice, liquid nitrogen, liquid helium, or thermoelectric coolers operating on the Peltier effect.

72.4.3 Thermal Detectors

Another class of infrared radiation detectors is called *thermal detectors*. Contrary to quantum detectors that respond to individual photons, thermal detectors respond to heat resulted from the absorption of thermal radiation by the surface of a sensing element that is capable of converting a heat level into an

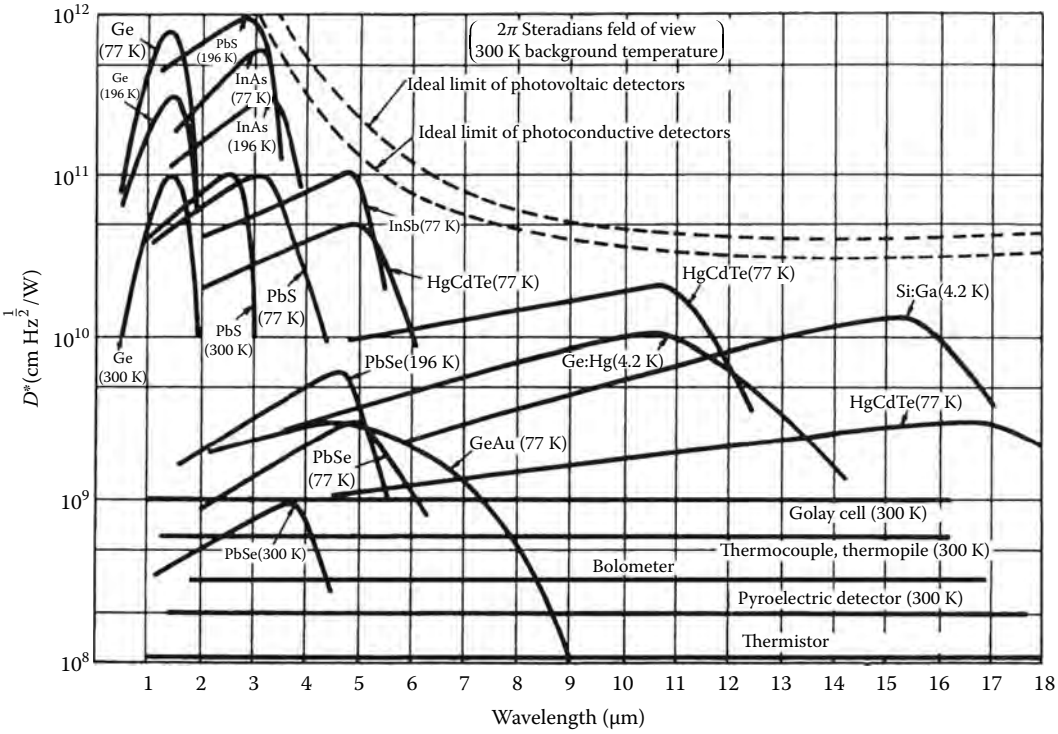


FIGURE 72.6 Operating ranges for the quantum and thermal infrared detectors.

electrical signal. The absorbed heat raises the temperature of the surface, and the resulting temperature increase becomes a measure of the incident thermal radiation. The thermal detectors are much less sensitive than quantum detectors because to elevate temperature of a detector to an appreciable level, a great number of photons is required, while a quantum detector produces an electron (and sometimes many) for a single absorbed photon. The thermal detectors are indicated at the lower portion of Figure 72.6. Note that unlike the quantum detectors, responses of the thermal detectors are flat over all spectral ranges.

Let us analyze the behavior of a thermal radiation detector when it is exposed to an object. Figure 72.7 illustrates an object having a temperature T and an emissivity ε. The object faces a detector having the temperature T_s and the emissivity ε_s. Note that the object radiates thermal flux Φ_{bo} toward the sensor while the sensor radiates its own flux Φ_{so} toward the object. This bidirectional radiation flow goes

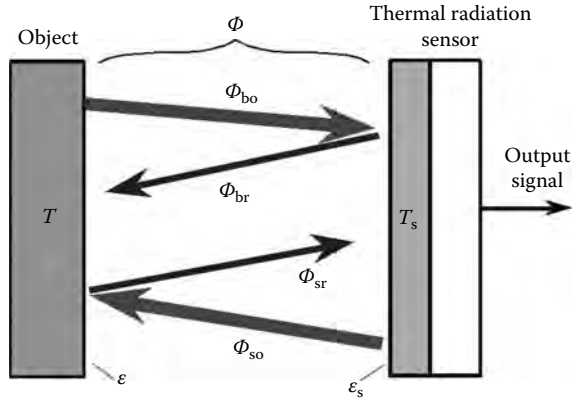


FIGURE 72.7 Radiative heat exchange between an object and a thermal radiation sensor.

through a coupling channel having a cross-sectional area A. According to the Stefan–Boltzmann law (Equation 72.6), these fluxes can be expressed as

$$\begin{aligned} \Phi_{bo} &= A\varepsilon\sigma T^4 \\ \Phi_{so} &= -A\varepsilon_s\sigma T_s^4 \end{aligned} \tag{72.14}$$

A negative sign indicates an opposite direction of flux from the detector Φ_{bo} with respect to flux Φ_{bo} from the object.

A larger part of flux Φ_{bo} is absorbed by the detector having an absorptivity (emissivity) ε_s:

$$\Phi_{bs} = A\varepsilon\varepsilon_s\sigma T^4 \tag{72.15}$$

While the detector absorbs energy from the object, it also loses energy toward the object. Thus, in turn, the object absorbs the detector's flux propagating at the opposite direction:

$$\Phi_{sb} = -A\varepsilon\varepsilon_s\sigma T_s^4 \tag{72.16}$$

These two fluxes combine into the final net flux that is absorbed or liberated by the detector's surface:

$$\Phi = \Phi_{bs} + \Phi_{sb} = A\varepsilon\varepsilon_s\sigma\left(T^4 - T_s^4\right). \tag{72.17}$$

This equation establishes a connection between thermal power, Φ, absorbed or liberated by the sensor, and the absolute temperatures of the object and the sensor, T and T_s, respectively. This net flux causes temperature of the sensing element in the detector to shift accordingly. Depending on which is larger, T or T_s, the sensing element temperature will go up or down. The temperature change can be used to determine the net thermal flux. To measure the temperature shift of the detector, the sensing element has an imbedded temperature sensor that produces the output electric signal proportional to the net flux, Φ, described by Equation 72.17.

The detector's output signal in the form of voltage can be defined as

$$V = g\Phi = Ag\varepsilon\varepsilon_s\sigma\left(T^4 - T_s^4\right), \tag{72.18}$$

where g is the conversion factor of the incident thermal flux to the sensor output voltage. The conversion factor generally is a function of the sensor's temperature T_s, albeit for a well-designed sensor, this dependence is small.

A temperature (in kelvin) of the object can be calculated from the inverted Equation 72.18:

$$T_b = \sqrt[4]{T_{sc}^4 + \frac{V}{Ag\varepsilon\varepsilon_s\sigma}}, \tag{72.19}$$

Note that in a sensor's relatively narrow temperature range, the denominator is a constant whose value is found during the thermometer calibration, while T_{sc} and T_b are the *computed* reference sensor and object's temperatures, respectively. These computed temperatures may somewhat differ from the true temperature T_s and T of Equation 72.18 because of the inevitable measurement uncertainties.

If we consider T being close to T_s, Equation 72.19 can be simplified to

$$T_b \approx T_{sc} + \frac{V}{4Ag\varepsilon\varepsilon_s\sigma T_{sc}^3} \tag{72.20}$$

Thus, to compute the object's temperature, one needs to know two variables—the detector's temperature and the detector's output voltage. The calibration factor (denominator) is found during the thermometer calibration.

There are a number of thermal detectors that can be used for measuring the sensor's temperature change described by the second summand. Such a detector is imbedded into the IR sensor's surface. One example is a thermistor-type sensor known as a *bolometer* [3]. But by far, the most popular is the so-called *thermopile* [4]. A thermopile belongs to a class of passive infrared detectors, that is, it generates electric output in relation to the absorbed thermal flux with no need for any additional source of energy (unlike the bolometers). Its operating principle is the same as that of a thermocouple (see the description of a thermoelectric effect in Chapter 70 in this book or in [1]). A thermopile consists of many serially connected thermocouples. Originally, it was invented by Joule to increase the output signal of a thermoelectric sensor. He connected several thermocouples in a series and thermally joined together all their "hot" and separately all "cold" junctions. Nowadays, a thermopile has a different configuration. Its prime application is a detection of thermal radiation. Note that a thermopile detector produces a signal that is nearly proportional to a temperature gradient between the "hot" and "cold" junctions but, unlike a bolometer, is not responsive to an absolute temperature. That is why a thermopile must be appended by an absolute temperature sensor, for example, a thermistor or a semiconductive temperature sensor. This additional absolute sensor sometimes is called a *reference temperature sensor*. It is essential for monitoring temperature T_s if Equation 72.19 or 72.20 is to be employed for computing temperature T_b of an object.

A cutout view of a thermopile IR sensor is shown in Figure 72.8 [1]. The sensor consists of a frame having a relatively large thermal mass. Typically, it is fabricated of silicon. To further stabilize it thermally, the frame is usually bonded to a metal header. The frame is the place where the "cold" junctions and reference temperature sensor are positioned. The frame supports a thin membrane (on the order of 1 μm) whose thermal capacity and thermal conductivity are small, so its temperature can change quickly and by a sufficient amount. The membrane is the area where all "hot" junctions are positioned. The membrane absorbs incident thermal radiation, thus its temperature shifts from T_s and subsequently is detected by the thermopile junctions. Even though a thermopile is just one part of the thermal IR sensor, the entire sensor is called a "thermopile."

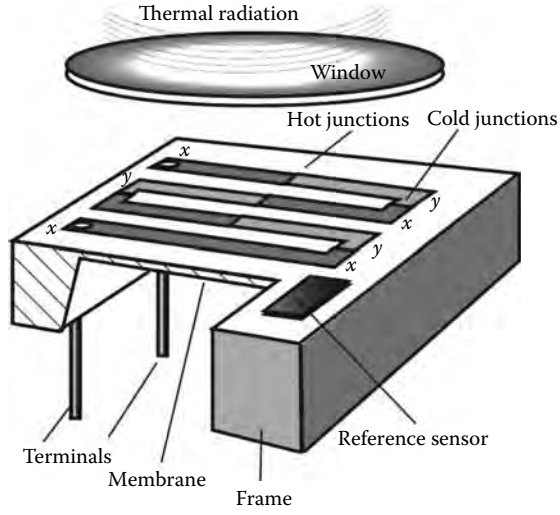


FIGURE 72.8 Thermopile sensor. The "hot" junctions are deposited on the membrane while the "cold" junctions on the frame. (Used with permission from Springer Science+Business Media: *Handbook of Modern Sensors,* 4th edn., 2010, Fraden, J., 2010.)



(a) (b)

FIGURE 72.9 (a) Thermal imaging camera from Heimann Sensor GmbH and (b) an example of a thermal image.

The best performance of a thermopile is characterized by a high sensitivity, fast response, and low noise, which may be achieved if the junction materials have high thermoelectric coefficients, low thermal conductivity, and low volume resistivity. Besides, the junction pairs should have the thermoelectric coefficients of the opposite signs. This dictates selection of the materials. Methods of construction of the junctions may differ to some extent, but all incorporate vacuum deposition techniques and evaporation masks to apply the thermoelectric materials on thin membranes [4]. The number of junctions varies from twenty to several hundreds. The membrane is given an additional coat on the top to increase the surface emissivity (absorptivity). The output voltage of a thermopile is nearly proportional to the net thermal flux when the IR flux is absorbed. As opposed to quantum detectors that require a cryogenic cooling, the thermopile sensors can operate at room temperatures. Since thermopiles produce rather small output voltages, the signal become susceptible to external interferences. This is the reason why these sensors often incorporate various interface electronic circuits in the same packaging [5]. The examples are the thermopile modules HID L1x FL5.5 from Heimann Sensor GmbH (Germany) and TMP006 from Texas Instruments (U.S.A.).

A thermopile sensor essentially is a "single-pixel" detector. However, multiple thermopile groups can be deposited onto a single membrane to form a multifaceted sensor array. This allows not only to measure temperature but also to create a thermal image of an object formed on the membrane by a special focusing infrared lens. Each pixel represents temperature of a small area of the object's surface [6]. At the time of this writing, the best thermal imaging sensors having up to 64 × 62-element arrays with the corresponding focusing IR optics and electronics are produced by Heimann Sensor GmbH (Wiesbaden, Germany: www.heimannsensor.de). Figure 72.9 illustrates a small Heimann imaging camera operating at room temperature and an example of a thermal image it produces.

72.5 Pyrometers

72.5.1 Disappearing Filament Pyrometer

The other names for this instrument are the *optical pyrometer* and *monochromatic-brightness radiation thermometer*. It is a rather old technology [7]. This type of a pyrometer is considered the most accurate radiation thermometer for temperatures over 700 °C and efficient to temperatures up to 4200 °C. The lower limit is a result of human eye sensitivity within a specific wavelength range. This limit can be lowered when a human eye is replaced with a light-sensitive detector as explained later.

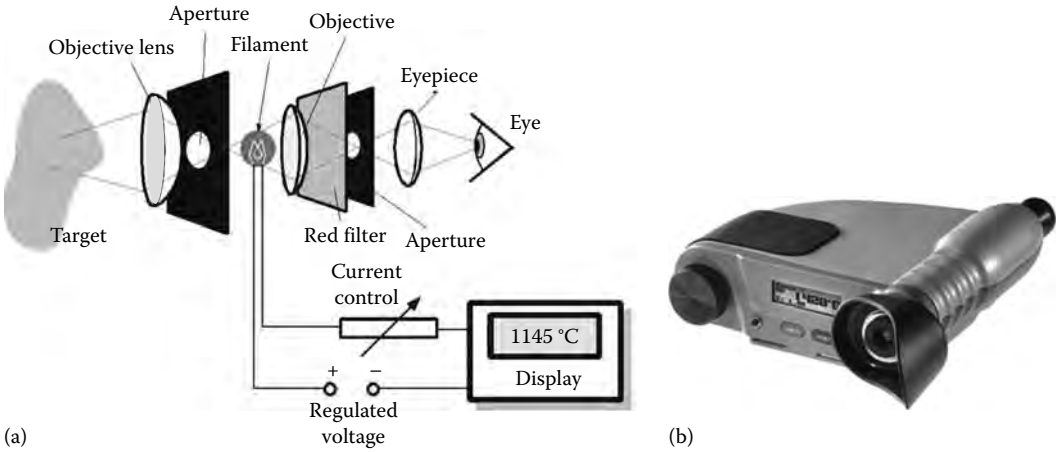


FIGURE 72.10 (a) Simplified block diagram of a disappearing-filament optical pyrometer. (b) A handheld DFP 2000 optical pyrometer. (Courtesy of Spectrodyne, Inc. www.spectrodyne.com.)

The operating principle of this thermometer is based on Planck's law (Equation 72.1 and Figure 72.1), which states that intensity and color of the surface change with temperature. The idea behind the design is to balance a radiation from a hot reference element having a known temperature against unknown radiation from a target. The pyrometer has a lens through which the operator views the target (Figure 72.10a).

An image of a tungsten filament (reference element) is superimposed on the image of a target. The filament is warmed up by electric current to glow. During calibration, the relationship between the heating current and the filament temperature was established by measuring brightness of a blackbody at known temperatures. The operator views the target through the eyepiece (see Figure 72.10b) and manually adjusts the heating current to the level when an image of the glowing filament visible in the foreground disappears, that is, when both the target and the filament have the same brightness and color. A color component complicates the measurement somewhat, so to remove this difficulty, a narrow-band red filter ($\lambda_f \approx 0.65$ μm) is inserted in front of an eyepiece. Therefore, the operator has to balance only the brightness of an image visible in red color. Another advantage of the filter is that emissivity ε_{bf} of the target needs to be known only at λ_f. There are several error sources in the pyrometer. One is difficulty in using it under dusty conditions. The other is known as the target emissivity. Fortunately, this is not a very large error since the error is given by

$$\frac{dT_b}{T_b} = -\frac{\lambda_f T_b}{C_2}\frac{d\varepsilon_{bf}}{\varepsilon_{bf}} \tag{72.21}$$

Thus, for a target at $T_b = 1000$ K, a 10% uncertainty in knowing emissivity of the target at the filter's wavelength results in only 0.45% (4.5 K) uncertainty in measured temperature. And of course the operator's ability to balance the filament brightness against the object may be a significant human factor.

The instrument may be further improved by removing an operator from the measurement process. Figure 72.11 shows an automatic version of the pyrometer where a rotating mirror tilted by 45° has a cutout sector that allows light from a target to pass through to the photo sensor. Such a mirror serves as a chopper that alternatively sends light to a photo sensor either from a target or from the filament. The microprocessor (μP) adjusts current through the filament to bring the optical contrast to zero. The optocoupler provides synchronization between the chopper and the microprocessor. In more advanced devices, instead of a chopper, a CCD receives an image of different portions of the field of view.

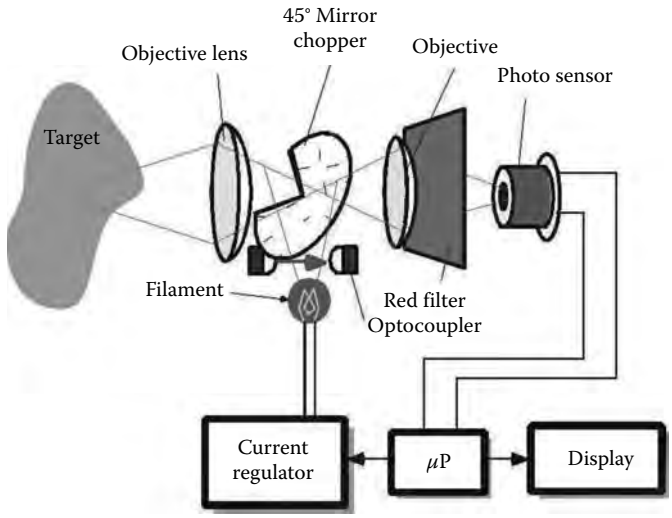


FIGURE 72.11 Block diagram of an automatic disappearing-filament optical pyrometer with a chopper.

72.5.2 Two-Color Pyrometer

Since emissivities of many materials, especially at high temperatures, are not known, measurement of a surface temperature may become impractical due to unacceptably large errors, unless the emissivity is excluded from the calculation. This can be accomplished by the use of a ratiometric technique in the so-called two-color radiation thermometer or *ratio thermometer*. In such a thermometer, the radiation is detected at two separate wavelengths λ_x and λ_y at which emissivities of the surface may be considered nearly the same. If the coefficients of transmission of the optical system at each wavelength respectively are λ_x and λ_y, then the ratio of two fluxes calculated for two wavelengths is

$$\phi = \frac{\Phi_x}{\Phi_y} = \frac{\gamma_x \varepsilon(\lambda_x)\frac{C_1}{\pi}\varepsilon(\lambda_x)\lambda_x^{-5}e^{-\frac{C_2}{\lambda_x T}}}{\gamma_y \varepsilon(\lambda_y)\frac{C_1}{\pi}\varepsilon(\lambda_y)\lambda_y^{-5}e^{-\frac{C_2}{\lambda_y T}}} \tag{72.22}$$

Since the emissivities are the same, $\varepsilon(\lambda_x) \approx \varepsilon(\lambda_y)$, after simple manipulations, Equation 72.22 can be rewritten for the computed temperature T_c, where ϕ represents a ratio of the thermal radiation sensor outputs at two wavelengths

$$T_c \approx C_2\left(\frac{1}{\lambda_y}-\frac{1}{\lambda_x}\right)\left(\ln\phi\frac{\gamma_y}{\gamma_x}\frac{\lambda_x^5}{\lambda_y^5}\right)^{-1} \tag{72.23}$$

It is seen that the equation for calculating temperature no longer depends on emissivity of the surface. Figure 72.12 shows a block diagram of a two-color pyrometer where an optical modulator is designed in the form of a disk with two filters. To eliminate a chopper, a CCD may be used instead of a single light sensor with a more complex signal processing [8].

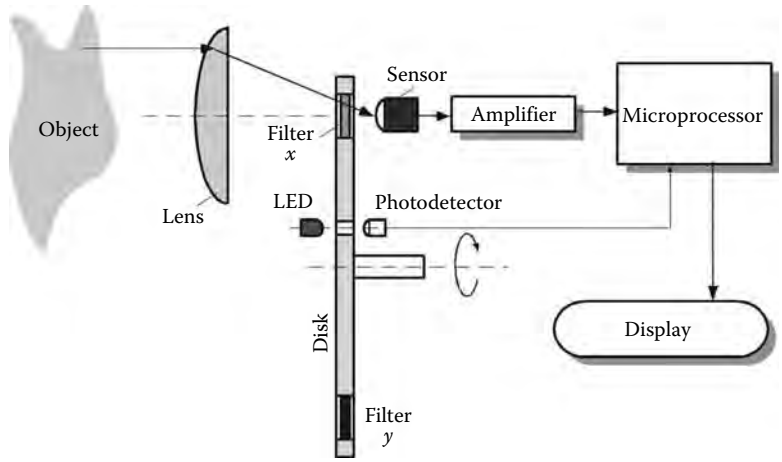


FIGURE 72.12 Block diagram of a two-color noncontact pyrometer.

72.6 IR Thermometry

72.6.1 Errors

To identify sources of errors in a noncontact temperature measurement by a thermal IR sensor, we need to take partial derivatives of Equation 72.20 for all factors that potentially may vary. For example, we take a partial derivative for the conversion factor *g*.

$$\frac{\partial T}{\partial g} \approx \frac{V}{4Ag^2\varepsilon\varepsilon_s\sigma T_{sc}^3} \tag{72.24}$$

It can be seen that the error is proportional to the IR sensor output signal V and it becomes smaller when emissivities of both the sensor and the object approach unity. On the other hand, if ε is small, error grows greatly. Since emissivities of metals are quite low and one never can be sure about their actual values, it is advisable, whenever practical, to coat a measured portion of a metal surface by a dielectric material, such as metal oxide or organic paint having known surface emissivity [9]. Alternatively, temperatures shall be taken not from a flat surface but rather from a hole or cavity in a metal object—taking an advantage of the cavity effect.

Another source of errors is a thermal instability of the IR sensor housing. It increases inaccuracy in measuring T_s, which is in the third power in Equation 72.24—meaning that even small variations in monitoring the sensor temperature will result in sizable errors. To minimize these errors, the IR sensors usually are encapsulated into metal housings having high thermal capacity and being well insulated from the environment (high thermal resistance). This helps stabilizing the temperature of the thermopile frame.

The optical coupling factor A is influenced by numerous variable components, such as transparency of the optical materials in the IR sensor and thermometer, angle of view, dust and scratches on the optical components, and so on.

A source of an error that is not part of the mathematical model of Equation 72.24 is a spurious heat source that may emanate its own thermal radiation either directly into the optical system of the IR thermometer or by means of reflection from the measured surface. Since no object is an ideal emitter (absorber), its reflectivity may have a significant value. For instance, an opaque object with emissivity 0.9 has a reflectivity of 0.1 (Equation 72.7), thus about 10% of radiation from a stray object in the vicinity of a measured surface is spuriously reflected toward the IR thermometer and may greatly disturb the measurement.

72.7 Special Applications

72.7.1 Measuring Semiconductor Materials

Temperature measurement of a semiconductor material during processing, such as growth of epitaxial films in vacuum chambers, always has been a difficult problem. Various process controls require accurate temperature monitoring, often without physical contact with the substrates. As a rule, heating of a substrate is provided by resistive heaters. The substrates are often loaded into the chambers through various locks and rotated during processing. Therefore, it is very difficult to attach a contact sensor, such as a thermocouple, to the wafer. And even if one does so, thermal gradient between the sensor and the substrate may be so large that one should never trust the result. The attractive alternative is the optical pyrometry; however, here we have another set of difficulties. The major problem is that semiconductors are substantially transparent in the spectral region where thermal radiation would be emanated. In other words, emissivity of a semiconductor is small and the amount of thermal radiation from a semiconductor is also not only small, but due to wafer transparency, radiation from the underlying devices (e.g., a heater) will go through to the noncontact thermometer.

One relatively simple method is to coat a small portion of the semiconductor with a material having high and known emissivity in the IR spectral range. Then, the thermal radiation can be measured from that small patch. An example of such a material is *Nichrome* (see Table 72.1).

An attractive method of temperature monitoring when no emissive patch can be deposited is the use of temperature dependence of band gaps of common semiconductors. The band gap is determined from the threshold wavelength at which the radiation from the heaters behind the substrate is transmitted through the substrate. Another method is based on temperature dependence of diffuse reflection of a semiconductor. In effect, this method is similar to the former; however, it relies on reflection, rather than on transmission of the semiconductor. An external lamp is used for the measurement of a threshold wavelength from a front polished and backside textured substrate. The temperature measurement arrangement is shown in Figure 72.13, where the diffused light is detected by cryogenically cooled quantum detector. The monochromator has a resolution of 3 nm and scans through the threshold area at a rate of 100 nm/s.

72.7.2 Medical Thermometers

Medical noncontact (infrared) thermometry covers two distinct types of measurements: *body* and *skin* surface temperature measurements. The skin temperature measurements have specific applications in

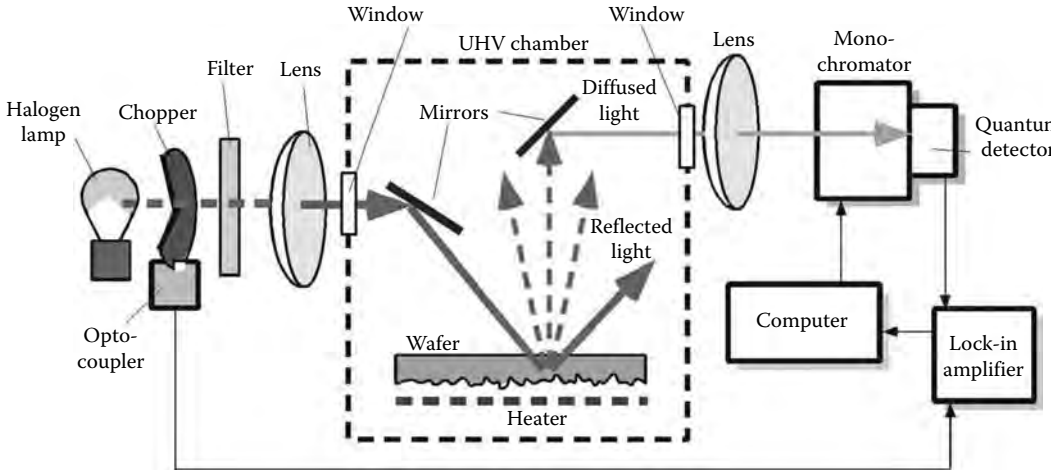


FIGURE 72.13 Diffused-light thermometer for measuring temperature of GaAs wafers.

diagnostics and are also used for estimation of the internal (core) body temperature. The skin temperature greatly depends on both the skin blood perfusion and environmental conditions. Therefore, it can be independently correlated with the temperature of the core only with some appreciable error; however, due to simplicity of use, the surface IR thermometers nowadays are quite popular for detecting fever.

A much more accurate way of measuring the core body temperature is by placing the probe of an IR thermometer into an ear canal and aiming it at a tympanic membrane [10]. The tympanic membrane has a temperature very close (within 0.1 °C) to that of blood, and by a correctly designed IR ear thermometer, it can be measured with high clinical accuracy.

Accuracy of the ear thermometer greatly depends on the accuracy of measuring flux Φ through the sensor's output voltage V, as in Equation 72.20. This is affected by the IR sensor sensitivity and its variations with temperature, optical components and their cleanliness, manufacturer's tolerances of the disposable sanitary probe covers, skill of an operator, and even the patient cooperation. All these often uncontrollable variables can cause an appreciable error that sometimes may be too large for an accurate medical diagnostic. Fortunately, Equation 72.24 suggests a solution that can dramatically improve accuracy in measuring patient's temperature. The error resulted from uncertainty of g is the so-called multiplicative, that is, it is proportional to the sensor's voltage V that, in turn, is proportional to a temperature gradient between the patient and the sensor (see Equation 72.18). Thus, by reducing a difference between T and T_s, one can reduce all multiplicative errors to smaller values and even close to zero. To do that, the sensor's temperature T_s should be artificially forced to come close to the patient temperature T. This was implemented in practice in the Braun-Thermoscan ear thermometer [11]. Figure 72.14 illustrates

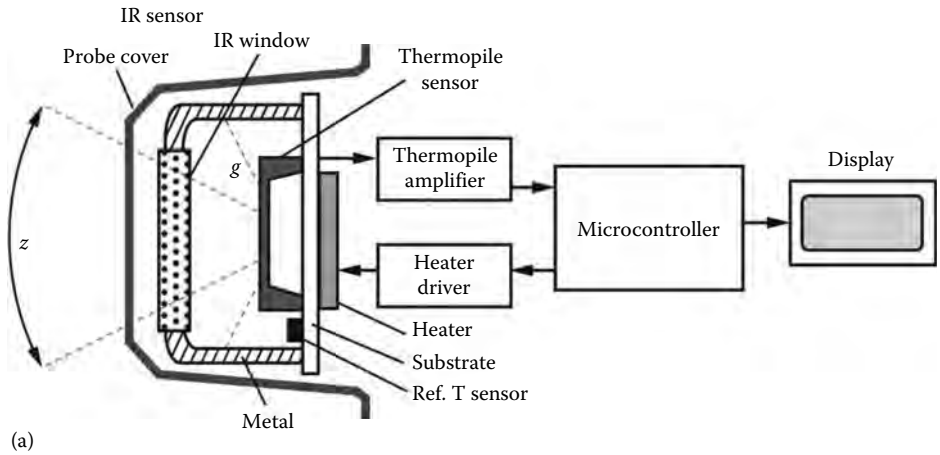


FIGURE 72.14 (a) Heated IR sensor in a medical ear thermometer and (b) the front end of Braun-Thermoscan ear thermometer.

a simplified block diagram of the thermometer. A thermopile IR sensor is encapsulated inside a metal header that is attached to a heater. Before taking temperature from a patient, the heater warms up the IR sensor from room temperature up to $t_{ss} = 37$ °C—a typical normal human body temperature. Thus, for a healthy patient, the thermal gradient between the patient and the sensor is close to zero and so is the measurement error. Even for a patient with fever, the temperature gradient is only 2 °C–4 °C and the error reduction is still very substantial, thus making this type of a medical thermometer highly accurate. Note that if the IR sensor is not warmed up and stays near room temperature, say 20 °C, the temperature gradient will be nearly an order of magnitude larger, and the errors will be larger as much. Since the skin emissivity is also multiplicative with *g*, a prewarmed sensor negates the effect of emissivity on accuracy as well.

A prewarming of the IR sensor to a temperature close to that of the patient is a powerful solution for medical applications since the patient's temperature varies in a narrow range (only from 35 °C to 42 °C) and a prewarming of the IR sensor to that range is not too difficult. Unfortunately, the method is not practical for use in much broader temperature ranges that are typical for nonmedical applications.

References

1. J. Fraden. *Handbook of Modern Sensors*, 4th edn., Springer Verlag, New York, 2010.
2. J. Fraden. Blackbody cavity for calibration of infrared thermometers. U.S. Patent 6447160, 2002.
3. J.L. Tissot, C. Trouilleau, B. Fieque, A. Crastes, and O. Legras. Uncooled microbolometer: Recent development at ULIS. *Opto-Electron. Rev*., 14(1), 25–32, 2006.
4. J. Schieferdecker, R. Quad, E. Holzenkämpfer, and M. Schulze. Infrared thermopile sensors with high sensitivity and very low temperature coefficient. *Sensor Actuat. A*, 46–47, 422–427,1995.
5. J. Schieferdecker, R. Quad, and M. Schulze. Temperature sensing module. U.S. Patent 5826982, 1998.
6. W. Leneke, M. Simon, M. Schulze, K. Storck, and J. Schieferdecker. Thermopile infrared sensor array. US Patent Publication No. 20080216883, 2008.
7. F.A. Cunnold. The optical system of the disappearing filament pyrometer. *Proc. R. Soc. Lond. A*, 152, 64–80, 1935.
8. T. Fu, X. Cheng, and Z. Yang. Theoretical evaluation of measurement uncertainties of two-color pyrometry applied to optical diagnostics. *Appl. Opt.,* 47(32), 6112–6123, November 10, 2008.
9. D.R. White and J.V. Nicholas. Emissivity and reflection error sources in radiation thermometry. In: *Temperature, Its Measurement and Control in Science and Industry*. Vol. 6, Part 2, pp. 917–922, AIP, New York, 1992.
10. J. Fraden. Medical infrared thermometry (review of modern techniques). In: *Temperature, Its Measurement and Control in Science and Industry*. Vol. 6, Part 2, pp. 825–830, AIP, New York, 1992.
11. B. Kraus and A. Klos. Radiation thermometer containing a heated measuring tip. U.S. Patent No. 7014358, 2006.

73

Pyroelectric Detectors

73.1 Pyroelectric Effect **73**-1
73.2 Pyroelectric Materials **73**-2
73.3 Manufacturing Process **73**-4
73.4 Pyroelectric Sensors **73**-5
73.5 Applications **73**-8
References **73**-9
Further Readings **73**-9

Jacob Fraden
Fraden Corporation

73.1 Pyroelectric Effect

The pyroelectric materials are crystalline substances capable of generating an electric charge in response to heat flow [1]. The pyroelectric effect is very closely related to the piezoelectric effect. The materials belong to a class of ferroelectrics. The name was given in association with ferromagnetics and is rather misleading because most such materials have nothing to do with iron.

A crystal is considered to be pyroelectric if it exhibits a spontaneous temperature-dependent polarization. Of the 32 crystal classes, 21 are noncentrosymmetric, and 10 of these exhibit pyroelectric properties. Besides pyroelectric properties, all these materials exhibit some degree of piezoelectric properties as well—they generate an electric charge in response to mechanical stress.

Pyroelectricity was observed for the first time in tourmaline crystals in the eighteenth century (some claim that the Greeks noticed it 23 centuries ago). Later, in the nineteenth century, Rochelle salt was used to make pyroelectric sensors. A large variety of materials became available after 1915: KDP (KH_2PO_4), ADP ($NH_4H_2PO_4$), $BaTiO_3$, and a composite of $PbTiO_3$ and $PbZrO_3$ known as PZT. Presently, more than 1000 materials with reversible polarization are known. They are called ferroelectric crystals. Most important among them are triglycine sulfate (TGS) and lithium tantalate ($LiTaO_3$).

A pyroelectric material can be considered a composition of a large number of minute crystallites, where each behaves as a small electric dipole. All these dipoles are randomly oriented, however, along a preferred direction. Above a certain temperature, known as the Curie point, the crystallites have no dipole moment.

When temperature of a pyroelectric material changes, the material becomes polarized. In other words, an electric charge appears on its surface. It should be clearly understood that the polarization occurs not as a function of temperature but only as function of a *change in temperature* of the material. There are several mechanisms by which changes in temperature will result in pyroelectricity. Temperature changes can cause shortening or elongation of individual dipoles. It can also affect the randomness of the dipole orientations due to thermal agitation. These phenomena are called *primary pyroelectricity*. There is also *secondary pyroelectricity*, which, in a simplified way, can be described as a result of the piezoelectric effect, that is, a development of strain in the material due to thermal expansion.

The dipole moment, M, of the bulk pyroelectric sensor is

$$M = \mu A h \tag{73.1}$$

where
μ is the dipole moment per unit volume
A is the sensor's area
h is the thickness

The charge, Q_a, which can be picked up by the electrodes, develops the dipole moment across the material:

$$M_0 = Q_a h \tag{73.2}$$

M must be equal to M_0 so that

$$Q_a = \mu A \tag{73.3}$$

As the temperature varies, the dipole moment also changes, resulting in an induced charge.

Thermal absorption can be related to a dipole change so that μ must be considered a function of both temperature, T_a, and an incremental thermal energy, ΔW, absorbed by the material:

$$\Delta Q_a = A\mu(T_a, \Delta W) \tag{73.4}$$

The aforementioned equation shows the magnitude of electric charge resulting from absorption of thermal energy. To pick up the charge, the pyroelectric materials are fabricated in the shapes of a flat capacitor with two electrodes on opposite sides and the pyroelectric material serving as a dielectric.

73.2 Pyroelectric Materials

To select the most appropriate pyroelectric material, energy conversion efficiency should be considered. It is, indeed, the function of the pyroelectric sensor to convert thermal energy into electrical. "How effective is the sensor?" is a key question in the design practice. A measure of efficiency is k_p^2 that is called the pyroelectric coupling coefficient [2,3]. It shows the factor by which the pyroelectric efficiency is lower than the Carnot limiting value $\Delta T/T_a$. Numerical values for k_p^2 are shown in Table 73.1.

Table 73.1 shows that TGS crystals are the most efficient pyroelectric converters. However, for a long time, they were quite impractical for use in sensors because of a low Curie temperature. If the sensor's temperature is elevated above that level, it permanently loses its polarization. In fact, TGS sensors proved to be unstable even below the Curie temperature, with the signal being lost quite spontaneously [4]. It was discovered that doping of TGS crystals with L-alanine (LATGS process patented by Philips) during its growth stabilizes the material below the Curie temperature. The Curie temperature was raised to 60 °C, which allows its use at the upper operating temperature of 55 °C, which is sufficient for many applications.

Other materials, such as lithium tantalate and pyroelectric ceramics, are also used to produce pyroelectric sensors. During recent years, deposition of pyroelectric thin films has been intensively researched. Especially promising is the use of lead titanate ($PbTiO_3$), which is a ferroelectric ceramic

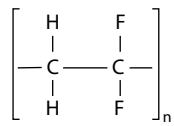
TABLE 73.1 Physical Properties of Pyroelectric Materials

Material	Curie Temperature, °C	Thermal Conductivity, W mK^{-1}	Relative Permittivity, ε_r	Pyroelectric Charge Coeff. C $(m^2K)^{-1}$	Pyroelectric Voltage Coeff. V $(mK)^{-1}$	Coupling k_p^2 (%)
Single crystals						
TGS	49	0.4	30	3.5×10^{-4}	1.3×10^{6}	7.5
$LiTaO_3$	618	4.2	45	2.0×10^{-4}	0.5×10^{6}	1.0
Ceramics						
$BaTiO_3$	120	3.0	1000	4.0×10^{-4}	0.05×10^{6}	0.2
PZT	340	1.2	1600	4.2×10^{-4}	0.03×10^{6}	0.14
Polymers						
PVDF polycrystalline layers	205	0.13	12	0.4×10^{-4}	0.40×10^{6}	0.2
$PbTiO_3$	470	2 (monocrystal)	200	2.3×10^{-4}	0.13×10^{6}	0.39

Source: Meixner, H. et al., *Siemens Forsch. Entwickl. Ber.*, 15(3), 105, 1986.
Note: The above-mentioned figures may vary depending on manufacturing technologies.

having both a high pyroelectric coefficient and a high Curie temperature of about 490 °C. This material can be easily deposited on silicon substrates by the so-called sol–gel spin-casting deposition method [5].

In 1969, Kawai discovered strong piezoelectricity in the plastic materials, polyvinyl fluoride (PVF) and polyvinylidene fluoride (PVDF) [6]. These materials also possess substantial pyroelectric properties. PVDF is a semicrystalline polymer with an approximate degree of crystallinity of 50% [7]. Like other semicrystalline polymers, PVDF consists of a lamellar structure mixed with amorphous regions. The chemical structure contains the repeat unit of doubly fluorinated ethane CF_2–CH_2:



The molecular weight of PVDF is about 10^5, which corresponds to about 2000 repeat units. The film is quite transparent in the visible- and near-IR regions and is absorptive in the far-infrared (IR) portion of the electromagnetic spectrum. The polymer melts at about 170 °C. Its density is about 1780 kg m^{-3}. PVDF is mechanically durable and flexible. In piezoelectric applications, it is usually drawn, uniaxially or biaxially, to several times its length. Elastic constants, for example, Young's modulus, depend on this draw ratio. Thus, if the PVDF film was drawn at 140 °C to the ratio of 4:1, the modulus value is 2.1 GPa, while for the draw ratio of 6.8:1, it was 4.1 GPa. Resistivity of the film also depends on the stretch ratio. For example, at low stretch, it is about 6.3×10^{15} Ω cm, while for the stretch ratio 7:1, it is 2×10^{16} Ω cm.

Since silicon does not possess pyroelectric properties, such properties can be added on by depositing crystalline layers of pyroelectric materials. The three most popular materials are zinc oxide (ZnO), aluminum nitride (AlN), and the so-called solid solution system of lead–zirconite–titanium oxides $Pb(Zr,Ti)O_3$ known as PZT ceramic, which is basically the same material used for fabrication of discrete piezoelectric and pyroelectric sensors. One of the advantages of using zinc oxide is the ease of chemical etching. The zinc oxide thin films are usually deposited on silicon by employing sputtering technology. Note, however, that silicon has a large coefficient of thermal conductivity. That is, its thermal time constant is very small (see as follows), so the pyroelectric sensors made with silicon substrates possess low sensitivity yet are capable of fast response.

73.3 Manufacturing Process

Manufacturing of ceramic PZT elements begins with high-purity metal oxides (lead oxide, zirconium oxide, titanium oxide, etc.) in the form of fine powders having various colors. The powders are milled to a specific fineness and mixed thoroughly in chemically correct proportions. In a process called "calcining," the mixtures are then exposed to an elevated temperature, allowing the ingredients to react to form a powder, each grain of which has a chemical composition close to the desired final composition. At this stage, however, the grain does not yet have the desired crystalline structure.

The next step is to mix the calcined powder with solid and/or liquid organic binders (intended to burn out during firing) and mechanically form the mixture into a "cake" that closely approximates a shape of the final sensing element. To form the "cakes" of desired shapes, several methods can be used. Among them are pressing (under force of a hydraulic-powered piston), casting (pouring viscous liquid into molds and allowing to dry), extrusion (pressing the mixture through a die or a pair of rolls to form thin sheets), and tape casting (pulling viscous liquid onto a smooth moving belt).

After the "cakes" have been formed, they are placed into a kiln and exposed to a very carefully controlled temperature profile. After burning out of organic binders, the material shrinks by about 15%. The "cakes" are heated to a red glow and maintained at that state for some time, which is called the "soak time," during which the final chemical reaction occurs. The crystalline structure is formed when the material is cooled down. Depending on the material, the entire firing may take 24 h.

When the material is cold, contact electrodes are applied to its surface. This can be done by several methods. The most common are fired-on silver (a silk screening of silver–glass mixture and refiring), electroless plating (a chemical deposition in a special bath), and sputtering (an exposure to metal vapor in partial vacuum).

Cristalities (crystal cells) in the material can be considered electric dipoles. In some materials, like quartz, these cells are naturally oriented along the crystal axes, thus giving the material sensitivity to stress. In other materials, the dipoles are randomly oriented, and the materials need to be "poled" to possess piezoelectric properties. To give crystalline material pyroelectric properties, several poling techniques can be used. The most popular poling process is thermal poling, which includes the following steps:

1. A crystalline material (ceramic or polymer film) that has randomly oriented dipoles (Figure 73.1a) is warmed to slightly below its Curie temperature. In some cases (for a PVDF film), the material is stressed. High temperature results in stronger agitation of dipoles and permits one to more easily orient them in a desirable direction.
2. The material is placed in strong electric field, E (Figure 73.1b), where dipoles align along the field lines. The alignment is not total. Many dipoles deviate from the field direction quite strongly; however, statistically predominant orientation of the dipoles is maintained.

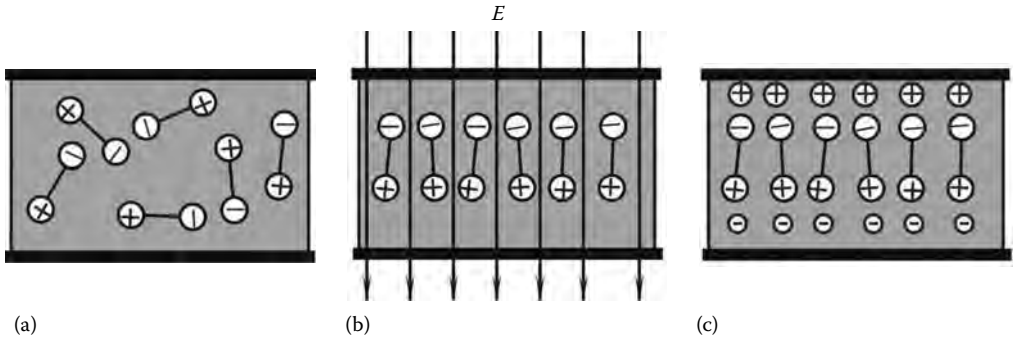


FIGURE 73.1 Poling of a pyroelectric crystal in a strong electric field: (a) randomly oriented dipoles, (b) dipoles aligning with electric field, and (c) dipoles frozen at high temperature without the electric field. The sensor must be stored and operated below the Curie temperature.

3. The material is cooled down, while the electric field across its thickness is maintained.
4. The electric field is removed, and the poling process is complete. As long as the poled material is maintained below the Curie temperature, its polarization remains permanent. The dipoles stay "frozen" in the direction that was given to them by the electric field at high temperature (Figure 73.1c). The aforementioned method is used to manufacture ceramic and plastic pyroelectric materials.

Another method, called a corona discharge poling, can be used to produce polymer piezoelectric/pyroelectric films. The film is subjected to a corona discharge from an electrode at several million volts per centimeter of film thickness for 40–50 s [8,9]. Corona polarization is uncomplicated to perform and can be easily applied before electric breakdown occurs, making this process useful at room temperature.

The final operation in preparing the sensing element is shaping and finishing. This includes cutting, machining, and grinding. After the piezo (pyro) element is prepared, it is installed into a sensor's housing, where its electrodes are bonded to electrical terminals and other electronic components.

After poling, the crystal remains permanently polarized; however, it remains electrically charged for a relatively short time. There is a sufficient amount of free carriers that move in the electric field setup inside the bulk material, and there are plenty of charged ions in the surrounding air. The charge carriers move toward the poled dipoles and neutralize their charges (Figure 73.1c). Hence, after a while, the poled piezoelectric material becomes electrically discharged as long as it remains under steady-state conditions. When temperature changes and thermally induced stress develops, the balanced state is degraded, and the pyroelectric material develops an electric charge. If the stress is maintained for a while, the charges again will be neutralized by the internal leakage. Thus, a pyroelectric material is responsive only to a changing temperature rather than to a steady level of it. In other words, a pyroelectric sensor is an ac device, rather than a dc device. Sometimes, it is called a *dynamic* sensor, which reflects the nature of its response.

73.4 Pyroelectric Sensors

To make sensors, the pyroelectric materials are used in the form of thin slices or films with electrodes deposited on the opposite sides to collect the thermally induced charges (Figure 73.2). The pyroelectric detector is essentially a capacitor that can be charged by an influx of heat. The detector does not require any external electrical bias (excitation signal). It needs only an appropriate electronic interface circuit to measure the charge. Contrary to thermoelectrics (thermocouples), which produce a steady voltage when two dissimilar metal junctions are held at steady but different temperatures, pyroelectrics generate charge in response to a change in temperature. Since a change in temperature essentially requires propagation of heat, a pyroelectric device is a heat flow detector rather than a heat detector. Figure 73.2 shows a pyroelectric detector (pyro-sensor) connected to a resistor R_b that represents either the internal leakage resistance or a combined input resistance of the interface circuit connected to the sensor. The equivalent electrical circuit of the sensor is shown on the right. It consists of three components: (1) the current source generating a heat-induced current, i (remember that a current is a movement of electric charges); (2) the sensor's capacitance, C; and (3) the leakage resistance, R_b. Since the leakage resistance

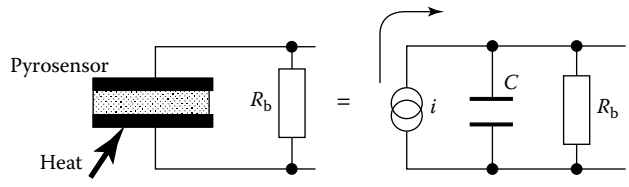


FIGURE 73.2 Pyroelectric sensor and its equivalent circuit.

is very high and often unpredictable, an additional bias resistor is often connected in parallel with the pyroelectric material. The value of that resistor is much smaller than the leakage resistance, yet its typical value is still on the order of 10^{10} Ω (10 GΩ).

The output signal from the pyroelectric sensor can be taken in the form of either charge (current) or voltage, depending on the application. Being a capacitor, the pyroelectric device is discharged when connected to a resistor, R_b (Figure 73.2). Electric current through the resistor and voltage across the resistor represent the heat flow–induced charge. It can be characterized by two pyroelectric coefficients [2]:

$$\begin{aligned} P_Q &= \frac{\mathrm{d}P_s}{\mathrm{d}T} \text{ Pyroelectric charge coefficient} \\ P_V &= \frac{\mathrm{d}E}{\mathrm{d}T} \text{ Pyroelectric voltage coefficient} \end{aligned} \tag{73.5}$$

where

P_s is the spontaneous polarization (which is the other way to say "electric charge")
E is the electric field strength
T is the temperature in K

Both coefficients are related by way of the electric permittivity, ε_r, and dielectric constant, ε_0:

$$\frac{P_Q}{P_V} = \frac{\mathrm{d}P_s}{\mathrm{d}E} = \varepsilon_r \varepsilon_0 \tag{73.6}$$

The polarization is temperature dependent, and, as a result, both pyroelectric coefficients in Equation 73.5 are also functions of temperature.

If a pyroelectric material is exposed to a heat source, its temperature rises by ΔT, and the corresponding charge and voltage changes can be described by the following equations:

$$\Delta Q = P_Q A \Delta T \tag{73.7}$$

$$\Delta V = P_V h \Delta T \tag{73.8}$$

Remembering that the sensor's capacitance can be defined as

$$C_e = \frac{\Delta Q}{\Delta V} = \varepsilon_r \varepsilon_0 \frac{A}{h} \tag{73.9}$$

then, from Equations 73.6, 73.8, and 73.9, it follows that

$$\Delta V = \frac{\Delta Q}{C_e} = P_Q \frac{h}{\varepsilon_r \varepsilon_0} \Delta T \tag{73.10}$$

It is seen that the peak output voltage is proportional to the sensor's temperature rise and pyroelectric charge coefficient.

Figure 73.3 shows a pyroelectric sensor whose temperature, T_0, is homogeneous over its volume. Being electrically polarized, the dipoles are oriented (poled) in such a manner as to make one side of the material positive and the opposite side negative. However, under steady-state conditions, free charge carriers

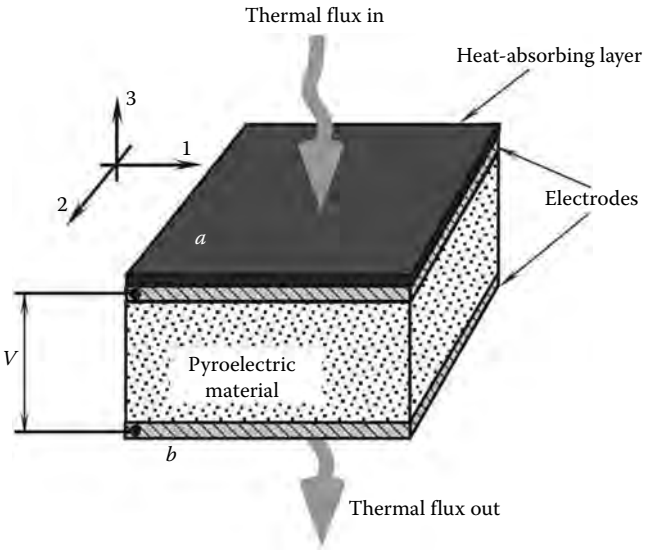


FIGURE 73.3 Pyroelectric sensor has two electrodes at the opposite sides of the crystalline material. Thermal radiation is applied along axis 3.

(electrons and holes) neutralize the polarized charge, and the capacitance between the electrodes does not appear to be charged. That is, the sensor generates zero charge. Now, assume that heat is applied to the top side of the sensor. Heat can enter the sensor in a form of thermal radiation that is absorbed by the top electrode and propagates toward the pyroelectric material via the mechanism of thermal conduction. The top electrode can be given a heat-absorbing coating, such as gold–black or organic paint. As a result of heat absorption, the top side becomes warmer (the new temperature is T_1), which causes the top side of the material to expand. The expansion leads to flexing of the sensor, which, in turn, produces stress and a change in dipole orientation. Being piezoelectric, stressed material generates electric charges of opposite polarities across the electrodes. Hence, one can regard secondary pyroelectricity as a sequence of events: thermal radiation, heat absorption, thermally induced stress, and electric charge.

The temperature of the sensor T_s is a function of time. That function is dependent on the sensing element: its density, specific heat, and thickness. If the input thermal flux has the shape of a step function of time, for the sensor freely mounted in air, the output current can be approximated by an exponential function so that

$$i = i_0 e^{-t/\tau_T} \tag{73.11}$$

where i_0 is the peak current.

Figure 73.4 shows timing diagrams for a pyroelectric sensor when it is exposed to a step function of heat. It is seen that the electric charge reaches its peak value almost instantaneously and then decays with a *thermal time constant,* τ_T. This time constant is a product of the sensor's thermal capacitance, C, and thermal resistance, r, which defines a thermal loss from the sensing element to its surroundings:

$$\tau_T = Cr = cAhr \tag{73.12}$$

where c is the specific heat of the sensing element

The thermal resistance r is a function of all thermal losses to the surroundings through convection, conduction, and thermal radiation. For low-frequency applications, it is desirable to use sensors with τ_T

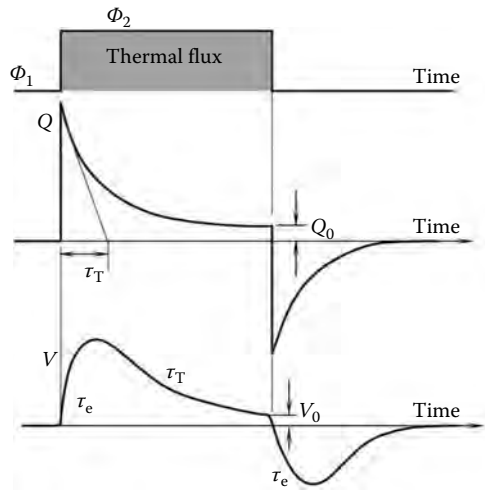


FIGURE 73.4 Response of a pyroelectric sensor to a thermal step function. The magnitudes of charge Q_0 and voltage V_0 are exaggerated for clarity.

as large as practical, while for the high-speed applications (e.g., to measure the power of laser pulses), a thermal time constant should be dramatically reduced. For that purpose, the pyroelectric material can be laminated with a heat sink: a piece of aluminum or copper.

When a heat flow exists inside the pyroelectric crystal, there is an outflow of heat from the opposite side of the crystal, as shown in Figure 73.3. Thermal energy enters the pyroelectric material from side *a*. Since the other side *b* of the sensor faces a cooler environment, part of the thermal energy leaves the sensor and is lost to its surroundings. Because the sides *a* and *b* face objects of different temperatures (one is the temperature of a target and the other is the temperature of the environment), a continuous heat flow exists through the pyroelectric material. As a result, in Figure 73.4, charge Q and voltage V do not completely return to zero no matter how much time has elapsed. Electric current generated by the pyroelectric sensor has the same shape as the thermal current through its material. An accurate measurement can demonstrate that as long as the heat continues to flow, the pyroelectric sensor will generate a constant voltage V_0 whose magnitude is proportional to the heat flow.

73.5 Applications

The pyroelectric sensors are useful whenever changing thermal radiation or heat flow needs to be measured. Examples are motion detectors for security systems and light control switches [1], instant medical IR thermometers, and laser power meters. Depending on the application, a pyroelectric sensor can be used either in *current* or in *voltage* mode. The voltage mode (Figure 73.5a) uses a voltage follower with a very high input resistance. Hence, JFET- and CMOS-input stages are essential. As a rule, in the voltage-mode sensor, the follower is incorporated inside the same package along with the element and bias resistor. Advantages of the voltage mode are simplicity and lower noise. The disadvantages are slower speed response due to high capacitance of the sensor (typically on the order of 30 pF) and other influences of the sensor capacitance on the quality of output voltage. The output voltage of the follower is shown in Figure 73.4 (V). It is seen that it rises slowly with electric time constant τ_e and decays with thermal time constant τ_T.

The current mode uses an electronic circuit having a "virtual ground" as its input (Figure 73.5b). An advantage of this circuit is that the output signal is independent of the sensor's capacitance and, as a result, is much faster. The signal reaches its peak value relatively fast and decays with thermal time

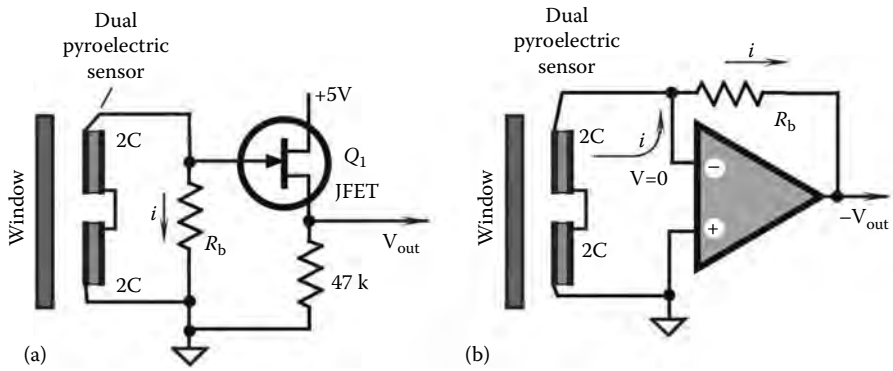


FIGURE 73.5 Interface circuits for pyroelectric sensors operating in (a) voltage and (b) current modes.

constant τ_T. The output voltage V_0 has the same shape of charge Q in Figure 73.4. The disadvantages of the circuit are higher noise (due to wider bandwidth) and higher cost.

Note that Figure 73.5 shows dual pyroelectric sensors, where two sensing elements are formed on the same crystalline plate by depositing two pairs of electrodes. The electrodes are connected in a serial opposite manner. If both sensors are exposed to the same magnitude of far-IR radiation, they will produce nearly identical polarizations, and, due to the opposite connection, the voltage applied to the input of the transistor or the current passing through resistor R_b will be nullified. This feature allows for cancelation of undesirable common-mode input signals in order to improve stability and reduce noise. Signals that arrive only to one of the elements will not be canceled.

References

1. J. Fraden, *Handbook of Modern Sensors*, 4th edn. New York: Springer Verlag, 2010.
2. H. Meixner, G. Mader, and P. Kleinschmidt, Infrared sensors based on the pyroelectric polymer polyvinylidene fluoride (PVDF), *Siemens Forsch. Entwickl. Ber.*, 15(3), 105–114, 1986.
3. P. Kleinschmidt, Piezo-und pyroelektrische Effekte. In: Heywang, W. (Ed.), *Sensorik*, Kap. 6. New York: Springer, 1984.
4. Semiconductor Sensors, *Data Handbook*. Eindhoven, the Netherlands: Philips Export B.V., 1988.
5. C. Ye, T. Tamagawa, and D.L. Polla, Pyroelectric $PbTiO_3$ thin films for microsensor applications. In: *Transducers'91. International Conference on Solid-State Sensors and Actuators*, San Francisco, CA, pp. 904–907, 1991.
6. H. Kawai, The piezoelectricity of poly(vinylidene fluoride), *Jap. J. Appl. Phys.*, 8, 975–976, 1969.
7. A. Okada, Some electrical and optical properties of ferroelectric lead-zirconite-lead-titanate thin films, *J. Appl. Phys.*, N. 48, 2905, 1977.
8. P.F. Radice, Corona discharge poling process, U.S. Patent No. 4,365,283, 1982.
9. P.D. Southgate, Room-temperature poling and morphology changes in pyroelectric polyvinylidene fluoride. *Appl. Phys. Lett.*, 28, 250, 1976.

Further Readings

Allied Industries, Pyroelectric. http://www.alliedelec.com/Emitters-and-Detectors/Pyroelectric (accessed on August 1, 2012).

EOC, Pyroelectric Sensor Detector. http://www.eoc-inc.com/pyroelectric_detectors.htm (accessed on August 4, 2012).

Infratec, How does a Pyroelectric Detector Work. http://www.infratec.de/en/sensor-division/literature/faq.html (accessed on August 4, 2012).



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

74

Liquid-in-Glass Thermometers

74.1 General Description **74**-1
74.2 Choice of Thermometers **74**-3
74.3 Operation **74**-3
74.4 Errors in Liquid-in-Glass Thermometry **74**-4
Scale Markings • Immersion Errors and Stem Corrections • Hysteresis and Stability • Time-Constant Effects • Pressure • Parallax • Stiction • Thermal Capacity Effects • Separated Columns
References **74**-12
Partial List of Vendors and Suppliers **74**-13

Rod White
Measurement Standards Laboratory of New Zealand

J.V. Nicholas
Measurement Standards Laboratory of New Zealand

Liquid-in-glass (LIG) thermometers are a remarkably long-lived technology. The first sealed LIG thermometers were based on the thermal expansion of wine spirit and developed in 1641 by the then Duke of Tuscany [1,2]. For the next 300 years, the technology gradually improved so that a bewildering variety of LIG thermometers are now available to suit almost every application. They cover temperature ranges from –190 °C to 600 °C, and some achieve accuracies approaching 0.001 °C. The distinctive form, self-contained nature, and apparent simplicity of the LIG thermometer have made it the archetypical thermometer for many generations of technical staff and the general public.

However, LIG thermometers have limitations: they are fragile, potentially toxic, and prone to instability. Nowadays, in practically all but the most specialized applications, LIG thermometers are being replaced by electronic resistance thermometers that outperform them in almost every respect, especially in cost of ownership. Nevertheless, LIG thermometers still have appeal for some general-purpose applications and are required in a wide range of product and materials testing applications covered by documentary standards.

74.1 General Description

The most common form of LIG thermometer is the solid-stem thermometer, similar to that illustrated in Figure 74.1. There are four main parts to the LIG thermometer:

- The thin glass bulb, which contains the bulk of the thermometric liquid, is usually designed for stability with temperature change and must be properly annealed to remove stresses and strains that may cause the bulb to slowly change shape. The bulb must be thin to ensure good thermal contact with the medium being measured, but this means it is susceptible to instabilities due to mechanical or thermal shock. Some low-accuracy, general-purpose thermometers are made with a thicker bulb to lower the risk of breakage.
- The stem is a glass capillary. Again, a suitable glass is necessary, and it may differ from that of the bulb. The volume of the capillary must be significantly smaller than the volume of the bulb for

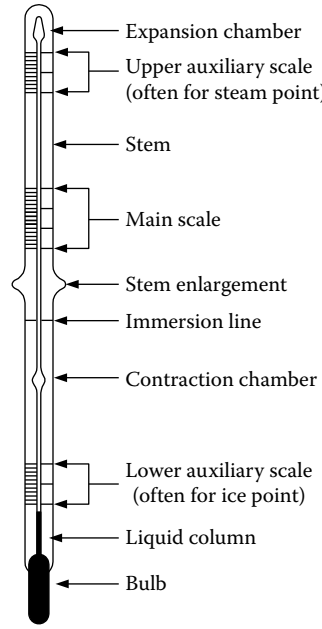


FIGURE 74.1 Main features of a solid-stem LIG thermometer.

good sensitivity, and it must be smooth and have a uniform cross section. The empty part of the capillary above the thermometric fluid can be gas filled or vacuous.

- The thermometric liquid is usually mercury in high-precision thermometers but is commonly an organic liquid in low-cost thermometers and in thermometers used for low temperatures. Mercury–thallium alloy is occasionally used for applications down to −55 °C, below the freezing point of mercury near −38 °C, but is also very toxic.
- The markings may be engraved, etched, or printed onto the stem. The markings include the scale, to allow direct reading of the temperature, as well as other useful information. In enclosed-scale thermometers, the scale is printed onto a separate glass plate and enclosed within the tubular body of the thermometer next to the capillary.

In addition to these basic features, many thermometers will have some of the additional features illustrated in Figure 74.1. The additional features may include

- An expansion chamber, which prevents the mercury from overpressuring the thermometer if it is exposed to temperatures above its maximum indication.
- A contraction chamber accompanied by an auxiliary scale near 0 °C, or sometimes 100 °C, which makes it possible to manufacture high-resolution thermometers that can be checked at the ice point or water boiling point. Good-quality thermometers include a 0 °C indication, either on the main scale or as an auxiliary scale.
- An immersion line or instructions for immersion, typically indicating that the thermometer should be immersed to 75 mm (or 3 in.) or 100 mm (or 4 in.) for partial-immersion thermometers, and the words "total immersion" for total-immersion thermometers.
- A short descriptor indicating whether the thermometer is gas filled or vacuous.
- A light-colored stripe up the back of the stem to aid scale visibility. The color of the stripe also indicates the type of glass [3].
- A stem enlargement, which is typically used to enable mounting of the thermometer in specialized apparatus.

Not all of these features will be found on all LIG thermometers.

TABLE 74.1 Summary of Specifications for ASTM Precision Thermometers

ASTM Thermometer Number	Range (°C)	Maximum Length (mm)	Graduations (°C)	Maximum Error (°C)
62C	−38 to +2	384	0.1	0.1
63C	−8 to +32	384	0.1	0.1
64C	−0.5 to +0.5 and 25 to 55	384	0.1	0.1
65C	−0.5 to +0.5 and 50 to 80	384	0.1	0.1
66C	−0.5 to +0.5 and 75 to 105	384	0.1	0.1
67C	−0.5 to +0.5 and 95 to 155	384	0.2	0.2
68C	−0.5 to +0.5 and 145 to 205	384	0.2	0.2
69C	−0.5 to +0.5 and 195 to 305	384	0.5	0.5
70C	−0.5 to +0.5 and 295 to 405	384	0.5	0.5

74.2 Choice of Thermometers

A large number of different types of LIG thermometer are available, with many of the variations designed with different dimensions and temperature ranges to suit specific applications. For best accuracy, mercury-in-glass thermometers should be used and restricted to operation over the range –38 °C to 250 °C. The purchase should be guided by a specification from a recognized standards body. Such bodies include the International Standards Organization (ISO) [4], the American Society for Testing and Materials (ASTM) [5], the British Standards Institute (BSI) [6], or the Institute for Petroleum (IP). Be aware that the standards bodies may use the same type number for different thermometers. For example, the IP 16C and the ASTM 16C are very different thermometers.

A choice of thermometer will probably be a compromise between the best range, resolution, and dimensions for the application. If good precision is required, then the thermometer range should be constrained to avoid long and unwieldy thermometers. Table 74.1 summarizes the ASTM specifications [5] for precision thermometers for which the smallest scale division is 0.1 °C. BS 1900 reference thermometers are available divided to 0.02 °C. The ASTM thermometers are also specified to be accurate to one scale division. Experience suggests that as many as 15% of all LIG thermometers are in error by more than one scale division at some point on their scales. However, with calibration, the scales on LIG thermometers can usually be interpolated to achieve accuracies of between 0.2 and 0.5 of a scale division, depending on the quality and resolution of the thermometer. The purchase cost of a good reference thermometer depends on the range and would typically be US$100 to US$200 for a good-quality thermometer. However, such costing neglects the cost of calibration that may be US$200 to US$2,000 or more. In general, good-quality electronic platinum resistance thermometers will have a lower cost of ownership and should be the first choice where measurement specifications allow. More information can be found in the references at the end of this chapter for further guidance on LIG thermometers [3,7,8], especially if high accuracy is required.*

74.3 Operation

The operation of LIG thermometers is most obviously based on the expansion of the thermometric liquid. As the volume of the liquid increases with temperature, it occupies a greater fraction of the space within the thermometer and is forced up the capillary where it provides the temperature indication. The flattest part of the liquid meniscus is used as the indicator: for mercury, this is the top of a convex meniscus; for organic liquids, the bottom of a concave meniscus.

* A good treatment of calibration practice for LIG thermometers, with a wider coverage than given here.

Perhaps less obviously, the glass is also a key part of the thermometer and a major source of the errors in LIG thermometry. Glass expands and contracts on heating sufficiently to offset the expansion of mercury by about 10%, and much of the nonlinearity, hysteresis, and possible erratic behavior of mercury-in-glass thermometers is due to the glass. Hysteresis, in particular, limits the basic accuracy of mercury-in-glass thermometers to about 0.1% of the temperature range, and special care and procedures are required to achieve greater accuracies. To avoid stress, mercury-in-glass thermometers should be stored horizontally on trays in cabinets, with care being taken to avoid any weight or pressure on the bulb. Vibration should also be avoided. Corrugated cardboard, or similar material, can be used as a liner for a tray to prevent the thermometers from rolling.

LIG thermometers employing organic fluids or "spirits" suffer less from the effects of the glass because the fluids have a much larger expansion coefficient. However, organic fluids have their own sources of error making spirit thermometers much less accurate and less reliable than mercury-in-glass thermometers. Spirit thermometers are useful below –38 °C, where mercury freezes, in situations where mercury is to be avoided, and in low-cost thermometers.

The sources of error associated with sprit thermometers include the thermal expansion of organic fluids, which is far less linear than for mercury. Secondly, the fluids are more volatile, which may lead to condensation on the upper part of the capillary. Spirit thermometers may also be difficult to read because of the clear liquid and concave meniscus. However, they do have a wider capillary, and, in combination with a suitable dye, they can be more readable than mercury thermometers. The wetting of the capillary by the liquid also means that many minutes may be required for the liquid to drain properly when measuring falling temperatures. Some of the liquids are also prone to polymerization when exposed to UV light.

Perhaps the largest source of error in LIG thermometry is due to the fact that not all of the thermometric fluid is at the measured temperature. Because some part of the liquid column must be visible, that part of the thermometric fluid must be outside the medium of interest and not at the correct temperature. This undesirable situation is made much worse when the operator fails to immerse a thermometer sufficiently and is also a feature of the normal use of partial-immersion thermometers. The same effect means that partial-immersion thermometers are sensitive to the stem temperature. The resulting errors can be large: a total-immersion thermometer partially immersed in boiling water with the entire stem outside the water exhibits an error of about 1 °C.

74.4 Errors in Liquid-in-Glass Thermometry

74.4.1 Scale Markings

During manufacture, the unmarked thermometer is exposed to two or more known temperatures, and "pointing" marks are placed on the stem corresponding to those temperatures, as shown in Figure 74.2. A linear scale is then placed on the stem to interpolate between adjacent pointing marks. In high-precision or wide-range thermometers, the scale may be placed in several segments to reduce nonlinearity errors. In general-purpose thermometers, there may only be two pointing marks, so the whole scale is placed at once. There are several factors that contribute to errors in the placement of the scale markings.

Firstly, in low-quality thermometers, the scale placement is not well controlled and may be up to one division in error. A good indicator of the quality of a thermometer is how close the pointing marks are to the scale markings. Sometimes, pronounced glitches in the scale markings can be seen where two scale segments meet. In high-quality thermometers, the scale will be well aligned with the pointing marks.

Secondly, the quality of the interpolation can vary. In good-quality thermometers, the scale is usually engraved or etched using a ruling machine to interpolate between the pointing marks. The scale markings are typically very fine and easy to interpolate between. In low-quality thermometers, the scale may be screen printed or stenciled. Such scales tend to have coarse scale markings and may be distorted through stretching of the screen or stencil.



FIGURE 74.2 Good-quality thermometer (left) and general-purpose thermometer (right) showing the difference in the quality of scale markings. In the good-quality thermometer, the 0 °C pointing mark is immediately alongside the corresponding scale mark. In the general-purpose thermometer, the 50 °C pointing mark is about ¼ scale divisions above the scale mark.

Thirdly, as noted earlier, the expansion of both the thermometric liquid and the glass is nonlinear to some degree. This means that most of the scale indications will be in error with the error varying over the entire range of the thermometer, and the scale indication may be correct only near the pointing marks.

Another major contributor to nonlinearity is the capillary. Ideally, the cross section of the capillary should be uniform over the length of the thermometer and, to some degree, uniform between thermometers of the same type so that their dimensions are the same. In high-quality thermometers, the uniformity of the capillary is specified and tested before being used for thermometers.

Collectively, these errors mean that LIG thermometer scales can have a very complicated pattern of errors, and effective calibrations must sample the errors over the anticipated range of use. For the most part, thermometers designed for specific applications will be used over a very narrow range of temperatures, perhaps only within ±0.3 °C of a single temperature. Examples include thermometers used for viscosity testing at 40 °C and 100 °C. General-purpose thermometers, on the other hand, tend to be used almost randomly over their whole range, and the calibration must sample the scale over the whole range.

74.4.2 Immersion Errors and Stem Corrections

If some of the thermometric liquid in a LIG thermometer is at a different temperature from the bulb, then immersion errors occur. In the context of LIG thermometry, they are often called "stem errors." Because LIG thermometers are sometimes used with varying immersion conditions, such errors are common. Indeed, in order to read a thermometer, a small part of the liquid column is normally required to be outside the region of interest. The errors due to immersion effects can be estimated and corrected by the procedures given as follows.

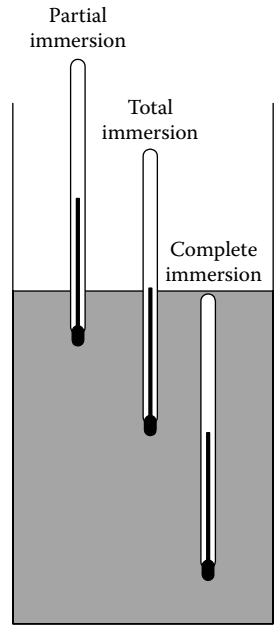


FIGURE 74.3 Immersion conditions used for LIG thermometers.

Three distinct immersion conditions are recognized for LIG thermometers, as illustrated in Figure 74.3:

- *Complete immersion*: By definition, if the bulb and the entire stem are immersed at the same temperature, the thermometer is completely immersed. This condition is not common, except at room temperature. Complete immersion should be avoided, especially at higher temperatures as the excess pressure above the liquid can cause the thermometer to rupture leading to a spread of mercury vapor. DO NOT put a mercury thermometer completely inside an oven to measure the temperature.
- *Total immersion*: This is the most common and most reliable operating condition. The description "total immersion" applies to the thermometric liquid, that is, all the liquid in the bulb, any contraction chambers, and the stem, is at the temperature of interest. The remaining portion of the stem where there is no thermometric liquid will be closer to room temperature. When thermometers are designed, calibrated, and used in total-immersion conditions, the immersion errors are very small, and usually no further corrections are required.
- *Partial immersion*: One way around the problem of scale visibility and the need that mount a thermometer within a complex apparatus is to immerse the thermometer to some fixed depth so that most, but not all, of the fluid is at the temperature of interest. The part of the fluid column not immersed is referred to as the emergent column. Many thermometers are designed and calibrated for partial immersion and are marked accordingly on the stem with an immersion depth or an immersion line (see Figure 74.1). The scale of a partial-immersion thermometer is properly defined only when the average temperature of the emergent column is also specified. Corrections may be required to compensate for the error arising from the emergent column not being at the specified temperature.

When LIG thermometers are used in incorrect immersion conditions, corrections should be made on the basis of the measured temperature of the emergent column. The traditional stem temperature measurement is made with a faden thermometer (the German word for thread thermometer). These are mercury-in-glass thermometers with a very thin long bulb, and a set of faden thermometers includes a wide range of bulb lengths. The bulb of the appropriate faden thermometer is mounted alongside the emergent column with the bottom of the bulb just in the fluid. The faden thermometer then indicates

the average stem temperature. Other ways of measuring the temperature profile are to use other small temperature sensors, such as thermocouples, mounted along the length of the emergent column. The thermometers should be evenly distributed over the length so that the average reading gives the correct length-weighted average stem temperature.

Thermometers are usually calibrated at the immersion conditions for which they have been designed. For partial-immersion thermometers, the stem temperatures during calibration should be measured and reported on the certificate. This ensures that the errors are negligible and can be ignored when the thermometer is used under its proper conditions. In many applications, a thermometer used for a standard test method, such as specified by the ASTM or IP, will have a specified immersion length and a specified stem temperature. Depending on the stem temperature during calibration, measurements reported in a calibration certificate may have to be corrected to meet the immersion conditions of the specification and to confirm that the thermometer meets any specified quality criteria.

When making a stem correction, the true temperature reading t is given by

$$t = t_i + N(t_2 - t_1)k \tag{74.1}$$

where

t_i is the indicated temperature

N is the length of emergent column expressed in degrees, as determined by the thermometer scale

t_2 is the mean temperature of the emergent column as specified or when calibrated (i.e., the stem temperature on a certificate for a partial-immersion thermometer or the reported reading for a total-immersion thermometer)

t_1 is the mean temperature of the emergent column in use, the fluid temperature if used in total immersion, or the measured stem temperature if used in partial immersion

k is the coefficient of apparent expansion of the thermometric liquid used in the glass of which the thermometer is made

See Table 74.2 for suitable values to use for normal temperature ranges and common glasses.

The use of Equation 74.1 with typical values of the liquid expansion coefficients from Table 74.2 is estimated to give about 10% accuracy in the correction. Consequently, a large correction is a major source of uncertainty. If the uncertainty due to the correction is considered too large, an alternative measurement method should be used.

Figure 74.4 shows a chart derived from Equation 74.1 for mercury thermometers. The chart is useful for making estimates of the correction and determining whether the immersion error is significant and therefore needs correction. For example, consider the case of measuring boiling water in a

TABLE 74.2 Working Range of Some Thermometric Liquids and Their Apparent Thermal Expansion Coefficient in Thermometer Glasses around Room Temperature

Liquid	Typical Apparent Expansion Coefficient $(°C)^{-1}$	Possible Temperature Range (°C)
Mercury	0.000 16	−35 to 510
Ethanol	0.001 04	−80 to 60
Pentane	0.001 45	−200 to 30
Toluene	0.001 03	−80 to 100

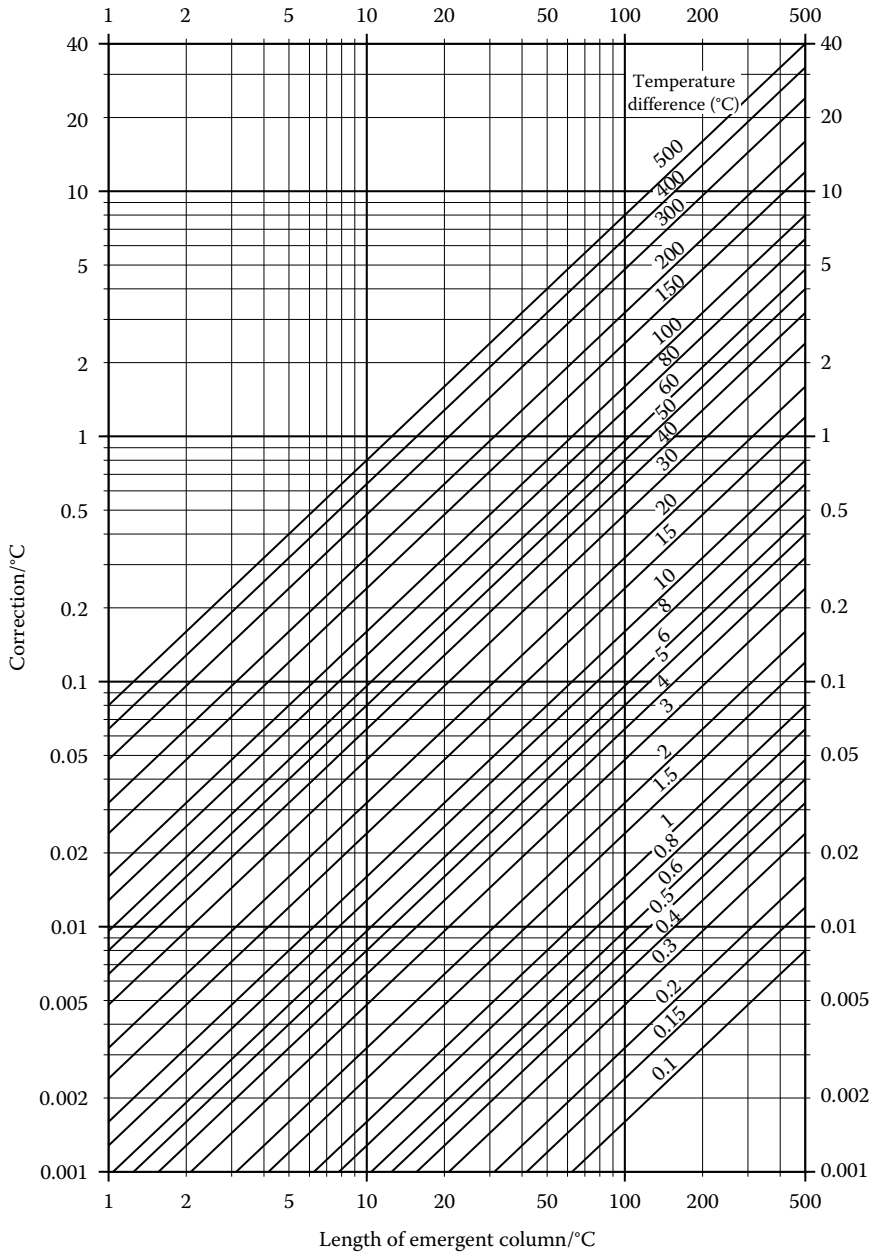


FIGURE 74.4 Stem correction chart for mercury-in-glass thermometers.

beaker with a total-immersion thermometer. The thermometer is immersed in the water to around the 20 °C mark. The emergent column is therefore 80 °C long. For the purpose of this calculation, we assume that the stem is close to the room temperature of 20 °C, resulting in an 80 °C temperature difference in the stem temperature from the calibration conditions. On the chart, the intersection of the line for the 80 °C emergent column length and line for the 80 °C temperature difference indicates a correction of just over 1 °C.

A similar application of the chart, with 1 °C emergent column length and a 5 °C temperature difference, shows that the error in the correct use of a total-immersion thermometer, with 1 °C of the column

visible, is well below 0.001 °C and can therefore be neglected. There are several different permutations of the correction equation depending on the thermometer type and how it is used. More examples can be found in the text of Nicholas and White [3].*

74.4.3 Hysteresis and Stability

Glass is an amorphous solid, which means that it does not have a well-defined atomic structure. A small part of the volume expansion and contraction of the glass that occurs with heating and cooling is due to small atomic rearrangements that occur slowly. If a thermometer is withdrawn from high temperatures and allowed to cool naturally, it will exhibit hysteresis or "temporary depression," the magnitude of which depends on the type of glass and the temperature to which the thermometer was exposed. The effect is easily measured by performing an ice-point check immediately before and after a measurement. The ice point, at 0.000 °C, is the temperature at which ice- and air-saturated water are in thermal equilibrium [9]. Always, after exposure to a higher temperature, the second ice-point reading will be lower than the first because the mercury contracts much faster than the glass. For good-quality mercury-in-glass thermometers, the effect is about 0.1% of the temperature change, that is, about 0.1 °C depression for a 100 °C temperature change in the glass. The thermometer will then relax quite rapidly over a period of days to close to its normal state although a residual effect will be measureable for months.

If accuracies of better than 0.1% are sought using mercury-in-glass thermometers, care must be taken to ensure that the thermal history of the thermometer is well controlled. This involves prescribed heating, cooling, and recovery processes that must be applied identically in use and calibration. With care, the temporary depression can be avoided. The atomic rearrangements responsible for the hysteresis occur more quickly at higher temperatures, and if the thermometer is cooled slowly, at 10 °C/h or slower [10], the effect is almost completely eliminated.

Besides the hysteresis, the glass bulb undergoes a slow contraction known as the secular effect, which causes the thermometer reading to slowly increase with age. Fortunately the effect is very slow, typically much less than a scale division per year, and regular calibration checks at the ice point will indicate the rate of drift. Because the effect originates in the bulb, the measured change in ice-point reading can be used to correct the whole of the thermometer scale.

The tendency for the glass to reorganize itself and seek the most stable amorphous state means that poorly annealed thermometers or thermometers that are subject to thermal or mechanical stress, such as being rested on their bulb, may never fully stabilize. Improperly annealed thermometers, for example, usually appear perfectly stable when received from the manufacturer but may exhibit random upward and downward changes of many scale divisions over periods as short as days or weeks following their first use.

74.4.4 Time-Constant Effects

In well-stirred media, the thermal time constant of an LIG thermometer is determined almost entirely by the diameter of the bulb as heat is conducted through the glass from the outside to the center of the bulb. A typical bulb of 5 mm diameter has a relatively short time constant of a few seconds when in well-stirred fluids. However, when the medium is slow or has a high thermal resistivity, the time constant is dominated by the medium itself and may be as large as many tens or hundreds of seconds. Table 74.3 gives the $1/e$ time constants in various media for a thermometer with a 5 mm diameter bulb. Time constants for other diameters can be estimated by scaling the times in proportion to the diameter.

* The material presented in this chapter section was extracted and adapted from this text. The text explains, for several different thermometer types, how to make traceable measurements and calibrate thermometers to meet international requirements.

TABLE 74.3 Time Constants for a Mercury-in-Glass Thermometer with a 5 mm Diameter Bulb

Medium	Still (s)	0.5 m/s Flow (s)	Infinite Flow Velocity (s)
Water	10	2.4	2.2
Oil	40	4.8	2.2
Air	190	71	2.2

LIG thermometers using organic thermometric fluids have an additional time-constant effect when measuring falling temperatures due to the time required for the fluid wetting the capillary to drain and join the main column. For precision thermometry, about 3 min/cm of wetted capillary is recommended.

74.4.5 Pressure

The thin bulb on most LIG thermometers unfortunately makes them sensitive to pressure. In good-quality mercury thermometers, the effect corresponds to about 0.1 °C for a one atmosphere change in pressure. This may seem like a small effect, but the pressure of a 760 mm high mercury column corresponds to one atmosphere. For this reason, most thermometers will give different readings in vertical and horizontal positions.

Another source of pressure variations is fluid turbulence in calibration or test baths. The pressure fluctuations from turbulence are readily visible as small and rapid variations in the reading. Ideally, fluid baths used with LIG thermometers should be constructed to provide a laminar flow along the axis of the thermometer.

74.4.6 Parallax

Parallax is the difference in the apparent position of an object viewed along two different lines of sight. The effect occurs in solid-stem LIG thermometers because the liquid column and the scale on thermometers are in different planes. Ideally, the thermometer should be set up so that the scale can be easily read while viewing at right angles to the stem. Optical aids can be useful to ensure a repeatable viewing position and to enlarge the image, but paradoxically they also increase the parallax effect, so care is still required.

Enclosed-scale thermometers are designed to minimize parallax by placing the capillary, and the scale markings in the same plane, nevertheless, always viewing at right angles will minimize the effects of any scale misplacement.

74.4.7 Stiction

A mercury thermometer with a fine capillary not only becomes difficult to read but also suffers from stiction: the mercury moving in fits and starts due to the surface tension between the mercury and the capillary walls. The effect can be overcome by lightly tapping the thermometer with a small pencil or paintbrush. The brush is also useful for clearing condensation when doing an ice-point check.

74.4.8 Thermal Capacity Effects

Glass thermometers are bulky and have considerable mass, especially high-precision thermometers with long bulbs. The transient heat flow accompanying the warming or cooling of the thermometer can upset temperature measurements, making high-precision measurements difficult. Preheating or cooling the

thermometer can alleviate the worst of the problem. For higher precision and lower masses, choose a platinum resistance thermometer or thermistor instead. Simple estimates of the heat requirements can be made by measuring the volume of thermometer immersed and assuming that 2 J are required to raise 1 cm^3 of the thermometer volume (glass or mercury) by 1 °C. This is a good rule of thumb for most solids and liquids.

74.4.9 Separated Columns

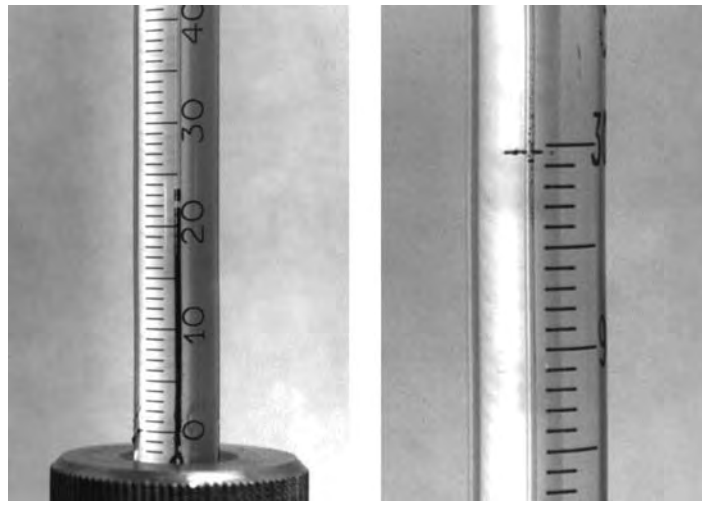
A common problem with LIG thermometers is for a part of the thermometric liquid to become separated from the main volume. This happens often when thermometers are shipped but may also happen with use at high temperature when the liquid vapor condenses near the cool end of the capillary (Figure 74.5). The effect is readily detectable as an ice-point shift, but the capillary of thermometers should be checked visually before use and especially after shipping.

With organic liquids, the problem can be more difficult to identify because the liquid is often clear and wets the surface of the capillary and may not be as visible. For this reason, spirit thermometers should be stored by suspending them vertically to allow the thermometric liquid to drain. The typical vertical temperature gradient found in most rooms will also warm the top of the thermometer to prevent condensation of vapor high up the capillary.

The main reason for including a gas fill in LIG thermometers is to prevent column separation. The gas not only "holds" a column together but also retards the diffusion of vapor at high temperatures and so prevents condensation. Usually, an inert gas like dry nitrogen is used to prevent oxidation. Note that an expansion chamber must be kept cooler than the bulb to prevent excessive gas-pressure buildup. The high pressure can permanently distort the bulb even if rupture does not occur.

Mechanical separation of the liquid column is, unfortunately, a common occurrence. The gas fill helps prevent separation, but, conversely, the gas makes the column more difficult to rejoin. There is also a risk of trapped gas bubbles in the bulb or contraction chambers, and careful inspection may be needed to locate them. A vacuous capillary will give rise to more breaks, but the column is easily rejoined.

With care, it is usually possible to rejoin the column and still have a viable thermometer. However, it must be realized that attempts to join a broken column may stress a thermometer sufficiently to render it permanently unstable. Column breaks that occur repeatedly when the thermometer is heated indicate the thermometer should be discarded.



(a) (b)

FIGURE 74.5 Separated columns in mercury-in-glass thermometers, showing (a) a break in the column and (b) condensation—near the pointing mark.

Various procedures for joining a broken column can be tried. The procedures as follows are given in order of preference:

- Lightly tap the thermometer while holding it vertically. This may work for a vacuous thermometer.
- For LIG thermometers using organic fluids, suspend the thermometer vertically for several hours with the upper end of the stem warmed slightly.
- Apply centrifugal force. This is best done by holding the bulb alongside the thumb, protecting it with the fingers, and with the stem along the arm. Raise the unbent arm above the head and bring it down quickly to alongside the leg. This method avoids a flicking action, which will cause breakage, and protects the thermometer from striking other objects.
- If the aforementioned methods are unsuccessful, try cooling the thermometer. This method relies on sufficient cooling for all the fluid to contract into the bulb, leaving none in the stem. The column should be rejoined when it has warmed to room temperature. More than one attempt might be needed. Tapping and centrifugal force might also be needed to assist the movement of the fluid.

Three readily available cooling media can be used:

- Salt, ice, and water (to –18 °C)
- Dry ice, that is, solid CO_2 (–78 °C)
- Liquid nitrogen (–196 °C)

Liquid nitrogen and dry ice require more care as they will freeze the mercury (near –38°C). *Cold burns to the user could also occur.* It pays to cool the thermometer slowly by alternately dipping the thermometer quickly into the media and then watching the mercury column fall. A slow cooling will also prevent excessive stress in the glass.

Heating a thermometer to force a rejoin in the expansion chamber is NOT RECOMMENDED because it exposes the thermometer to temperatures above its annealing temperature, which will cause it to become unstable, and it may cause the bulb to rupture.

If the broken column has been successfully rejoined, then an ice-point check (or other reference check) should be made. If the reading is the same as obtained previously, then the join can be considered completely successful, and the thermometer is ready for use. However, a significant ice-point shift indicates that the join was not successful and that the thermometer should be discarded. If a small ice-point shift has occurred, then the shift might simply be due to hysteresis, so treat the thermometer with suspicion until there is evidence of long-term stability, that is, no further significant ice-point changes occur.

References

1. W. E. K. Middleton, *A History of the Thermometer and Its Uses in Meteorology*, John Hopkins Press, Baltimore, MD, 1966.
2. T. J. Quinn, *Temperature*, 2nd edn., Academic Press, London, U.K., 1990.
3. J. V. Nicholas and D. R. White, *Traceable Temperatures*, 2nd edn., John Wiley & Sons, Chichester, U.K., 2001.
4. ISO issued documentary standards related to the liquid-in-glass thermometer:
   ISO 386-1977 Liquid-in-glass laboratory thermometer—Principles of design, construction and use
   ISO 651-1975 Solid-stem calorimeter thermometers
   ISO 653-1980 Long solid-stem thermometers for precision use
   ISO 654-1980 Short solid-stem thermometers for precision use
   ISO 1770-1981 Solid-stem general purpose thermometer
5. ASTM in their standards on Temperature Measurement, Vol. 14.03, include two standards related to liquid-in-glass thermometers:
   E1-95 Specification for ASTM Thermometers
   E77-92 Test Method for Inspection and Verification of Liquid-in-glass Thermometers

6. BSI published a series of documentary specifications for thermometers, including
   BS 593:1989 Laboratory Thermometers
   BS 791:1990 Thermometers for Bomb Calorimeters
   BS 1704:1985 General Thermometers
   BS 1900:1976 Secondary Reference Thermometers
7. J. A. Wise, *Liquid-in-Glass Thermometer Calibration Service*, National Institute of Standards and Technology Special Publication, 250-23, 1988.
8. E. C. Holligan, Calibration and use of liquid-in-glass thermometers. In: R. E. Bentley (ed.), *Handbook of Temperature Measurement*, Vol. 2, pp. 107–133, Springer, New York, 1998.
9. ASTM International, *ASTM E 563-08 Standard Practice for Preparation and Use of an Ice-Point Bath as a Reference Temperature,* ASTM International, West Conshohocken, PA, 2008.
10. S. J. Van Dijk, J.A. Hall, and V.M. Leaver, The influence of cooling rate on the zeros of mercury-in-glass thermometers, *J. Sci. Instrum*., 35, 334–348, 1958.

Partial List of Vendors and Suppliers

Kimble Chase, 8500 Spruce St., Vineland, NJ 08360-1502, Phone: 856-692-8500, 888-546-2531 (toll free), Fax: 856-794-9762, www.kimble-chase.com/index.html

Palmer Wahl, 234 Old Weaverville Rd., Asheville, NC 28804-1228, Phone: 828-658-3131, 800-421-2853 (toll free), Fax: 828-658-0728, www.palmerwahl.com

Tel-True Mfg. Co., Inc., 408 St. Paul St., Rochester, NY 14605-1734, Phone: 585-232-1440, 800-232-5335 (toll free), Fax: 585-232-3857, www.teltru.com/

ThermoProbe, Inc., 112-A Jetport Dr., Pearl, MS 39208, Phone: 601-939-1831, Fax: 601-355-1831, www.thermoprobe.net

ThermoWorks, 1762 W. 20 S., Suite 100, Lindon, UT 84042, Phone: 801-756-7705, Fax: 801-756-8948, www.thermoworks.com

VEE GEE Scientific, Inc., 13600 N.E. 126th Place, "A", Kirkland, WA 98034-8720, Phone: 425-823-4518, 800-423-8842 (toll free), Fax: 425-820-9826, www.veegee.com



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

75

Manometric Thermometers

Franco Pavese
National Institute of Meterology Research

75.1 Vapor Pressure 75-2
75.2 Gas Thermometry 75-4
Constant-Volume Gas Thermometer • Acoustic Gas Thermometer
References 75-12

Manometric thermometers are defined in this handbook as those thermometers that make use of the pressure of a *gaseous* medium as the physical quantity to obtain temperature. Very seldom are they available from commercial sources; for example, the temperature control of a home freezer is often of this kind. Consequently, instead of simply buying one, every user must build his or her own when this type thermometer is needed. They can be a quite useful choice since, in the era of electronic devices and sensors, it is still possible to make a totally nonelectronic thermometer, though of modest precision, which in addition keeps its calibration indefinitely, as long as the quantity of substance sealed in it remains unchanged. The range of temperatures that can be covered by this kind of thermometer depends on the principle and on the substance used. When the thermodynamic equilibrium between the *condensed* phase of a substance (either liquid or solid) and its vapor is used, one has a "vapor-pressure thermometer," and the temperature range spanned by each substance is generally narrow. In addition, only substances that are gaseous at room temperature (i.e., condensed at temperatures lower than 0 °C) are normally used, confining the working range to below room temperature; however, some substances that are liquid at room temperature and have a high vapor pressure (i.e., which easily evaporate) have been used, but do not result in a sizable extension of the working range much above room temperature. A special case of vapor pressure being used at high temperature is the device called a "heat pipe," which is not used as a thermometer, but instead as an accurate thermostat [1]; using sodium, the working range is pushed up to ~1100 °C. When a pure substance is used only in its gaseous state, one has a "gas thermometer," whose temperature range can be very wide, especially for moderate accuracy, depending mainly on the manometer; on the other hand, its fabrication is somewhat more complex and its use normally much less straightforward.

Both thermometers can be built to satisfy the state-of-the-art accuracy of national standards (uncertainty better than ±0.001 K) or for lower accuracies, down to an uncertainty of ±1% or higher. Both thermometers require the measurement of pressure, in the range from less than 1 Pa up to 100 bar. A complete and specialized treatment on both vapor-pressure and gas thermometers up to room temperature and on pressure measurement instruments and techniques for gaseous media can be found in [2]. Gas thermometry above room temperature is treated in [3,4].

In consideration of the fact that these kinds of thermometers typically must be built by the users, the following will concentrate on the basic guidelines for their design and fabrication.

75.1 Vapor Pressure

Figures 75.1 and 75.2 show the pressure values and the sensitivities in the temperature range allowed for each of the most common gaseous substances, considering also, in addition to the liquid phase, the use of the solid phase (where vapor pressure is lower) below the triple point. The lower end of the range is determined by the manometer uncertainty (for a given accuracy), the upper end by the full-scale pressure of the manometer (or by the full evaporation of the condensed phase).

Table 75.1 reports "certified" vapor-pressure equations, linking the measured pressure p to the unknown temperature T. The reader might prefer them to the plethora of equations found in the

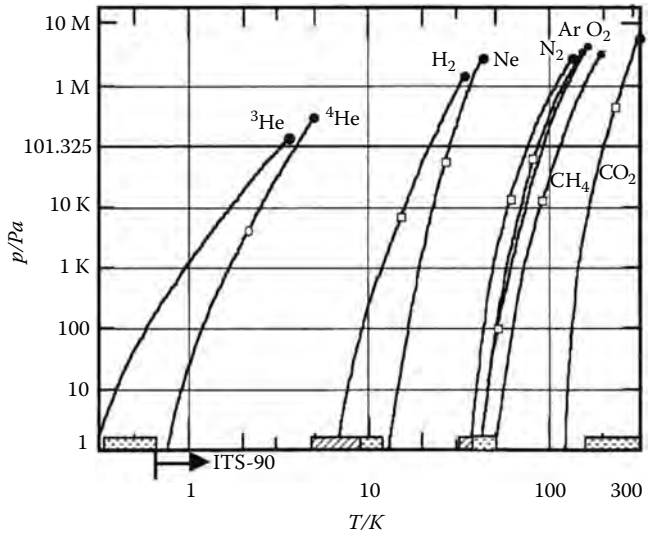


FIGURE 75.1 Range for vapor-pressure thermometry of various gases. The shaded parts indicate regions where it is less common or less accurate. ▨, not available; ⁙, lower accuracy; ●, critical point; ○, triple point; 0, lambda point.

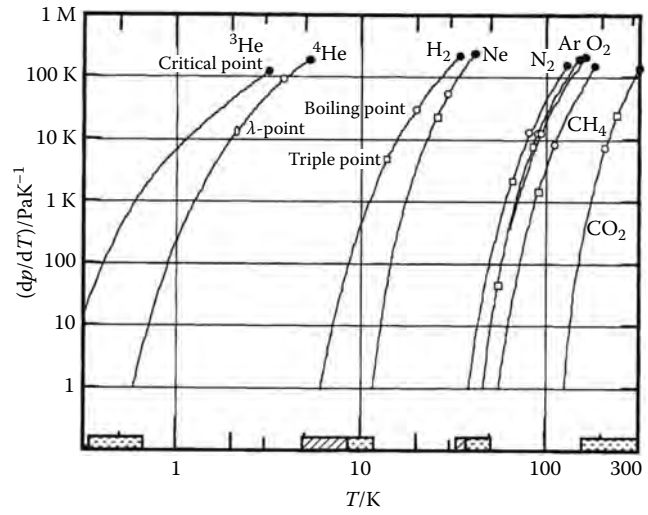


FIGURE 75.2 Sensitivity dp/dT of vapor-pressure thermometry for selected gases. The shaded parts indicate regions where it is less common or less accurate. ▨, not available; ⁙, lower accuracy; ●, critical point; ○, triple point; 0, lambda point.

TABLE 75.1 Vapor-Pressure Equations

Equilibrium State	T_{90} (K)	Uncertainty $\pm\delta T$ (mK)	Purity of Material[a] (vol%)
Liquid–vapor phases of helium-4	1.25–2.1768	0.1	99.9999
	$T_{90}/K = A_0 + \sum_{i=1}^{9} A_i \left[\left(\ln(p/Pa) - B \right) / C \right]^i$		
	$A_0 = 1.392408$	$A_1 = 0.527153$	$A_2 = 0.166756$
	$A_3 = 0.050988$	$A_4 = 0.026514$	$A_5 = 0.001975$
	$A_6 = -0.017976$	$A_7 = 0.005409$	$A_8 = 0.013259$
	$B = 5.6$	$C = 2.9$	
	2.1768–5.0	0.1	99.9999
	$A_0 = 3.146631$	$A_1 = 1.357655$	$A_2 = 0.413923$
	$A_3 = 0.091159$	$A_4 = 0.016349$	$A_5 = 0.001826$
	$A_6 = -0.004325$	$A_7 = -0.004973$	$B = 10.3$
	$C = 1.95$		
Liquid–vapor phases of equilibrium hydrogen	13.8–20.3	1[b]	99.99
	$p/Pa = (p_0/Pa)\exp\left[A + B/(T_{90}/K) + C\,T_{90}/K \right] + \sum_{i=0}^{5} b_i (T_{90}/K)^i$		
	$A = 4.037592968$	$B = -101.2775246$	
	$C = 0.0478333313$		
	$b_0 = 1902.885683$	$b_1 = -331.2282212$	$b_2 = 32.25341774$
	$b_3 = -2.106674684$	$b_4 = 0.060293573$	$b_5 = -0.000645154$
Liquid–vapor phases of natural neon[c]	24.6–40	2	99.99
	$\log(p/p_0) = A + B/(T_{90}/K) + C(T_{90}/K) + D(T_{90}/K)^2$		
	$A = 4.61948943$	$B = -160.478268$	
	$C = -0.0369937132$	$D = 0.00004256101$	
Solid–vapor phases of nitrogen	56.0–63.1	2	99.999
	$\log(p/p_0) = A + B/(T_{90}/K) + C(T_{90}/K)$		
	$A = 12.07856655$	$B = -858.0046109$	$C = -0.009224098$
Liquid–vapor phases of nitrogen	63.2–125	5	99.999
	$A = -6.10273365$	$B = 1.153844492$	$C = -1.087106903$
	$D = -1.759094154$	$T_c = 126.2124$ K	$P_c = 3.39997$ MPa
Liquid–vapor phases of oxygen	54–154	2	99.999
	$\mathrm{In}(p/p_c) = T_c/T_{90}\left[A\tau + B\tau^{1.5} + C\tau^3 + D\tau^7 + D\tau^9 \right]$		
	$\tau = 1 - T_{90}/T_c$	$A = -6.044437278$	
	$B = 1.176127337$	$C = -0.994073392$	$D = -3.449554987$
	$E = 3.343141113$	$T_c = 154.5947$ K	$P_c = 5.0430$ MPa
	$\mathrm{In}(p/p_c) = T_c/T_{90}\left[A\tau + B\tau^{1.5} + C\tau^3 + D\tau^6 \right]; \quad \tau = 1 - (T_{90}/T_c)$		

(*continued*)

TABLE 75.1 (continued) Vapor-Pressure Equations

Equilibrium State	T_{90} (K)	Uncertainty $\pm\delta T$ (mK)	Purity of Material[a] (vol%)
Liquid–vapor phases of argon	83.8–150	5	99.999
	$A = -5.906852299$	$B = 1.132416723$	$C = -0.7720072001$
	$D = -1.671235815$	$T_c = 150.7037$ K	$P_c = 4.8653$ MPa
Liquid–vapor phases of methane	90.7–190	5[d]	99.99
	$\ln(p/p_c) = T_c/T_{90}\left[At + B\tau^{1.5} + C\tau^{2.5} + D\tau^5\right];\quad \tau = 1-(T_{90}/T_c)$		
	$A = -6.047641425$	$B = 1.346053934$	$C = -0.660194779$
	$D = -1.304583684$	$T_c = 190.568$ K	$p_c = 4.595$ MPa
Liquid–vapor phases of carbon dioxide	216.6–304	15	99.99
	$\ln(p/p_c) = A_0\left(1-(T_{90}/T_c)\right)^{1.935} + \sum_{i=1}^{4} A_i\left((T_c/T_{90})-1\right)^i$		
	$p_c = 7.3825$	$T_c = 304.2022$ K	$A_0 = 11.37453929$
	$A_1 = -6.886475614$	$A_2 = -9.589976746$	$A_3 = 13.6748941$
	$A_4 = -8.601763027$		

Note: For the relevant references and more gases, see [6]. p_0 = 101325 Pa, except when otherwise indicated.

[a] Minimum purity of the material to which the listed values of temperature and uncertainty apply.

[b] The summation term in the equation adds to the value of p a pressure amounting to the equivalent of 1 mK maximum.

[c] These values are for neon with an isotopic composition close to that specified in the ITS-90 definition.

[d] Above 100 K. It increases to 15 mK at 91 K and to 10 mK near the critical point.

literature, since they have been checked by official bodies and T is expressed as T_{90}, in the ITS-90, the International Temperature Scale [5, text in 2]. More checked equations can be found in [6]. The vapor-pressure equations are generally empirical equations; therefore, the temperature measured with a vapor-pressure thermometer is not the thermodynamic temperature. T_{90} and thermodynamic temperature T must be used (see Section 75.6).

Figure 75.3 shows the general layout of a vapor-pressure thermometer. The fabrication of a vapor-pressure thermometer is not exceedingly difficult if a few guidelines are followed. Table 75.2 summarizes the most critical ones—design criteria and filling information—in a compact form [2]. Much more constructional details can be found in [1,3]. In most cases, the manometer is located at room temperature, and the bulb is connected to it via a small-bore tube (the "capillary tube") without critical drawbacks. The accuracy of these thermometers ranges from ±0.0001 K using very pure substances in calorimeters and precision mercury manometers to ≈ ±1% using dial manometers.

75.2 Gas Thermometry

As indicated in Chapter 65 of this handbook, since 2005, one is allowed to consider the direct use of thermodynamic temperature as an internationally agreed method to measure temperature. Thermodynamic thermometers can be of different kinds. The manometric ones can be of different types. The historically more common is the constant-volume gas thermometer (CVGT), but in recent years also, the acoustical thermometer has greatly improved its accuracy and, to a minor extent, the dielectric-constant gas thermometer.

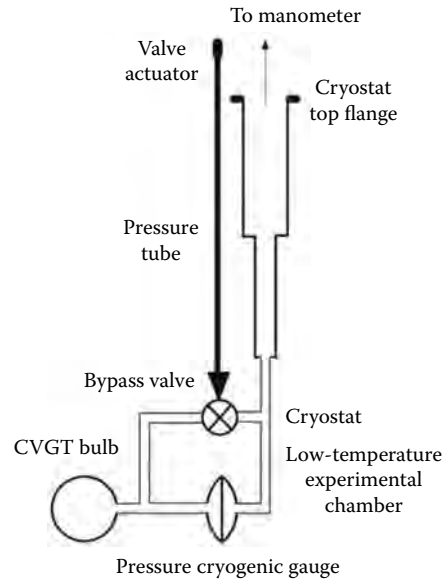


FIGURE 75.3 General layout of a manometric thermometer. It is shown with a cryogenic diaphragm pressure transducer; when the transducer is placed instead at room temperature, the bypass valve is also placed at room temperature. *Vapor-pressure thermometer:* The diameter of the pressure tube increases in steps when pressures lower than ≈0 Pa must be measured, in order to decrease the thermomolecular pressure drop. Gas *thermometer* (constant volume): The diameter of the pipes connecting the bulb to the cryogenic pressure transducer must be small in order to reduce the so-called dead-volume. This requirement is much more stringent when the pressure transducer is moved up to room temperature. In this case, in order to reduce the error due to the "dead volume," the bulb volume must be increased significantly, as shown in Figure 75.4.

75.2.1 Constant-Volume Gas Thermometer

The layout of Figure 75.3 also applies to the design of a CVGT (more common than the constant-pressure type), with the differences indicated in the relevant caption.

For a pressure gauge placed at room temperature, the design becomes that of Figure 75.4. The lower temperature end of the working range of a gas thermometer of this type is stated, well before condensation of the substance takes place, by the increase of the uncertainty due to the increase in the nonideality of the gas—that is, deviation from linearity of the relationship $p(T)$—which takes place when approaching the condensed state or for increasing gas densities or due to excessive absorption of gas on the bulb surface, thereby changing the quantity of the thermometric substance. All these conditions act at the lower end of the range. The upper end is stated by technological reasons or by the manometer full-scale capability. The best substances are, as listed, helium (either isotopes), hydrogen, and nitrogen.

From a design point of view, Table 75.3 summarizes the most critical issues. The major problem, apart from gas purity and ideality, is meeting the constant-volume requirement. Being that the manometer is generally at room temperature, the fraction of gas filling the connecting "capillary" tube is subtracted from the total amount of thermometric substance amount filling the system, and since this fraction is not constant, but depends on temperature and on technical conditions, it tends to increase the measurement uncertainty, which is contrary to the case of the vapor-pressure thermometer; this error is called the *dead-volume error,* which determines the largest contribution to the total uncertainty. In modern realizations, the use of a cryostat based on a close-cycle refrigerator can sensibly reduce this

TABLE 75.2 Summary of Design Criteria for Vapor-Pressure Thermometers

	Example
1. Choice of working substance	$T_{max} = P_c$
Temperature range: Each substance spans only a narrow temperature interval.	T_{min} = 100 Pa K^{-1}
T_{max}/T_{min} < 2–3 (including solid–vapor range), except helium. The limit	
T_{max} set by maximum manometer pressure	$T_{max}/T_{min} = {}^3He \approx 10$
T_{max} set by manometer sensitivity	$^4He \approx 9$
Accuracy	$H_2 \approx 3$
Manometer: No single manometer spans whole range from ≈ 1 Pa ($dp/dT \approx 100$ Pa K^{-1}) and critical point ($p_c > 10^6$ Pa, except helium) with high or constant accuracy or with sufficient sensitivity.	Ne ≈ 3
Substance: Not all substances allow for maximum accuracy, due to purity or to thermal problems related to a low thermal diffusivity value.	$N_2 \approx 2.5$
2. Choice of pressure measuring system	$O_2 \approx 2.5$
Sensitivity and accuracy must be matched to the range of dp/dT and of p, that is, T, to be measured.	Ar ≈ 2.5
Without separating diaphragm: can be used only for low to medium accuracy, as thermometric gas also fills the entire manometric apparatus, with problems of contamination and increases in vapor volume.	$CO_2 \approx 2$ (solid ≈ 1.5)
Dial *manometers*: used only for accuracy > ±1%.	
Metal *diaphragm* or bellows (electronic) manometers: can achieve a ±0.1%–0.03% accuracy.	
Quartz Bourdon gauges: can approach a ±0.01% accuracy, but helium leaks through quartz.	
Cryogenic *pressure* transducers: None is commercially available with accuracy better than ± 0.1% (after cryogenic calibration). Eliminate the need of the connecting tube in sealed thermometers, but transducer must withstand high room-temperature pressure.	
With separation diaphragm: mandatory for high or top accuracy. Only zero reproducibility and a moderate linearity near zero are important.	
Capacitive diaphragms: Several commercial models, when properly used, allow zero sensitivity and reproducibility better than ±0.1 Pa.	
Cryogenic diaphragms: Only laboratory-made diaphragms are available, some with high zero reproducibility. Allow to confine thermometric gas at low temperatures, but the tube connecting the diaphragm to room-temperature manometer is still necessary.	
Room-temperature manometers: When a cryogenic diaphragm is used, only manometers allowing helium as manometric gas can be used.	
3. Choice of sealed versus "open" thermometer	
Sealed: low-accuracy-only (e.g., dial) thermometers.	
Medium-accuracy sealed thermometers are still very simple when using cryogenic manometer and reducing vapor volume, but room-temperature pressure can be higher than 10 MPa. Therefore, only low-sensitivity manometers can be used, and thermometer measures only upper part of vapor-pressure scale.	
High-accuracy sealed thermometers can be made, using ballast room-temperature volume and precision room-temperature diaphragm.	
"Open": Vapor-pressure thermometers using gases are open only since working substance does not stay permanently in working bulb, but (new) samples are condensed in it only during measurements. Requires permanent use of a gas-handling system.	

TABLE 75.2 (continued) Summary of Design Criteria for Vapor-Pressure Thermometers

	Example
4. Gas purity, isotopic composition, and spin equilibrium	
Purity: must be known and possibly checked, for example, by performing a triple-point temperature measurement. Dew boiling point difference measurement must also routinely be performed, before sealing in the case of sealed devices.	Problems only for high-accuracy Kr, Xe (for isotopic composition), H_2, D_2 (for spin-equilibrium composition)
Isotopic composition: Some gases show irreproducibility in results due to sample-to-sample changes in isotopic composition. It is impossible to obtain top accuracy with these substances, unless pure isotopes are used.	
Spin equilibrium: With some gases, showing different spin species, equilibrium must be ensured with use of a suitable catalyst.	
5. Thermometer filling	^{4}He thermometer
Amount of substance n_{max} at $T_{min} \rightarrow V^L \approx V_b$	$T_{min} = 2.2$ K
$n_{max} \le \frac{p_{max}}{RT_r}\left[\frac{2V_cT_r}{T_f + T_{min}} + V_f\right] + \frac{V_b}{M}\rho_{min}$	$p_{min} = 5.263$ kPa $\rho_{min} = 146$ kg m^{-3}
Amount of substance n_{max} at $T_{min} \rightarrow V^L = V_\varepsilon^L; \approx 0$	$T_{max} = 5.2$ K
$n_{min} \ge \frac{V_\varepsilon^L \rho_{max}}{M} + \frac{\rho_{max}}{RT_t}\left[\frac{2V_cT_t}{T_t + T_{min}} + V_r + \frac{T_r}{T_{max}}(V_{max} - V_\varepsilon^L)\right]$	$p_{max} = 227.5$ kPa
Bulb volume V_b	$p_{max} = 67.5$ kg m^{-3}
$V_b\left[\frac{\rho_{min}}{M} - \frac{\rho_{max}}{RT_{max}}\right] \le \frac{V_r}{RT_r}\left[p_{max} - p_{min}\right] + \left[\frac{\boldsymbol{\rho_{max}}}{\boldsymbol{M}} - \frac{\boldsymbol{\rho_{max}}}{\boldsymbol{RT_{max}}}\right]$	$T_f = 4.2$ K $p_f = 100$ kPa $T_t = 300$ K $p_t = 200$ kPa $V_t = 220$ cm^3 $V_c = 16$ cm^3 $V_T = 500$ cm^3 $M = 4$ g mol^{-1}
$+ \frac{\mathbf{2V_c}}{\boldsymbol{R}}\left[\frac{\boldsymbol{p_{min}}}{\boldsymbol{T_t + T_{max}}} - \frac{\boldsymbol{p_{max}}}{\boldsymbol{T_t \xi V_{min}}}\right]$	
to a first approximation the terms in bold can be omitted.	It follows $V_b \ge \approx 2$ cm^3. Taking the minimum volume $0.074 \ge n$ $n \ge 0.034$
Calculation of the amount of substance n to condense in the thermometer: The gas is stored at P_r in the room-temperature reservoir of volume V_T. When the substance is condensed in the thermometer bulb at a temperature T_f, a residual $n_0 = p_f V_T / R_T$ remains in V_T. Therefore, in order to condense a quantity N, one must have in the system	In order to seal in 0.05 mol, the filling system must contain 0.069 mol.
$n' = \frac{p_r V_T}{RT_r}\left[1 - \frac{p_f}{p_r}\right] + \left(V_b + V_c + V_r\right)$	

V^L = volume of the liquid phase; ρ, density; p, pressure; V, volume; subscript r, room temperature; c, capillary; b, bulb.

error, allowing for a shorter total length of the capillary. Also, the bulb volume itself changes with temperature, due to thermal expansion and, to a much smaller extent, to the change in internal pressure. Design and fabrication criteria and measuring procedures are given in great detail in [2]. The case where the gas thermometer is *calibrated* at a number of fixed points is also described in [2], with a discussion of the simplification in the use of the gas thermometer introduced with this instrument (called the *interpolating CVGT* (ICVGT), defined in the ITS-90 for use between 3 K and 26 K), though it is not

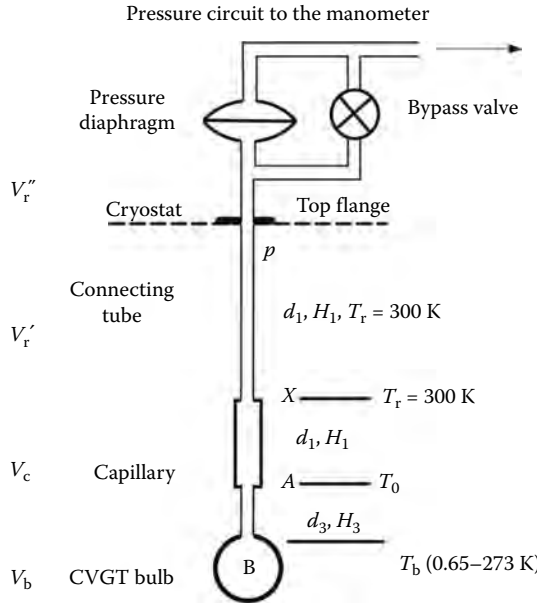


FIGURE 75.4 General layout of a constant-volume gas thermometer with room temperature diaphragm pressure transducer. d and H = pipe diameter and length.The temperature gradient should be concentrated in a short length of capillary tube AX whose diameter must be larger to limit the thermomomecular effect, if placed horizontally, the aerostatic head pressure gradient can also be avoided. In order to reduce the correction due to the "dead-volume" (sum of V_c, V_r', and V_r'') and its uncertainty, the bulb volume must be increased significantly.

more a thermodynamic thermometer, but a semiempirical one, providing only T_{90} values. Tables 75.4 and 75.5 summarize the most important differences between an ICVGT and a CVGT.

75.2.2 Acoustic Gas Thermometer

Figure 75.5 shows the layout of this type of thermometer: only the bulb is shown since all the rest of the gas circuit is much less critical than for the CVGT. For details, see [2].

The main problems to be taken into account in this type of gas thermometer are the following:

Correction of the effect of the acoustic boundary layer: Ideally, the normal component of the particle velocity is expected to vanish at the boundary (wall), and for a viscous fluid (gas) also, the tangential component should vanish. Further, the solid boundary should disturb the temperature field associated with the velocity field. Thus, near the interface, the acoustic propagation would be more isothermal than adiabatic, due to the very high thermal conductivity of the wall with respect to the gas. The calculation of this correction requires the knowledge of the density and heat capacity of the gas as a function of temperature and pressure. These can be obtained by inversion of the second-order virial equation. The importance of the boundary layer correction is indicated by the fact that it leads overall to the largest correction and that it can assume values as high as 5×10^{-4}.

Measurement of the acoustic thermometer volume: The use of a spherical bulb (the resonator) allows the employment of resonances and an improved quantification of the boundary layer effect through an appropriate choice of the available acoustic modes. The volume of the resonator can be measured by filling it with mercury and then weighing it. A much better method is to measure the radius of the resonator and its variation due to thermal expansion, by the application of microwaves. The microwave measurements can be performed separately in vacuum. Using a quasi-spherical resonator, in order to

TABLE 75.3 Summary of Design Criteria for an Absolute CVGT in the Low-Temperature Range ($T < 273.16$ K)

1. Choice of temperature range and of span $T_{min} \leftrightarrow T_{max}$
   This choice is preliminary to the choice of most of the design parameters.

   Below 273.16 K, ^{4}He gas thermometry is limited down to 2.5 K. With ^{3}He, accurate virial corrections available down to 1.5 K.

   Only CVGTs of special design can be used in full span. Being that $p \propto T$, the 2.5–273.16 K range corresponds to $p_{max}/p_{min} > 100$. For top accuracy, $\delta p/p < 0.01\%$, corresponding at p_{min} to $\delta p < 10^{-6}$. p_{max}, generally not achievable.

   Being that $p \propto = n/V$, molar density must generally be changed over the range to optimize accuracy, but n/V must be limited to avoid third virial correction, especially below = 2 K.

   In general, a CVGT is designed for work only below or only above a temperature between 25 K and 100 K.

2. Choice of reference temperature T_0

   Truly absolute thermometer: only one choice possible—273.16 K.

   Two-bulb CVGT: avoids necessity to bring up to T_0 the bulb measuring $T_{min} > T < T_{max}$. Useful with thermometers designed for use at $T \ll T_0$.

   Single-bulb CVGT: Same bulb spans the entire range up to T_0.

   Low-temperature reference temperature T_0^* (≈from 25 K to 90 K)

   Single-bulb CVGT commonly used. T_0^* value assigned by an independent experiment and, therefore, not exact by definition. However, the additional uncertainty is a minor inconvenience with respect to the advantage of limiting bulb temperature within the span $T_{min} \leftrightarrow T_{max}$.

3. Choice of thermometric gas and filling density
   Thermometric gas

   Nitrogen: low-medium accuracy.

   e-Hydrogen: not used for over 50 years but still suitable for low-medium accuracy and temperature range above ≈20 K.

   Helium-4: commonly employed in recent gas thermometers. Use limited to above 2.5 K.

   Helium-3: considered more in modern gas thermometry. Use presently limited to above 1.5 K; potential for use down to <1 K.

   Filling density: $p \propto n/V$ and $\mathrm{d}p/\mathrm{d}T \propto n/V$ (1 kPa K$^{-1} \triangleq$ 121 mol m^{-3}). Always advantageous increasing n/V, up to an upper boundary set by need of third virial correction. As a rule, $n/V < 250$ mol m^{-3} above = 2.5 K; $n/V < 160$ mol m^{-3}, down to 1.2 K; and $n/V < 30$ mol m^{-3} at 0.8 K.

4. Choice of the pressure measuring system
   See Table 75.1.

5. CVGT parameter design
   Room-temperature pressure transducer
   Bulb: top accuracy, 1 L volume typical; low accuracy, as low as 50 cm^3

   Dead volume: top accuracy, <10 cm^3; low accuracy, up to 10% of bulb volume

   Cryogenic pressure transducer
   Bulb: minimum volume limited only by its invariance, affected by adsorption and wall deflection. It can reduce to the transducer chamber

   Dead-volume: none (or very small)

6. Bulb design
   Volume may not be constant because of

   Compression modulus: Walls must be thick to limit deflection due to pressure, or bulb must be enclosed in a guard chamber kept at bulb pressure. Stress in bulb material must be relieved by annealing after machining.

   Thermal expansion: Nothing can be done to suppress this effect, except using glass; must be corrected for. Small effect below ≈30 K.

   Amount of "active" gas might not be constant because of

   Gas adsorption: Physicochemical interaction of bulb walls with the gas determines the amount adsorbed. Copper often gold plated to limit adsorption: this prevents heating the bulb above 50°C–70°C.

   Impurity molecules on walls and leaks: clean machining used for metal bulbs, followed by physicochemical cleaning. The bulb sealing gaskets must be stable in shape and leakproof at working temperatures.

(*continued*)

TABLE 75.3 (continued) Summary of Design Criteria for an Absolute CVGT in the Low-Temperature Range ($T < 273.16$ K)

7. Dead-volume design (case of room temperature pressure transducer)
   Dead-volume effect comes from combination of geometrical volume, working pressure, and gas density distribution, that is, from the amount of substance contained in it.

   Room-temperature dead volume: consists of all volumes of the gas measuring system at room temperature. Must be kept at uniform temperature (except diaphragm, often thermostated at ≈40 °C), to be measured within 0.1 °C–1 °C.

   Low-temperature dead volume: (part of) capillary tube between room and bulb temperatures. Temperature and density change from one end to the other. Tube diameter is a trade-off between geometrical volume and thermomolecular pressure effect: typical values between 0.5 mm and 3 mm. Advantageous in keeping the parts of tube where temperature variations occur as short as possible. For medium-high accuracy, temperature distribution must be known accurately.
8. Gas handling and measuring system (for non-sealed CVGTs)
   Handling system: must ensure purity, checked online with a mass spectrometer for the highest accuracy, and include gas recovery with cryogenic pumps and clean storage (or purification).

   Measuring system (case A): Separating diaphragm requires valve system for zero check, including constant-value valves and provisions to avoid (or to restore) thermometric gas losses and contamination from the manometric gas. For this purpose, a second diaphragm separator can be used.

resolve the degenerate microwave resonance triplets TE1n and TM1n, allows a reduction in the uncertainty: In this case, gas helium filling is used or argon; the use of different gases is an important check on the correctness of the corrections. They must also account for the heat exchange between the gas and the wall. Similarly, the use of different metals for the resonator allows a check on the correctness of the term accounting for the penetration of the e.m. field into the walls. Copper was used, as also stainless steel and aluminum, with the thermal conductivities of copper and stainless steel differing by a factor of about 300 at 77 K and of about 50 at the triple point of water. The use of helium has one drawback with respect to argon: it is much more sensitive to impurities. It is therefore advisable in such a case to apply a small flow of the measuring gas for the determination of impurities in a gas chromatograph, passing through a liquid-helium-cooled trap with the drawback, however, that the total amount of gas n is not perfectly constant anymore.

Effects from other acoustic parameters: Some effects include increase in acoustic noise from the bubbling helium bath, perturbation of the acoustic modes by the coil-terminated coaxial cables used for coupling microwaves into the cavity, and the condition of the inner surface of the resonator. It is expected that an electropolished surface, as compared with the as-machined one, would reduce the apparent precondensation of argon at 95 K and of helium at 4 K. High-purity argon is used at high temperatures, while for temperatures below, say, 100 K, the use of helium is more appropriate.

Effect of the amount of gas and of impurities in the gas: The influence of the quantity of gas enclosed in the resonator is by far less critical than in a CVGT, a great advantage that has been used to allow a real-time verification of an important influence factor, the effect of chemical impurities in the gas. These issues can be resolved with the apparatus including continuous (slow) flushing of the gas (3×10^{-5} to 3×10^{-4} mol s^{-1}) through the resonator and the outgoing gas being analyzed by an inline gas chromatograph, able to measure the presence of impurities at a level of 0.3×10^{-6} amount of substance fraction. Impurities affect helium much more than argon (10^{-6} of H_2O causes a shift of -3.93×10^{-6} and 0.12×10^{-6}, respectively). Also isotopic composition affects the measurements when argon is used.

TABLE 75.4 Differences and Simplifications in Design Criteria of an ICVGT with Respect to a CVGT

1. Choice of temperature range and span $T_{min} \ll T_{max}$
   An ICVGT can usefully replace other types of thermometers between 5 K and 30 K.

   ITS-90 defines two types, between 4.2 K and 24.5 K and between 3 K and 24.5 K.

   ICVGT definitions always assume constant molar density n/V in the whole range. According to the CVGT limits, T_{min} is set at ≈2.5 K for $n/V = 250$ mol m^{-3} and at ≈1.2 K for $n/V = 160$ mol m^{-3}. Pushing the definition as follows requires much lower density (i.e., sensitivity) values in *whole* range (e.g., for $T_{min} = 0.8$, K is $n/V = 30$ mol m^{-3}), or lower accuracy must be accepted.
2. Choice of definition range and of reference temperatures $T_{r,i}$
   An ICVGT is always of the single-bulb type. Its definition range can be extended below ≈1.2 K, down to 0.8 K using lower values of n/V, thus with lower accuracy.

   The number, N, of reference temperatures $T_{r,i}$, where the ICVGT has to be *calibrated*, set by the number of free parameters in definition equation. Fixed points are selected among the available (not always choice is possible) to minimize nonuniqueness, allow easiest realization, and allow maximum *extrapolation* interval (in order of importance).

   ITS-90 definitions: $n = 3$. Two triple points (Ne and e-H_2). The lowest is a ^{4}He or ^{3}He vapor-pressure point of ITS-90 definition.

   Other possible definitions (see details in Pavese and Molinar, first edition 1992): $n = 2$–4.

   Three main categories, using

   Only triple points: $n = 2$. Definitions with some stipulated parameters, in the 2–30 K range; only triple points of Ne and e-H_2.

   Only fixed points, not requiring pressure measurements: $n = 3$. Definitions in 2–30 K range; triple points of Ne and e-H_2 and the λ-point of ^{4}He.

   Only triple points and superconducting transitions: $n = 3$–4. Definitions in 0.8–30 K range; triple points of Ne and e-H_2 and superconducting transitions: zinc, aluminum, or indium.
3. Choice of thermometric gas and filling density
   Thermometric gas
   No difference with respect to a CVGT.

   Filling density: can be higher than with a CVGT, as calibration can take into account also the third virial coefficient effect.[a] For the ITS-90 definitions, the same limits of a CVGT apply.
4. Choice of pressure measuring system
   No difference with respect to a CVGT, but ICVGT calibration can also incorporate (re)calibration of some of the pressure transducer parameters.[a]
5. ICVGT parameter design
   Room-temperature pressure transducer
   Bulb: Volume V_b can be smaller than that of a CVGT, even for top accuracy, since calibration takes into account dead-volume effects.[a]
   Dead volume V_d: The ratio V_b/V_d can be smaller.

   Cryogenic pressure transducer
   No difference with respect to a CVGT.
6. Bulb design
   Variations of ICVGT volume V_b: due to wall modulus of elasticity and to thermal expansion, taken into account by calibration[a].

   Change of the quan.tity of "active" gas n_b: due to gas adsorption, taken into account by calibration[a], if reversible and reproducible. As regards to impurity molecules on the walls and to leaks, same problems as a CVGT.
7. Dead-volume design (case A)
   All dead-volume effects fully taken into account by calibration[a].
8. Gas handling and measuring system (for non-sealed CVGT)
   No difference with respect to a CVGT.

[a] Within the validity of the mathematical model used for the definition.

TABLE 75.5 Differences and Simplifications in Measurement Procedures for an ICVGT with Respect to a CVGT

1. Preliminary procedures and measurements
Only gas purity check is necessary. Then the ICVGT filling can be performed.
Realization of the fixed points: One or more thermometers must be calibrated at the fixed points realizing the ICVGT reference temperatures.
2. Measurements of the manometer pressure (p^*)
Measurement at the reference temperatures T_i: The bulb is first stabilized at each T_i. After capillary temperature distribution stabilization, which can take more time than bulb, pressure measurement and all manometric system corrections are performed; a p_i^* value is obtained. No auxiliary measurements need be done.
Measurement at any other temperature θ: same procedure; a p^* value is obtained. With the ICVGT, temperature distributions and, in general, all corrections taken into account by calibration procedure must remain the same in subsequent measurements, at least in between calibrations.
3. Calculation of temperature θ_b
The p_i^* values are *directly* used in the defining equation:
The numerical values of the equation coefficients are obtained by solving the system of equations obtained by substitution of the p_i^* values.
The defining equation (e.g., T_{90}) is then used with each p^* value to compute the corresponding value of θ_b of the bulb where thermometers are fitted.

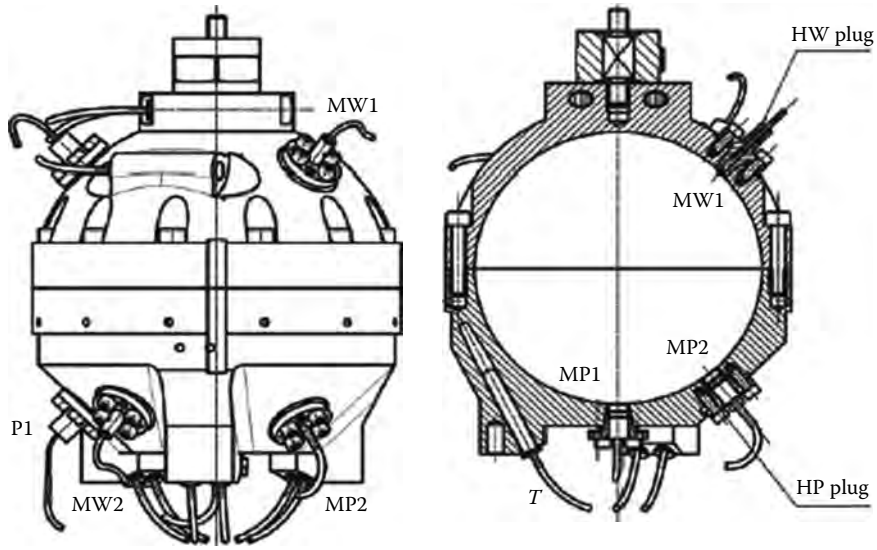


FIGURE 75.5 Layout of acoustic gas thermometers with quasi-spherical (tri-elliptical) resonator type. The symbols indicate the various acoustic, microwave and temperature transducers: MP, microphone; MW, microwave; *T*, thermometer. (After Gavioso, R. M. et al., *Int. J. Thermophys*., 32, 1339, 2011.)

References

1. R. E. Bedford, G. Bonnier, H. Maas, and F. Pavese, Techniques for approximating the ITS-90, Monograph 90/1 of the Bureau International des Poids et Mesures, Sevres: BIPM, 1997. http://www.bipm.org/en/publications/its-90_techniques.html (accessed on March 17, 2013).
2. F. Pavese and G. F. Molinar, *Modern Gas-Based Temperature and Pressure Measurements*, International monograph series on cryogenic engineering, New York: Plenum Publishing, 1992, pp. 514, and references therein; Augmented 2nd edn., 2012, New York: Springer, pp. 660, also available as e-Book.

3. J. F. Schooley, *Thermometry*, Boca Raton, FL: CRC Press, 1986.
4. T. J. Quinn, *Temperature*, 2nd edn., London, U.K.: Academic Press, 1990.
5. R. E. Bedford, G. Bonnier, H. Maas, and F. Pavese, Recommended values of temperature on the ITS-90 for a selected set of secondary reference points, *Metrologia*, 33, 133–154, 1996.
6. F. Pavese, Recalculation on ITS-90 of accurate vapour-pressure equations for e-H_2, Ne, N_2, O_2, Ar, CH_4 and CO_2, *J. Chem. Thermodynam.,* 25, 1351–1361, 1993.
7. R. M. Gavioso, G. Benedetto, D. Madonna Ripa, P. A. Giuliano Albo, C. Guianvarc'h, A. Merlone, L. Pitre, D. Truong, F. Moro, and R. Cuccaro, Progress in INRIM experiment for the determination of the Boltzmann constant with a quasi-spherical Resonator. *Int. J. Thermophys*. 32, 1339–1354, 2011.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

76

Temperature Indicators

Jan Stasiek
Technical University of Gdansk

Tolestyn Madaj
Technical University of Gdansk

Jaroslaw Mikielewicz
Institute of Fluid Flow Machinery

76.1 Melting and Shape/Size Changing Temperature Indicators.... **76**-1
Seger Cones • Thermoscope Bars • Bullers Rings • Temperature-Indicating Pellets, Liquids, and Crayons and Temperature Monitors
76.2 Color-Change Temperature Indicators..... **76**-11
Temperature-Indicating Paints and Crayons • Thermochromic Liquid Crystals
References..... **76**-16
Further Readings..... **76**-16
Partial List of Vendors and Suppliers..... **76**-16

Temperature indicators serve for approximate determination of bodies' temperatures and are used to control a variety of temperature treatment processes. The temperatures are determined based on knowledge of characteristic rated temperatures, which are mean critical temperatures of the indicator. However, it should be stressed that the accuracy of these measurements is satisfactory only if the measurement conditions are similar to the standard conditions for which the temperature indicators were calibrated. Otherwise, the critical temperatures of the indicators can be different from their rated temperatures listed in the standards, and the measurements can have considerable errors.

The temperature indicators can be classified into two groups, each group using different physical properties for the determination of the temperature. The indicators belonging to the first group melt at certain temperatures. For some of these indicators, such as pyrometric cones, thermoscope bars, and rings, the process of melting manifests itself as a shape/size deformation for which the temperature is determined by measuring the degree/rate of deformation of the indicator. For others, such as melting pellets, liquids, crayons, and monitors, the melting means turning entirely into a liquid smear. This can also be accompanied by color changing. The second group consists of color-change indicators containing pigments that at different temperatures show different colors by selectively reflecting incident white light. Among this group are reversible and irreversible paints, color-change crayons, and liquid-crystal indicators.

76.1 Melting and Shape/Size Changing Temperature Indicators

The latest British Standard BS 1041, part 7, 1988 [1], lists the following temperature indicators: Seger cones, thermoscope bars, and Bullers rings. Some temperature indicators used in the past—such as Watkin cylinders and Holdcroft bars—are no longer used and are of historical value only. In the United States, melting pellets, crayons, liquids, and monitors are available on the market and widely used.

The rated temperature for Seger cones (pyrometric cones) and thermoscope bars is defined by appropriate shape deformation resulting from the transformation of a certain amount of the indicator substance from solid to liquid state. For chemically pure elements and compounds at a constant pressure, the temperature during the entire process of phase change remains constant. If the pressure changes

within the range of changes for atmospheric pressure, then the temperature changes are insignificant and can be neglected even during precise measurements. The pyrometric cones and thermoscope bars are prepared from complex mixes of frits, fluxes, clays, calcium and magnesium compounds, silica, etc. The melting temperature of the indicators also referred to as the critical or rated temperatures varies with the proportions of the aforementioned compounds. Therefore, a set of indicators differing in the proportions of the compounds is capable of covering a required range of rated temperatures. The melting temperatures can also change, to a degree, with the proportion of phases. For mixes that constitute pyrometric cones and thermoscope bars, the temperature difference between the beginning and the end of the melting process can be as large as 25 °C–40 °C. At the rated temperature, one can assume that the temperature is either at the beginning or at the end of the melting process. In practice, an intermediate value is assumed, referring to an expected shape deformation of the temperature indicator.

For Bullers rings, the rated temperature is determined by a shape deformation that can be described as temperature shrinkage. An indication of the required temperature is a proper contraction of the outer diameter of the ring made of special clay (a mix of appropriate materials) that contracts uniformly with the increase in temperature throughout the operating range.

For melting pellets, crayons, liquids, and monitors, the rated temperature is that of the beginning of the melting process when the indicator turns entirely into a liquid smear. Usually, on cooling, the liquid mark solidifies and becomes glossy transparent or translucent in appearance. The entire process can be accompanied by a change in color—mostly because the color of the workpiece surface or the back of an adhesive label, which enables the contact of the indicator with the surface, will show up from under the transparent mark. However, the moment of melting—not a color change—is the temperature signal.

76.1.1 Seger Cones

The pyrometric cones are typically slender, truncated, trihedral pyramids, about 25–60 mm in height. The base of the pyramid is a regular triangle of side 7–16 mm. One edge of the pyramid is vertical or slightly leaned outward (see Figure 76.1a). The recommended height of the standard cones is 60 mm; the laboratory cones are 30 mm high.

The pyramids are manufactured by pressing a powder mixture of a number of minerals mixed in different proportions throughout the required range of rated temperatures. The main components are silicon oxide (SiO_2), aluminum oxide ($A1_2O_3$) with additives in the form of oxides (MgO, K_2O, Na_2O, CaO, B_2O_3, PbO), and an organic binder. The following equation is an example of chemical constitution of the pyrometric cone for the temperature range of 600 °C–900 °C:

$$X(2SiO_2 + A1_2O_3) + (1 - X)(0.5Na_2O + 0.5PbO + B_2O_3 + 2SiO_2)$$

where X is a mass unit. The pyramids designed for higher rated temperatures are prepared based on similar equations.

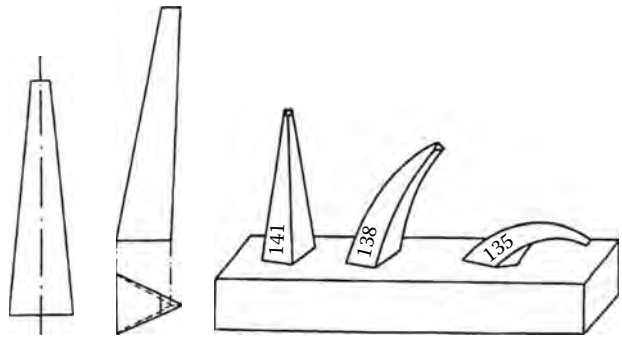


FIGURE 76.1 Pyrometric cones (a) in cross sections and (b) on a plaque during firing.

76.1.1.1 Touchdown Temperatures for Seger Cones

As the heating progresses, the cone used for the measurement begins to soften and bends until its tip touches down on the surface on which it was placed. The rated temperature referring to this deformation is called the *touchdown temperature*. The touchdown temperatures for the Seger cones are determined in an electric kiln with clean atmospheric air at a heating rate of 60 °C h^{-1}.

According to the earlier German standard DIN 51063 [2], the range of touchdown temperatures from 600 °C to 2000 °C at 10 °C–50 °C steps is covered by a series of Seger cones denoted traditionally by numbers from 022 to 42 (see Table 76.1).

According to the latest British Standard BS 1041, part 7, 1988, the touchdown temperatures within the range of 600 °C–1535 °C at temperature intervals of 15 °C–35 °C are realized by a series of Seger cones numbered from 022 to 20 (see Table 76.1).

TABLE 76.1 Approximate Touchdown Temperatures of Pyrometric Cones

Cone No.	Temperature (°C)	Cone No.	Temperature (°C)	Cone No.	Temperature (°C)	Cone No.	Temperature (°C)
DIN 51063							
022	600	07a	960	9	1280	29	1650
021	650	06a	980	10	1300	30	1670
020	670	05a	1000	11	1320	31	1690
019	690	04a	1020	12	1350	32	1710
018	710	03a	1040	13	1380	33	1730
017	730	02a	1060	14	1410	34	1750
016	750	0la	1080	15	1435	35	1770
015a	790	la	1100	16	1460	36	1790
014a	815	2a	1120	17	1480	37	1825
013a	835	3a	1140	18	1500	38	1850
012a	855	4a	1160	19	1520	39	1880
011a	880	5a	1180	20	1530	40	1920
010a	900	6a	1200	26	1580	41	1960
09a	920	7	1230	27	1610	42	2000
08a	940	8	1250	28	1630		
BS 1041							
022	600	011	880	1	1135	11	1310
021	615	010	900	2	1150	12	1330
020	630	09	925	3	1165	13	1350
019	665	08	950	4	1180	14	1380
018	700	07	975	5	1195	15	1410
017	730	06	1000	6	1210	16	1435
016	760	05	1030	7	1230	17	1460
015	790	04	1060	8	1250	18	1485
014	810	03	1085	9	1270	19	1510
013	830	02	1105	10	1290	20	1535
012	860	01	1120				

Notes: (1) Each temperature given in the table is that at which the tip of a cone will bend sufficiently to touch the base in an electric kiln with a heating rate of 60°C h^{-1}. (2) The touchdown temperature depends on the rate of heating: reports on firing behavior should quote the cone number, not the temperature taken from the aforementioned table. (3) Intermediate degrees of bending can be referred to the hands of a clock; e.g., 3 o' clock would represent a cone bent halfway to the stand.

The precision of determination of the touchdown temperatures for the industrial cones should be better than ±15 °C; for the laboratory cones, better than ±10 °C. If the heating rate undergoes change within the range of 20 °C h^{-1} to 150 °C h^{-1}, then the rated temperatures can change by −40 °C for the above lower limiting value of the heating rate up to 60 °C for the upper value. A 0.35% content of SO_2 in the atmosphere increases the rated temperatures by about 35 °C. Also, the presence of soot in the atmosphere slightly raises the rated temperatures.

How to use the materials: While single cones are sometimes used, usually three or four consecutively numbered cones, including a cone whose rated temperature is equal to the required temperature of the heat treatment and two cones of neighboring numbers (one less and one more), are employed for the temperature determination (see Figure 76.1b). They are installed into specially unfired plaques with tapered holes and protrusions that hold the cones firmly. The plaques are mounted to a workpiece surface to allow observations. A cone can be set up in other ways, such as inserting its base into refractory clay. However, it is necessary to assure a correct angle and firm hold of the cones during the firing cycle. Failure in these respects will cause the cone to bend unpredictably and give incorrect assessment of the heat treatment. If the process of heating takes place with a standard heating rate, then the rated temperature is reached when the tip of the central cone touches the base of the plaque. With further prolongation of the firing cycle, the cone will melt completely to form a blob on the plaque. The process of reaching the rated temperature is signaled in advance by the cone of one-less number. The cone of one-more number is there to prove that the required temperature value is not exceeded. Placing a series of cones with lower numbers (lower rated temperatures) provides the opportunity to carry out the process of heating at a required rate.

Typical application: Seger cones are used for the control of firing processes in the ceramics industry and artistry.

76.1.2 Thermoscope Bars

These indicators have the shape of bars of rectangular cross sections. The typical dimensions of the bars are length, 57 mm; width, 8 mm; and height, 6 mm. Bars of consecutive numbers (rated temperatures) are placed horizontally on a refractory stand as in Figure 76.2. The set of thermoscope bars is a more convenient and slightly modified form of Holdcroft bars. The bars are made of the same composites (mineral mixes and organic binder) as the pyrometric cones. The mixed powders are pressed and can be hardened by prefiring at relatively low temperatures, below those at which bending should occur.

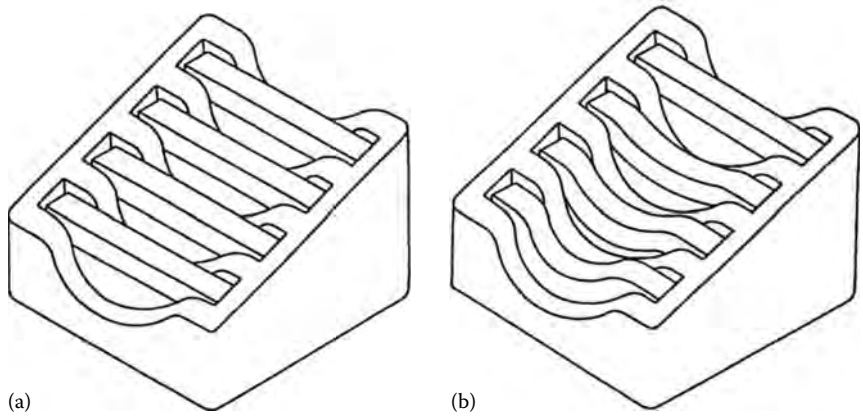


FIGURE 76.2 Thermoscope bars on a stand (a) before and (b) after firing.

TABLE 76.2 Approximate Bending Temperatures of Thermoscope Bars (BS 1041)

Bar No.	Temperature (°C)	Cone No.	Temperature (°C)	Cone No.	Temperature (°C)	Cone No.	Temperature (°C)
1	590	12	870	23	1130	33	1300
2	610	13	890	24	1145	34	1320
3	625	14	915	25	1160	35	1340
4	650	15	940	26	1175	36	1365
5	685	16	965	27	1190	37	1395
6	715	17	990	28	1205	38	1425
7	745	18	1015	29	1220	39	1450
8	775	19	1045	30	1240	40	1475
9	800	20	1075	31	1260	41	1500
10	820	21	1095	32	1280	42	1525
11	845	22	1115				

Notes: (1) Each temperature given in the table is that at which the bar starts to bend in an electric kiln with a heating rate of 60 °C h^{-1}. (2) The bending temperature depends on the rate of heating: reports on firing behavior should quote the bar number, not the temperature taken from the aforementioned table. (3) The bar can be expected to bend sufficiently to touch the stand at a temperature of 10 °C–30 °C higher than the values given in the table, depending on the composition of the bar.

76.1.2.1 Bending Temperatures of Thermoscope Bars

The rated temperatures of thermoscope bars, referred to as the bending temperatures, are found during the calibration in an electric kiln with a heating rate of 60 °C h^{-1} when the bars start to exhibit deformation (i.e., begin to bend).

According to the British Standard BS 1041, the range of rated temperatures from 590 °C to 1525 °C at temperature intervals of 15 °C–35 °C is covered by 42 thermoscope bars (see Table 76.2).

The precision of determination of the bending is about ±15 °C; the other properties of the thermoscope bars referring to changes of the standard conditions are the same as for the pyrometric cones.

How to use the materials: Four thermoscope bars of consecutive numbers—the first two having lower bending temperatures, the third one having the bending temperature equal or close to the required temperature, and the fourth one having a higher bending temperature—are placed in sequence on a special refractory stand as in Figure 76.2a. The set is mounted to the workpiece surface where observations take place. If the process of heating takes place with a standard heating rate, then the beginning of deformation (bending) of the third bar indicates that the required temperature is reached. The process of reaching the rated temperature for the third bar is signaled in advance by the deformation of the proceeding bars whose behavior allows the evaluation of the heating rate. The unbent fourth bar testifies that the required temperature is not exceeded (see Figure 76.2b).

Typical application: The application of the thermoscope bars is identical to that of the Seger cones.

76.1.3 Bullers Rings

These temperature indicators in the form of rings have the following dimensions: outer diameter, 63 mm; inner diameter, 22 mm; and width, 8 mm. The appropriate measuring unit consists of a Bullers ring and a specially prescaled device for measurement of the temperature shrinkage of the ring. This contraction gage measures the outer diameter of the heated ring, based on which the

heating work is assessed. The full measuring range of rated temperatures from 960 °C to 1440 °C is covered by four types of rings manufactured by pressing powders of ceramics mixes, with a binder and without prefiring.

1. Rings denoted as 55 of brown color, suitable for temperatures from 960 °C to 1100 °C, are used in the firing of glost ware and common building bricks where the finishing temperatures are relatively low.
2. Rings numbered 27/84, colored green, suitable for temperatures from 960 °C to 1250 °C, are used for firing earthenware at the medium finishing temperatures, as well as tiles and bricks refractory with respect to low temperatures.
3. Rings numbered 75/84, colored natural, recommended for firing temperatures from 960 °C to 1320 °C, are allowed for higher-finishing temperatures and are used for firing electrical porcelain, china, grinding wheels, and bricks refractory with respect to higher temperatures.
4. Rings numbered 73, colored yellow, recommended for temperatures from 1280 °C to 1440 °C for slow firing conditions, are used in the manufacture of high-temperature ceramics and heavy refractories.

Approximate rated temperatures for the Bullers rings and corresponding readings of the contraction gage according to the British Standard BS 1041 are presented in Table 76.3.

How to use the materials: One or more rings of the same type are placed vertically in a prefired stand and mounted to a workpiece surface. To determine the temperature as the firing progresses, the heated rings are withdrawn from their stands, cooled to the ambient temperature, and then measured for contraction. This measurement is carried out on a gage consisting of a brass baseplate on which a radial arm with a pointer moving over a scale and the two steel dowel pins, against which the ring is pressed by the movable arm, are mounted (see Figure 76.3). A contraction of the ring gives rise to an amplified movement of the pointer over the scale. The divisions on the scale are numbered from −5 to 60. More heavily fired rings contract more and give higher readings on the gage. The divisions below 0 indicate expansion of the ring; above 0 indicates contraction. Rings should be measured across several diameters by turning them around in the gage so as to find the mean value to which the temperature can be assigned with the help of Table 76.3. Placing several rings in the stand in a manner that allows their easy withdrawal gives the possibility of measuring the heating rate. In a similar way, distribution of a number of rings throughout the furnace enables the determination of the temperature field in the furnace.

Usually, one or more test pieces from a series of rings are picked out for the sake of calibration so as to compare the obtained readings with the standard values enclosed in Table 76.3. Intermediate measurements are also carried out to evaluate the effect of the heating rate on the temperature shrinkage of the rings. The accuracy of the temperature determination for the standard heating conditions is ±0.5 of a single division of the scale.

Typical application: The application of the Bullers rings is similar to the Seger cones and thermoscope bars. An inconvenience is the fact that the Bullers rings require gage measurements and the temperature cannot be solely determined based on naked-eye observations.

76.1.4 Temperature-Indicating Pellets, Liquids, and Crayons and Temperature Monitors

76.1.4.1 Temperature-Indicating Pellets

Temperature-indicating pellets are manufactured by pressing powders of mineral mixes of certain melting temperatures and an indifferent binder. Melting pellets are recommended as standard tablets $\phi 7/16 \times 1/8$ and miniature tablets $\phi 1/8 \times 1/8$ (see Figure 76.4). There are 112 different pellets that cover

TABLE 76.3 Approximate Rated Temperatures for Bullers Rings

	Gage Readings			
Temperature (°C)	Ring No. 55	Ring No. 27/84	Ring No. 75/84	Ring No. 73
960	3	0	0	
970	7	1	1	
980	11	2.5	2	
990	15	4	3	
1000	18	5.5	4	
1010	21	7	5	
1020	24	8.5	6	
1030	27	10	7	
1040	30	11.5	8.5	
1050	32	13	10	
1060	34	14	11	
1070	36	15.5	12.5	
1080	37	17	14	
1090	38	18.5	15.5	
1100	39	20	17	
1110		21.5	18	
1120		23	20	
1130		24.5	21	
1140		26	22	
1150		27	23	
1160		28.5	24.5	
1170		30	26	
1180		31.5	27	
1190		33	28	
1200		34.5	29	
1210		36	30	
1220		37.5	31	
1230		38.5	32	
1240		40	33	
1250		41.5	34.5	
1260			36.5	
1270			38.5	
1280			40	29.5
1290			42	30
1300			44	31
1320			46	34
1340				37
1360				40.5
1380				44
1400				48
1420				51
1440				54

Notes: These values should be used with caution because they are dependent on the firing cycle to which the rings are subjected.

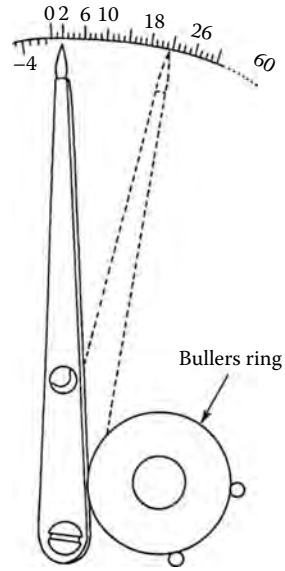


FIGURE 76.3 Contraction gage.

the temperature range from 40 °C (100 °F) to 1650 °C (3000 °F); see Table 76.4. The accuracy of the temperature determination is ±1% of the rated temperature. There are also available pellets for temperature control in strongly reducing atmospheres.

How to use the materials: A pellet of the rated temperature equal to the required temperature of heat treatment is placed on the investigated surface before the heating starts. When the heating progresses, the beginning of melting signals that the rated temperature is reached. Placing more pellets with the rated temperatures lower and higher than the required temperature enables more precise control of the heating process.

Typical application: Typical applications are checking furnace temperatures, control of heat treating of large units, as well as other applications involving long-duration heating.

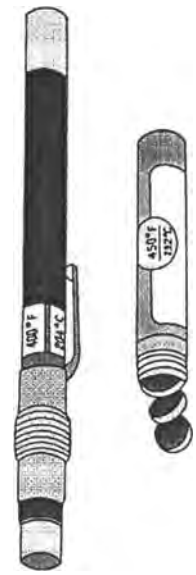


FIGURE 76.4 Temperature-indicating crayon and pellets.

TABLE 76.4 Rated Temperatures for Temperature-Indicating Pellets, Liquids, and Crayons

°F	°C	°F	°C	°F	°C
100	38	325	163	1200[a]	649
103	39	331	166	1250[a]	677
106	41	338	170	1300[a]	704
109	43	344	173	1350[a]	732
113	45	350	177	1400[a]	760
119	48	363	184	1425	774
125	52	375	191	1450[a]	788
131	55	388	198	1480	804
138	59	400	204	1500[a]	816
144	62	413	212	1550	843
150	66	425	218	1600	871
156	69	438	226	1650	899
163	73	450	232	1700	927
169	76	463	239	1750[a]	954
175	79	475	246	1800	982
182	83	488	253	1850	1010
188	87	500	260	1900[a]	1032
194	90	525	274	1950	1066
200	93	550	288	2000	1093
206	97	575	302	2050	1121
213	101	600	316	2100	1149
219	104	625	329	2150[a]	1177
225	107	650’	343	2200[a]	1204
231	111	675	357	2250[a]	1232
238	114	700	371	2300[a]	1260
244	118	725	385	2350[a]	1288
250	121	750[a]	399	2400	1316
256	124	800[a]	427	2450	1343
263	128	850[a]	454	2500[a]	1371
269	132	900	482	2550[b]	1390
275	135	932	500	2600[b]	1427
282	139	950	510	2650[b]	1454
288	142	977	525	2700[b]	1482
294	146	1000	538	2800[b]	1538
300	149	1022	550	2900[b]	1593
306	152	1050[a]	566	3000[b]	1649
313	156	1100	593		
319	159	1150	621		

[a] Series “R” pellets for use in strongly reducing atmospheres.
[b] Available in pellets only.

76.1.4.2 Temperature-Indicating Liquids

Temperature-indicating liquids are solutions of powdered mineral mixes in indifferent highly volatile solvents. They are available for use by brushing or spraying. There are over 100 different liquids with the rated temperatures from 40 °C (100 °F) to 1371 °C (2500 °F); see Table 76.4. The accuracy of the temperature determination is ±1% of the rated temperature.

How to use the materials: A thin coating of the liquid is put on the clean and dry surface by brushing or spraying before the heating starts. It dries almost instantly to a dull opaque mark. When the required temperature is reached, this mark liquefies. The melted coating does not revert to its original opaque appearance but remains glossy transparent on cooling. It should be noted that color changes do not signal the required temperature. The melting, not the color change, is the temperature signal.

Typical application: They are recommended for temperature control on fabrics, rubber, plastics, and smooth surfaces such as glass or polished metals, as well as for monitoring critical temperatures in electronic fields.

76.1.4.3 Temperature-Indicating Crayons

Temperature-indicating crayons are sticks manufactured from powders of mineral mixes of certain melting temperatures and an indifferent binder. The crayons are put in specially made adjustable metal holders with labels saying their rated temperatures; see Figure 76.4. Similar to the temperature-indicating liquids, there are over 100 different crayons that cover the temperature range from 40 °C (100 °F) to 1371 °C (2500 °F); see Table 76.4. The accuracy of the temperature determination is also ±1% of the rated temperature.

How to use the materials: During heating, the workpiece should be struck repeatedly by the crayon. Below its rated temperature, the crayon leaves a dry opaque mark. When the rated temperature is reached, the crayon leaves a liquid smear. On cooling, the liquid mark will solidify with a transparent or translucent appearance. For temperatures below 700 °F, the mark can be put on the workpiece surface before the heating process. The mark will liquefy when the rated temperature is reached. It should be remembered that the moment of melting, not any change in color, is the temperature signal.

Typical application: The crayons can be applied in welding, forging, heat treating and fabrication of metals, and molding of rubber and plastics, wherever the workpiece is accessible during the heating process. Very smooth surfaces are excluded.

76.1.4.4 Temperature Monitors (Labels)

These temperature indicators are adhesive-backed labels with one or more heat-sensitive indicators under transparent circular windows. The indicators turn black (show black paper backing) when the rated temperature is reached. The rated temperatures are from 40 °C (100 °F) to 320 °C (600 °F). They are available as single temperature or multi-temperature indicators with 10°, 25°, or 50° steps. The tolerance of the temperature determination is ±1 °C (±1.8 °F) below 100 °C and ±1% of the rated temperature above 100 °C. Exemplary rated temperatures for a series of 4-temperature (4 dots) indicators are presented in Table 76.5. A 4-dot temperature monitor is displayed in Figure 76.5.

How to use the material: After removing the backing, the label is pressed firmly to the dry and clean workpiece surface. A change in color to black is the temperature signal. Application of multi-temperature-indicating labels allows more precise temperature determination.

Typical application: They are especially applied for monitoring the safe operating temperature of equipment and processes, safeguarding temperature-sensitive materials during storage and transport.

The melting temperature indicators are described in the catalogs of their manufacturers [3,4]. As the melting indicators are widely used in the United States, temperatures in Fahrenheit are also given.

TABLE 76.5 Rated Temperatures for 4-Temperature (4 Dots) Labels

Model No.	°F	°C	°F	°C	°F	°C	°F	°C
4A-100	100	38	110	43	120	49	130	54
4A-110	110	43	120	49	130	54	140	60
4A-120	120	49	130	54	140	60	150	66
4A-130	130	54	140	60	150	66	160	71
4A-140	140	60	150	66	160	71	170	77
4A-150	150	66	160	71	170	77	180	82
4A-160	160	71	170	77	180	82	190	88
4A-170	170	77	180	82	190	88	200	93
4A-180	180	82	190	88	200	93	210	99
4A-190	190	88	200	93	210	99	220	104
4A-200	200	93	210	99	220	104	230	110
4A-210	210	99	220	104	230	110	240	116
4A-220	220	104	230	110	240	116	250	121
4A-230	230	No	240	116	250	121	260	127
4A-240	240	116	250	121	260	127	270	132
4A-250	250	121	260	127	270	132	280	138
4A-260	260	127	270	132	280	138	290	143
4A-270	270	132	280	138	290	143	300	149

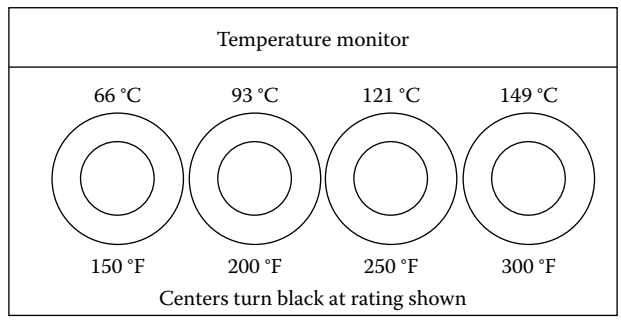


FIGURE 76.5 Four-dot label.

76.2 Color-Change Temperature Indicators

Color-change indicators comprise temperature-indicating paints, crayons, as well as liquid-crystal indicators.

76.2.1 Temperature-Indicating Paints and Crayons

Temperature-indicating paints are basically acrylic lacquers containing finely dispersed temperature-sensitive inorganic pigments. The principle of operation of these indicators draws on the change in color of incident light reflected from the surface of the paint due to chemical reactions that the dispersed pigments undergo and creation of new compounds at specific transition temperatures. The color-change temperatures are also determined by the heating time. According to the British Standard, it is assumed that the rated temperatures of the paints correspond to the change in color at a heating interval of 10 min.

To make the characteristics of temperature-indicating paints complete, the manufacturers also provide, together with the paints, graphs of trigger temperature versus heating-time relationships.

Temperature-indicating paints and crayons can be divided into two groups:

- Irreversible indicators: where the change of color becomes permanent.
- Reversible indicators: after cooling and some time, they revert to previous colors.

Irreversible color-change indicators are complex compounds containing various metals, including cobalt, chromium, molybdenum, nickel, copper, vanadium, or uranium. However, they are lead-free and sulfur-free. They are available on the market in the form of paints and crayons.

76.2.1.1 Irreversible Color-Change Paints

The irreversible paints can change color once or several times during the heating process. Thus, another division can be made on single-change and multichange paints. The range of rated temperatures is from 40 °C to 1350 °C at 10 °C–200 °C steps. At the standard conditions, the tolerance of measurements is ±5 °C for lower temperature values and ±1% for higher temperatures. Exemplary single-change paints with two critical temperatures—the initial trigger temperature for which the paint changes color after 10 min heating and the cutoff temperature being the lowest temperature for which the color change is achieved for long-duration heating—are collected in Table 76.6. Color changes and critical temperatures for some multichange paints (changing color 2, 3, or 6 times throughout the heating cycle) are presented in Table 76.7.

How to use the materials: A thin layer of the paint is applied to a workpiece surface by brushing or spraying like an ordinary paint and allowed to dry before the heating starts. During the heating, when a point of the surface reaches or exceeds the critical temperature, a color change will take place. To determine the distribution of temperature, a multichange paint can be applied. With a nonuniform temperature rise, a number of colored bands separated by isothermal lines will appear on the workpiece surface, allowing the thermal record to be made of the temperature gradient across the surface.

Typical application: The temperature-indicating paints are widely used in industrial applications for observing heat patterns, detecting high- and low-temperature points on surfaces of heat engines, pipelines, and refrigeration fins. They can be also used for controlling temperatures of powered elements and surfaces that are inaccessible or revolve at high speeds.

TABLE 76.6 Single-Change Paints

Original Color	Signal Color	Initial Trigger Temperature[a] (°C)	Cutoff Temperature (°C)
Pink	Blue	48	30
Pink	Blue	135	110
Mauve pink	Blue	148	120
Blue	Dark green	155	46
Yellow	Red	235	180
Blue	Fawn	275	150
Mauve red	Gray	350	220
Mauve	White	386	290
Green	Salmon pink	447	312
Green	White	458	312
Orange	Yellow	555	482
Red	White	630	450

[a] Color-change temperature for 10 min heating.

TABLE 76.7 Multichange Paints

Original Color	Signal Color	Initial Trigger Temperature[a]	Cutoff Temperature °C
Light tan	Bronze green	160	150
Bronze green	Pale Indian red	230	210
Reddish orange	Dark gray	242	193
Dark gray	Medium gray	255	211
Medium gray	Dirty white	338	228
Purple	Pink	395	355
Pink	Fawn	500	386
Fawn	Blue	580	408
Red	Dusty gray	420	310
Dusty gray	Yellow	555	328
Yellow	Orange	610	450
Orange	Green	690	535
Green	Brown	820	621
Brown	Green/gray	1050	945

[a] Color-change temperature for 10 min heating.

76.2.1.2 Color-Change Crayons

Color-change crayons, available in more than 10 distinct colors, similar in shape to regular crayons for drawing, cover the temperature range from 65 °C to 670 °C at 10 °C–100 °C temperature intervals. Exemplary single-change crayons are presented in Table 76.8. The accuracy of the temperature determination is the same as for the temperature-sensitive paints. They can be used for evaluating the temperature on already heated surfaces. They change color 2 min after reaching the rated temperature. Easy to use and inexpensive, they are invaluable for occasional temperature control in auto repairs, soldering, welding, electrical wiring, and enameling and for any operation involving boiling, baking, and other forms of heating.

TABLE 76.8 Single-Change Crayons

Original Color	Signal Color	Initial Trigger Temperature (°C)
Ivory	Light green	65
Yellow and green	Light green	75
Light pink	Blue	100
Gray and white	Light blue	120
Light ivory	Pink	150
Light blue	Black	200
Green	Black	280
Light green	Gray and brown	300
Blue	White	320
Brown	Red orange	350
White	Yellow	410
Light pink	Black	450
Ochre	Black	500
Blue	White	600
Green	White	670

76.2.1.3 Reversible Color-Change Indicators

Reversible color-change indicators are available on the market as paints and in label form. The thermal pigments of these temperature indicators are mercury-based complexes. Therefore, they cannot be applied directly to metal surfaces as this causes decomposition. They also tend to decompose after long exposure to heat, but the decomposition can be retarded by using a clear over-lacquer. The pigments find their most successful application when encapsulated into labels.

The rated temperatures for the reversible color-change paints do not exceed 170 °C. For temperatures up to 70 °C under standard conditions, the tolerance of measurements is ±1 °C; for 70 °C–150 °C, ±2 °C; and for 150 °C–170 °C, ±3 °C.

How to use the materials: A thin layer of a reversible paint is applied to a workpiece surface by brushing or spraying, or a label is pressed to the surface. During the heating, a color change will take place when the temperature of the surface reaches or exceeds the critical temperature.

Typical application: The reversible color-change paints are widely used in the electrical industry, especially on bus bars, live conductors, and connectors in high-current switches and in electronic fault-finding. They also find application as warning and indicating devices of domestic appliances. They are invaluable for controlling lower temperatures when it is necessary to detect undesirable temperature excursions, correct faults, and revert to normal conditions.

76.2.2 Thermochromic Liquid Crystals

Liquid crystals constitute a class of matter unique in exhibiting mechanical properties of liquids (fluidity and surface tension) and optical properties of solids (anisotropy to light, birefringence). Certain liquid crystals are thermochromic and react to changes in temperature by changing color. They can be painted on a surface or suspended in a fluid and used to make the distribution of temperature visible. Normally clear or slightly milky in appearance, liquid crystals change in appearance over a narrow range of temperatures called the color-play bandwidth (the temperature interval between first red and last blue), centered around the nominal event temperature (mid-green temperature). The displayed color is red at the low-temperature margin of the color-play interval and blue at the high end. Within the color-play interval, the colors range smoothly from red to blue as a function of temperature; see Figure 76.6. Liquid crystals or mesophases have been classified as smectic, chiral nematic, cholesteric, and blue. The structure of liquid crystals is shown schematically in Figure 76.7.

76.2.2.1 Temperature-Sensitive and Shear-Sensitive Formulations

Temperature-sensitive liquid crystals show colors by selectively reflecting incident white light. Conventional temperature-sensitive mixtures turn from colorless (or black against a black background) to red at a given temperature and, as the temperature is increased, pass smoothly through the other colors of the visible spectrum in sequence (orange, yellow, green, blue, violet) before turning colorless (or black) again in the ultraviolet at a higher temperature. The color changes are reversible, and on cooling, the color-change sequence is reversed.

Temperature-insensitive (sometimes called shear-sensitive) formulations can also be made. These mixtures show just a single color below a given transition temperature (called the clearing point) and change to colorless (black) above it. The working temperature range is thus below the clearing point. Both reversible and hysteretic (memory) formulations can be made. All liquid-crystal mixtures should be viewed against nonreflecting backgrounds (ideally black, totally absorbing) for best visualization of the colors.

76.2.2.2 Color-Play Properties and Resolution

Temperature-sensitive thermochromic mixtures have a characteristic red-start or mid-green temperature and color-play bandwidth. The bandwidth is defined as the blue-start temperature minus the red-start temperature. The color play is defined by specifying either the red-start or mid-green temperature and the bandwidth. For example, R35C1W describes a liquid crystal with a red start at 35 °C and a

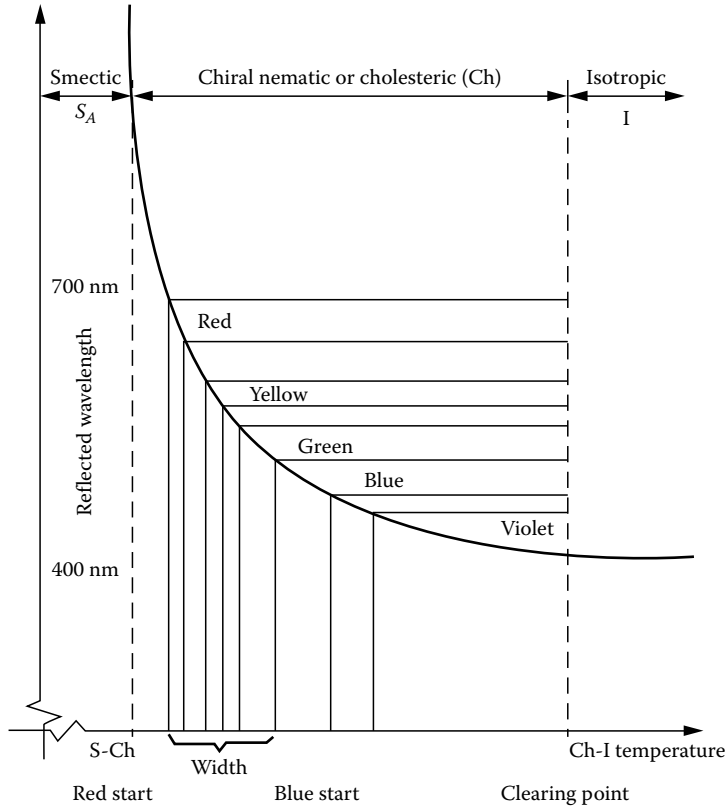


FIGURE 76.6 Typical pitch versus temperature response of thermochromic liquid crystals.

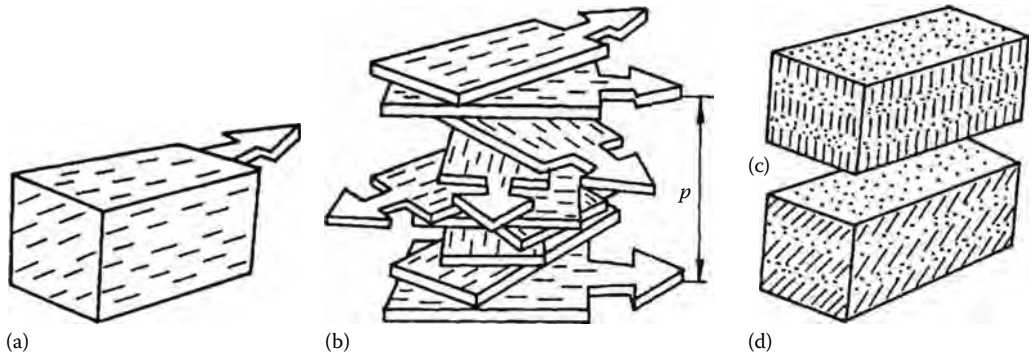


FIGURE 76.7 Structures of liquid crystals (a) nematic, (b) cholesteric, (c) smectic A, and (d) smectic B.

bandwidth of 1 °C, that is, a blue start 1 °C higher, at 36 °C; G100F2W describes a liquid crystal with a mid-green temperature at 100 °F and a bandwidth of 2 °F.

Both the color-play bandwidth and the event temperature of a liquid crystal can be selected by its proper chemical composition. The event temperatures of liquid crystals range from –30 °C to 115 °C with color-play bands from 0.5 °C to 20 °C, although not all combinations of event temperature and color-play bandwidth are available. Liquid crystals with color-play bandwidths of 1 °C or less are called narrowband materials, while those whose bandwidth exceeds 5 °C are referred to as wide band. The type of material to be specified for temperature indicating should depend very much on the type of available image interpretation technique—human observers and intensity-based image-processing or true-color

image-processing systems [5]. The uncertainty associated with direct visual inspection is about 1/3 the color-play bandwidth, given an observer with normal color vision—about ±0.2 °C to 0.5 °C. The uncertainty of true-color image-processing interpreters using wideband liquid crystals is of the same order as the uncertainty assigned to human observers using narrowband materials and depends on the pixel-to-pixel uniformity of the applied paint and the size of the area averaged by the interpreter (about ±0.05 °C can be achieved). Using a multiply filtered, intensity-based system, the resolution is better than ±0.1 °C.

How to use the materials: Liquid-crystal indicators can be used in a number of different forms: as unsealed liquids (also in solutions), in the microencapsulated form (as aqueous slurries or coating formulations), and as coated (printed) sheets. The different forms of the materials have selective advantages and suit different temperatures and flow-visualization applications. Individual products are described in more detail in relevant booklets issued by the manufacturers of liquid crystals [6].

Typical application: Liquid-crystal indicators are ideal for monitoring temperatures of electronic parts, transformers, relays, and motors. They are invaluable for a fast visual indication of temperatures.

References

1. BSI BS 1041-7: Temperature Measurements—Part 7: Guide to Selection and Use of Temperature/Time Indicators, 1988.
2. DIN EN 993-12: 1997-06: Methods of test for dense shaped refractory products—Part 12: determination of pyrometric cone equivalent refractoriness.
3. Temperature Products, Sensors, OMEGA, http://www.omega.com/temperature/tsc.html (accessed on March 14, 2013).
4. Tempil, Data Library, *Technical Data Sheets*, http://www.tempil.com/resources/ (accessed on March 14, 2013).
5. Moffat, R.J., Experimental heat transfer, *Proceedings of the 9th International Heat Transfer Conference,* Jerusalem, Israel, pp. 187–206, 1990.
6. LCR Hallcrest, *Products, Thermometry, Labels*, http://hallcrest.com/cat.cfm?cat_id=67 (accessed on March 14, 2013).

Further Readings

Roger Lindsay Hand, Temperature Indicators, http://www.devileye.net/catalog/laser_facial_rejuvenation/temperature_indicators.html, accessed on July 21, 2012.

Temperature and Humidity Indicators, eHow, http://www.ehow.com/list_7522896_temperature-humidity-indicators.html, accessed on July 26, 2012.

Partial List of Vendors and Suppliers

ISE, Inc., 10100 Royalton Rd., Cleveland, OH 44133, Phone: 440-237-3200, 800-642-1670 (toll free), Fax: 440-237-1744, www.instserv.com

Liberty Industries, 133 Commerce St., East Berlin, CT 06023-1106, Phone: 860-828-6361, 800-828-5656 (toll free), Fax: 860-828-8879, www.liberty-ind.com

Palmer Wahl, 234 Old Weaverville Rd., Asheville, NC 28804-1228, Phone: 828-658-3131, 800-421-2853 (toll free), Fax: 828-658-0728, www.palmerwahl.com

Pyrodynamics, 3522 Ashford Dunwoody Rd., Suite 423, Atlanta, GA 30319-2002, Phone: 657-444-7976, www.pyrodynamics.com

ShockWatch, 5501 LBJ Freeway, Suite 350, Dallas, TX 75240, Phone: 214-630-9625, 800-527-9497 (toll free), Fax: 214-638-4512, www.shockwatch.com/

Watlow, 12001 Lackland Rd., Saint Louis, MO 63146-4001, Phone: 314-878-4600, 800-928-5692 (toll free), www.watlow.com

77

Fiber-Optic Thermometers

77.1 Introduction 77-1
77.2 Fiber-Optic Temperature Sensors 77-3
77.3 Fiber-Optic Point Temperature Measurement Systems 77-4
77.4 Distributed and Quasi-Distributed Optical Fiber Temperature Measurement Systems 77-7
77.5 Application for Optical Fiber Temperature Probes 77-10
Further Information 77-10
Further Reading 77-10
Partial List of Vendors and Suppliers 77-11

Brian Culshaw
University of Strathclyde

77.1 Introduction

Optical fiber sensing is a remarkably versatile approach to measurement. A fiber sensor guides light to and from a measurement zone where the light is modulated by the measurand of interest and returned along the same or a different optical fiber to a detector at which the optical signal is interpreted. The measurement zone can be intrinsic within the fiber that transports the optical signal or can be extrinsic to the optical waveguide. The versatility of the fiber sensing medium arises in part because of the range of optical parameters that can be modulated and in part because of the diversity of physical phenomena that involve environmentally sensitive interactions with light.

For example, highly coherent light from a laser source can be introduced into a fiber and its phase modulated by a parameter of interest. The resulting phase changes can then be detected interferometrically. The phase change is simply a modification to the optical path length within the fiber and can be modulated by shifts in temperature, strain, external pressure field, or inertial rotation. A well-designed interferometer can detect 10^{-7} rad—equivalent to 10^{-14} m!

The laser light could also be Doppler shifted through reflection from a moving object. Its state of polarization can be changed. Its throughput intensity can be modified, or the light can be used to stimulate some secondary emissions, which in turn can be monitored to produce the relevant optical signal. If the light is incoherent, then its wavelength distribution (color) can be modified, in addition, of course, to the possibilities for polarization changes and intensity changes.

The physical phenomena capable of imposing this modulation are again many and varied. They include, for example, periodically bending an optical fiber to introduce a localized loss that depends on the sharpness of the bend (usually referred to as microbend loss); changing the relative refractive indices of the core and the cladding of the optical fiber and thereby changing the guiding properties and again introducing a loss; modifying an optical phase delay by introducing a change in refractive index or a change in physical length; examining changes in birefringence introduced through modifications to physical stress and/or temperature; and using external indicators to color modulate a broadband source and relate the color distribution to temperature, chemical activity, etc. These are all linear effects where

the input optical frequency is the same as the output optical frequency (regarding Doppler shift as a rate of change of phase of an optical carrier) and where, for a given system setting, the output at all frequencies is directly proportional to the input.

Nonlinear effects are also widely exploited. Of these, the most important are fluorescence, observed usually in fluorophores external to the optical fiber, and Raman and Brillouin scattering, usually observed within the fiber itself. In all these phenomena, the light is absorbed within a material and reemitted as a different optical wavelength from the one that was observed. The difference in optical wavelengths depends on the material and usually on strain and temperature fields to which the material is subjected. These major features of optical fiber sensors are encapsulated in Figure 77.1.

Optical fiber sensors have an additional feature that is unique to the medium—namely, the abilities for intrinsic networking in either distributed, quasi-distributed/multiplexed, or discrete (point) configurations. The essential features of these architectures are sketched in Figure 77.2. For intrinsic sensors,

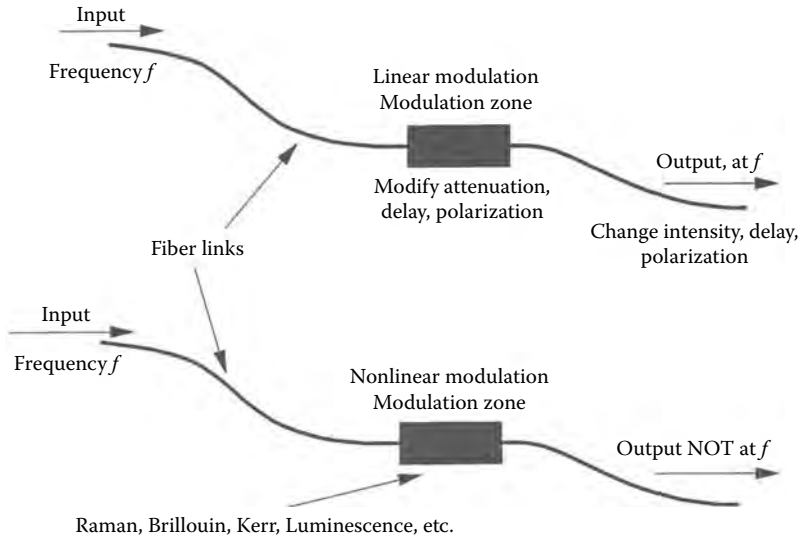


FIGURE 77.1 Linear and nonlinear optical processes for measurement using optical fiber sensors.

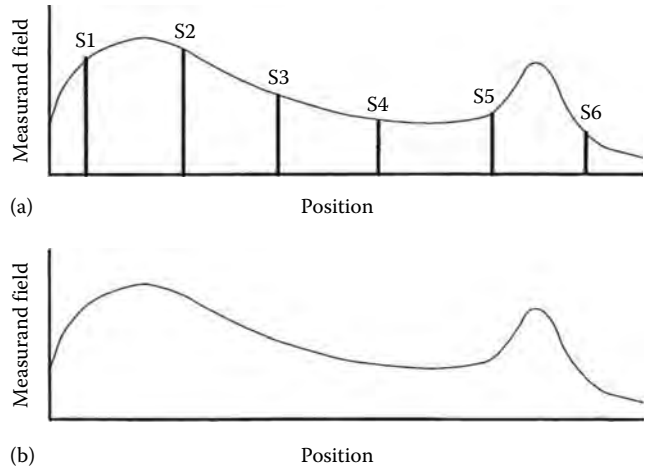


FIGURE 77.2 Sensor system outputs for (a) point array and (b) distributed sensor systems.

the fiber responds to the measurand throughout its length, and the output in transmission is an integral of this linear response. However, using an interrogation scheme in reflection that incorporates a delay proportional to distance along the fiber enables the system to retrieve the measurand as a function of position. These are distributed sensors. The quasi-distributed architecture examines separately identified individual (usually adjacent) lengths of fiber and extracts the integral of the measurand along each of these individual lengths. Distributed and quasi-distributed sensors effectively convolve the measurand field along the interrogation fiber with a window determined by either the time resolution of the interrogating electronics (distributed architectures) or the defined lengths of the individual fiber sections (quasi-distributed systems).

Point and multiplexed systems address the measurement as essentially a point sample located at a specific distance along the interrogating fiber. All these architectures can be realized in all optical fiber form and have been demonstrated to address a very wide range of measurements, often within a single network. The availability of distributed sensing is unique to optical fiber technology, as indeed are optical fiber-addressed passive arrays.

In summary, the optical fiber approach to measurement has the demonstrated capability to address a wide range of physical, chemical, and biological parameters. It must take its place alongside the other competing technologies against which its merits must be assessed. The principal benefits of using fiber optics include:

- Immunity to electromagnetic interference within the sensor system and within the optical feed and return leads
- The capacity for intrinsic distributed measurements
- Chemical passivity within the sensor system itself and inherent immunity to corrosion
- Small size, providing a physically, chemically, and electrically noninvasive measurement system
- Mechanical ruggedness and flexibility, optical fibers are exceptionally strong and elastic—they can withstand strains of several percent
- High temperature capability—silica melts at over 1500 °C

There remain cost and user acceptability deterrents within the exploitation of optical fiber sensor technology. Consequently, the majority of field experience in optical fiber sensors is targeted at addressing the specialized problems where these aforementioned benefits are paramount. Many of these lie in the area of temperature measurement.

77.2 Fiber-Optic Temperature Sensors

The important phenomena that have been exploited in the optical techniques for temperature measurement include:

- Collection and detection of blackbody radiation
- Changes in refractive index of external media with temperature
- Changes in fluorescence spectra and/or fluorescence rise times with temperature
- Changes in Raman or Brillouin scatter with temperature
- Phase transitions in carefully selected materials imposing mechanical modulation on optical fiber transmission properties
- Changes within an optical path length with temperature, either within the fiber or an external interferometer element

Within these phenomena, Brillouin and Raman scatter and mechanical phase transitions have been primarily used in distributed measurement systems. Some distributed measurement/quasi-distributed measurement systems based on modulated to phase delay have also been evaluated, although they have yet to reach commercial reality. The remaining phenomena are almost exclusively used in point sensor systems.

77.3 Fiber-Optic Point Temperature Measurement Systems

One of the first commercial optical fiber sensors was a fluorescence-based temperature probe introduced in the early 1980s by the Luxtron Corporation of Mountain View, CA. Successors to these early sensors are still commercially available and are a very effective, but expensive, approach to solving specific measurement problems. These include monitoring temperature profiles in both domestic and industrial microwave ovens, examining temperatures in power transformer oils, motor/generator windings, and similar areas where (primarily) the issue is the operation of a reasonably precise temperature probe within very high electromagnetic fields. In such circumstances, a metallic probe either distorts the electromagnetic field significantly (e.g., in microwave ovens) or is subjected to very high levels of interference, producing spurious readings. Other applications sectors exploit the small size or chemical passivity of the device, including operation within corrosive solvents or examination of extremely localized phenomena such as laser heating or in determining the selectivity of radiation and diathermy treatments.

The principles of the probe are quite simple and are shown in Figure 77.3. The rare earth phosphor is excited by an ultraviolet light source (which limits the length of the silica-based feed fiber to a few tens of meters), and the return spectrum is divided into "red" and "green" components, the intensity ratios of which are a simple single-valued function of phosphor temperature. For precision measurement, the detectors and feed fiber require calibration, and, especially for the detectors, the calibration is a function of ambient temperature. However, this can be resolved through curve fitting and interrogation of a thermal reference. The instrument, which has now gone through several generations to improve upon the basic concept, is capable of accuracies of about ±0.1 °C within subsecond integration times over a temperature range extending from approximately −50 °C to +200 °C. Since its introduction, this particular probe has accumulated extensive field experience in a wide variety of applications and remains among the most widely exploited fiber-optic sensor concepts.

A number of temperature probes based on fluorescence decay time measurements have also been demonstrated. The level of commercial activity exploiting these concepts has, to date, been very modest, partly because the accurate measurement of decay times can be problematic.

Measuring the temperature response of dyes and other thermally sensitive color-selective materials can afford a very simple approach to temperature measurements. Among the most successful of these has been the temperature probe examining the band edge of gallium arsenide introduced by ASEA (now ABB),

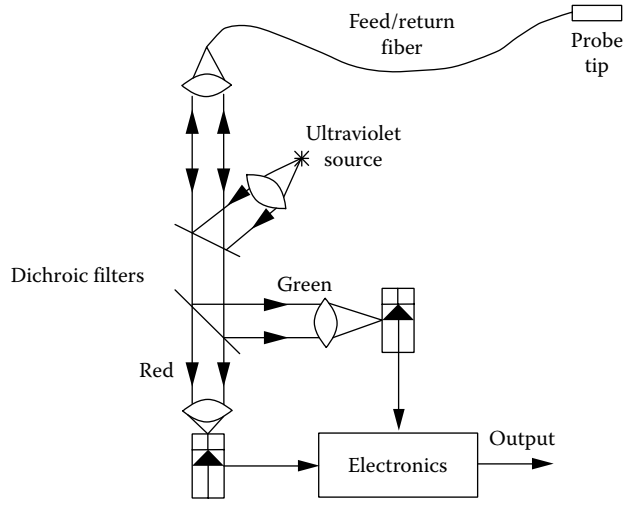


FIGURE 77.3 Optical fiber fluorescent thermometer.

again in the early 1980s and now transferred to Takaoka. The band edge can either be monitored through examining the spectra of induced fluorescence or through interrogating the absorption characteristics of the material when subjected to a constant spectrum excitation. The accuracy and temperature range of this probe are comparable with those of the Luxtron system, and this particular version of the band-edge probe has the additional benefit of operating primarily in the near-infrared range of the spectrum, thereby accessing the best transmission characteristics of the optical fiber medium. The probe was originally conceived to address ASEA's internal needs in monitoring electrical power system components. Similar band edge probes have also been demonstrated based on absorption edge detection in materials such as ruby.

Refractometry and interferometry are potential extremely sensitive thermal probes. Several have been demonstrated, some of which achieve microkelvin resolution. Interferometric detection or exploitation of sensitive mode coupling phenomena is the source of this very high sensitivity, although rarely is such high sensitivity required in practice. The relatively simple Fabry–Perot probe shown in Figure 77.4 has been introduced commercially with simplified spectral analysis and a semiconductor source, although as yet its market penetration has been relatively modest.

Optical pyrometry is a well-established approach to measuring temperatures in the hundreds to thousands of degrees centigrade. The disappearing filament pyrometer has been used in this fashion for over half a century. The optical fiber equivalent has also found a few niche applications. The general form of such a sensor is to place the blackbody at the end of the fiber and place it with the fiber into the hot zone. The consequent radiation within the transmission spectrum of the fiber is then monitored using a semiconductor photodetector that can be based on III–V materials or silicon. The received radiation is then primarily within the red and near infrared from about 600 nm to, depending on the detector, 1.8 μm. Blackbodies radiate significantly in this range at temperatures in the hundreds of degrees centigrade and above. The most significant success of this approach has been in the fabrication of the reference standard temperature probe at NIST for temperatures above 1200 °C. This uses a sapphire rather than silica collection fiber because of its superior optical and thermal properties within the temperature and the wavelength ranges of interest. It defines these high temperatures with subdegree precision.

Optical fibers also have the capacity to make unique nanoprobes—the opposite end of the scale by orders often from the distributed sensors discussed as follows. These (Figure 77.5) are tapered optical fibers with the end reduced in diameter to tens of nanometers. The tapered region is coated with a

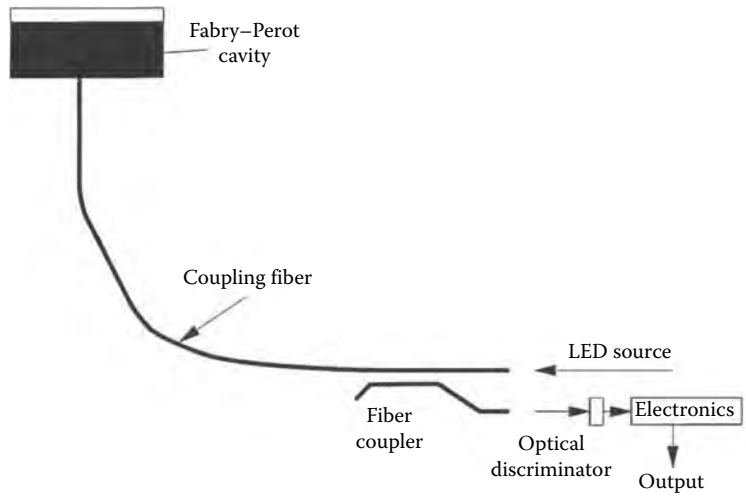


FIGURE 77.4 Optical fiber thermometry using short temperature-sensitive Fabry–Perot cavity.

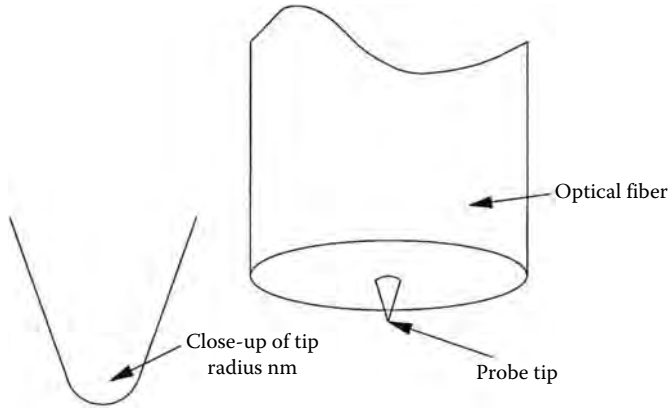


FIGURE 77.5 Probe for photon tunneling microscope and nanooptrode.

metal, often aluminum or silver, to confine the optical field. This produces an intense spot of light at the fiber tip, which irradiates an area in nanometers dimensions. The tip can be coated with the dye or the fluorescent thermally sensitive material and used to monitor temperature over extremely small areas. This enables thermal profiles within cellular dimensions to be assessed. In other formats, the same probe can also be used to address chemical activity and chemical composition.

Fiber-optic temperature sensing can be realized using a wide variety of techniques primarily, but not exclusively, based on the variation of optical properties of materials with temperature. An example of the exceptions is the optically excited vibrating element probe shown in Figure 77.6. This probe has been primarily used for pressure measurement and is now available commercially for pressure assessment downhole in oil wells. It can also be configured to exhibit extremely high temperature sensitivity with accuracies and resolutions in the millikelvin region. It uses the beneficial features of mechanical resonators and the consequential frequency readout in parallel with optomechanical excitation and direct optical interrogation to produce a probe that can be reliably exploited over interrogation lengths of tens of kilometers.

Fiber-optic point sensors for temperature measurement are now a relatively mature technology. Most of the devices mentioned previously were first introduced and characterized 10 or more years ago and have since been refined to address specific applications sectors. They remain expensive, especially when compared to the ubiquitous thermocouple, but their unique capability for noninvasive electrically passive interference immune measurement gives them a very specific market address that cannot be accessed using alternative technologies. Within these market areas, the probes have been extremely successful.

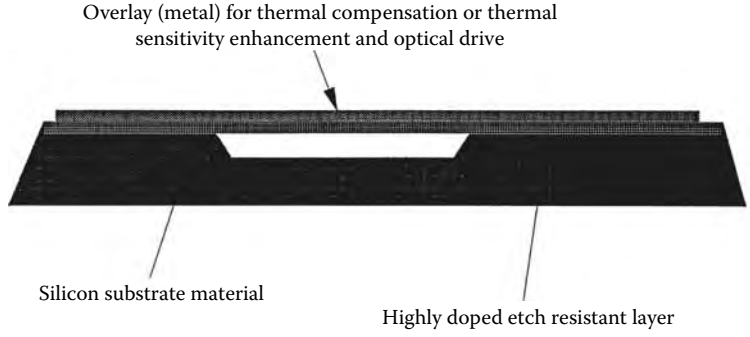


FIGURE 77.6 Longitudinal section of silicon optically excited microresonator.

77.4 Distributed and Quasi-Distributed Optical Fiber Temperature Measurement Systems

These systems all exploit the unique capability for optical fibers to measure and resolve environmental parameters as a function of position along the fiber length. This generic technology is unique to optical fiber systems, and, while there are a few commercial distributed temperature sensor systems available, the research in this sector continues.

The stimulated Raman scatter (SRS) distributed temperature probe is the most well established of these and, in common with many of the point sensors, was originally introduced commercially in the late 1980s. The principle (Figure 77.7) is quite simple. Within the Raman backscatter process (and also within the spontaneous Brillouin backscatter process), the amplitudes of the Stokes and anti-Stokes lines are related to the energy gap between these lines by a simple $\exp(-\Delta E/kT)$ relationship. Therefore, measuring this ratio immediately produces the temperature. Furthermore, this ratio is uniquely related to temperature and cannot be interfered with by the influence of other external parameters. The system block diagram is shown in Figure 77.8. The currently available performance from such systems enables resolutions of around 1 K in integration times of the order of 1 min, with resolution lengths of one to a few meters over total interrogation lengths of kilometers. The interrogation can extend to tens of kilometers if either the interrogation times are increased or the temperature and/or spatial resolutions are relaxed.

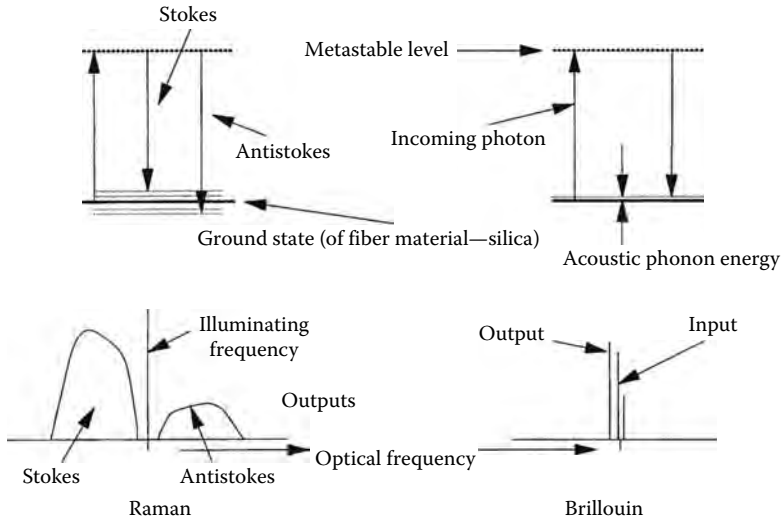


FIGURE 77.7 Thermally sensitive nonlinear scattering processes in optical fibers.

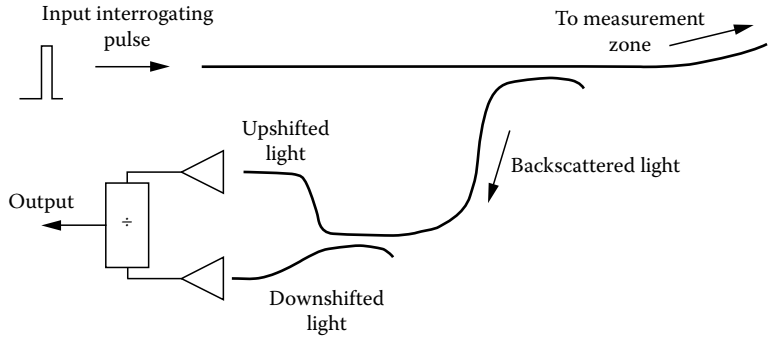


FIGURE 77.8 Raman distributed temperature probe: basic schematic.

The system is available from both European and Japanese manufacturers. The applications are very specific, as indeed they must be to accommodate an instrument price that is typically in tens of thousands of dollars. The instruments have been used in a variety of highly specific areas, ranging from monitoring temperature profiles in long process ovens to observing the thermal characteristics within large volumes of concrete during the curing process.

Distributed temperature alarms triggering on and locating the presence of either hot or cold spots along the fiber can be realized at significantly lower costs and have been modestly successful as commercial systems. The first of these—and probably the simplest—was originally conceived in the 1970s. This uses a simple step index fiber in which the refractive index of the core material has a different temperature coefficient than that of the cladding material. The temperature coefficient is designed such that at a particular threshold temperature, the two indices become equal and thereafter that of the cladding exceeds that of the core. Within this section, light is no longer guided. Simple intensity transmittance measurement is then very sensitive to the occurrence of this threshold temperature at a particular point along the fiber. If used with an optical time domain reflectometer, the position at which this first event occurs can be located. This system is now in use as a temperature alarm on liquefied natural gas storage tanks. Here, the core and cladding indices for a plastic-clad silica fiber cross at a temperature in the region of 50 °C. Such temperatures can only be achieved when a leak occurs. Further, the system has the obvious benefit of intrinsic safety and total compatibility with use within potentially explosive atmospheres.

A heat—as opposed to cold—alarm system that has also been introduced commercially is shown in Figure 77.9. In this system, the central tube is filled with a wax that expands by typically 20% when passing through its melting point. This expansion in turn forces the optical fiber against the helical binding, introducing a periodic microbend and thereby increasing the local loss within the fiber. The wax transition temperatures can be defined over a relatively wide range (say 30 °C–70 °C) and a low-cost OTDR system enables location of the hot spot to within a few meters. This system presents a very cost-effective overheat or fire alarm when such systems are required and are in intrinsically safe areas or in regions of very high electromagnetic interference. Again, it is the unique properties of the optical fiber sensing medium—especially intrinsic safety and electromagnetic immunity—that provide this system with its market address.

Brillouin scatter is very similar in character to Raman scatter except that in Brillouin scatter, the interaction is with an acoustic phonon rather than an optical phonon. The frequency shifts are then correspondingly significantly smaller (typically 10–15 GHz). Additionally, in Brillouin scatter, the frequency of the scattered light depends on the acoustic velocities in the medium within which the light is scattering. Consequently, the Brillouin scatter spectrum is a function of both temperature (through variations of

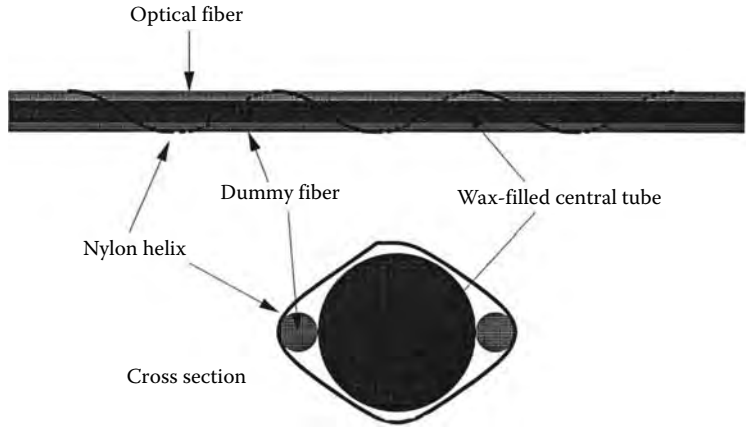


FIGURE 77.9 Fiber-optic distributed heat (fire) alarm.

modulus and density) and strain applied to the optical fiber. Usually this is exploited as a complex but very effective means for measuring strain distributions along an optical fiber. The Brillouin scatter cross section is much higher than that for Raman scatter so that distributed strain distributions can be measured over distances well in excess of 100 km. This measurement is particularly effective when exploiting stimulated Brillouin scatter that guides the scattered light back toward the source. However, since the apparent value of strain depends on temperature through the variation of acoustic velocity with temperature, temperature correction is required in most practical situations, and, in principle, this correction can be implemented by measuring spontaneous Brillouin scatter and specifically the intensity ratio of this in the upper and lower sidebands. This particular correction technique is currently in its infancy, and accuracies in the degree Kelvin range are the current state of the art. The difficulty in temperature measurement is that the energy gap between the two sidebands is very small so that the ratio of the amplitudes is close to unity but must be measured very accurately in order to invert the exponential.

The optical Kerr effect manifests itself as an intensity-dependent refractive index. Consequently, this nonlinearity gives rise to either second harmonic generation or sum and difference frequencies. It has been investigated for distributed temperature sensing using pump/probe configurations and birefringent fiber from which the beat length is a function of temperature and strain. This beat length in turn determines the frequency offset through phase-matching conditions of the mixed pump and probe signal (Figure 77.10). The overall situation is conceptually similar—this offset frequency depends on temperature and strain although in principle, dual measurements and adequate calibration can retrieve both, or alternatively the probe can measure a temperature field in the absence of strain. Again, the actual experimental results that have been achieved remain in the laboratory, and the accuracies and resolutions are modest.

In quasi-distributed sensing and point multiplexed systems, temperature probes have, as yet, been but little exploited. There are many variations on the basic theme of a marked optical fiber within which the optical interrogation system measures the optical path length between the marks. These marks can be introduced using partially reflective gratings, mode coupling Bragg gratings, partially reflective splices or connectors, low reflectivity directional couplers, or a multitude of other arrangements. Similarly, the optical delay between the markings can be measured directly as an optical or subcarrier phase or indirectly through monitoring dispersion between adjacent modes typically in low-mode or birefringent fibers. Yet again, the different delays depend on both temperature and strain so that for temperature measurement, a strain-free mounting is ideal. The context within which most, if not all, of this class of system has been evaluated is that of smart structures, and here the technique does offer some promise as a means for deconvolving strain, mechanical, and thermal effects and assessing structural integrity. It could also function as a temperature measurement probe, but to date has been minimally addressed in this application.

In multiplexed systems, the current fashion, again primarily for combined strain and temperature measurement, is to incorporate arrays of Bragg gratings used here as wavelength filters rather than

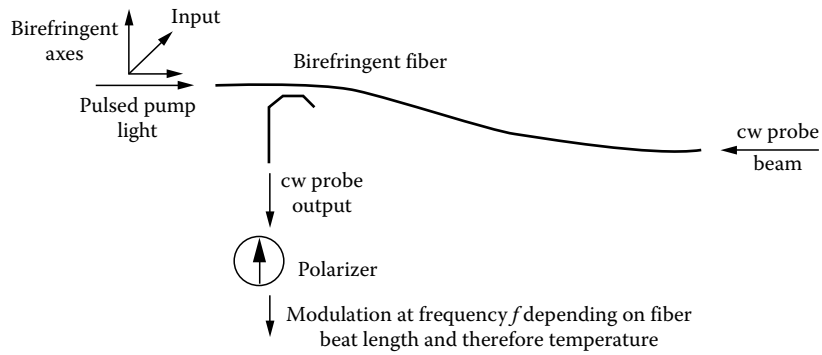


FIGURE 77.10 Distributed Kerr effect probe for temperature or strain field measurements.

as broadband reflectors. In this configuration, the Bragg grating presents a combined temperature/strain field at its location encoded within the reflection wavelength. Multiple addressing schemes can deconvolve temperature and strain sensitivities. There have also been a few demonstrations of discrete temperature-sensitive elements inserted at strategic points along an optical fiber. Of these the use of band-edge shifting in ruby crystals interrogated using a pulsed source observed in reflection has probably been the most successful. In this arrangement, the reflectors are replaced by the crystals and sample the temperature field at these points. The receiver then determines the return to spectrum as a function of time.

77.5 Application for Optical Fiber Temperature Probes

Instrumentation is a very application-specific discipline, and, in particular for sensors, a particular technology is usually only relevant in a limited number of application sectors. As the technology becomes more and more specialized and expensive, these applications niches become much more tightly defined.

For optical fiber sensors and their use in temperature probes, the more conventional approaches (thermocouples, thermistors, platinum resistance thermometers, etc.) are always easier and simpler. The fiber-optic technology must exploit one or more of electromagnetic immunity, small size, noninvasiveness, chemical immunity, or the capacity for distributed measurement.

Point optical sensors are therefore primarily used as measurement probes in regions of very high electromagnetic fields, in zone zero intrinsically safe areas, and as in vivo medical probes.

The distributed capability of fiber sensors is especially relevant in structural monitoring and in other specialized areas such as measuring the temperature distribution along underground power lines, tunnels, or similar structures or in experimental circumstances such as the measurement of curing processes in large volumes of concrete.

Fiber optics then is exactly similar to all other sensing techniques—it is an inappropriate temperature probe for the majority of applications; but for those for which it is appropriate, it offers a unique and effective solution to frequently otherwise intractable measurement challenges. As a technology evolves and becomes both more widely accepted and readily available, the applications address will no doubt expand.

Further Information

Additional information on optical fiber temperature measurements can be obtained from the following:

Culshaw, B., *Optical Fiber Sensing and Signal Processing,* Stevenage, U.K.: IEE/Peter Perigrinus, 1983.

Culshaw, B. and J. P. Dakin, *Optical Fiber Sensors, Vol. I-IV,* Boston, MA: Artech House, Vol. 1, 1988; Vol. II, 1989; Vols. III and IV, 1997.

Grattan, K. T. V. and B. T. Meggitt, *Optical Fiber Sensor Technology*, Kluwer Academic Publisher, Dordrecht, the Netherlands, 2000.

Udd, E. and W. B. Spillman Jr., *Fiber Optic Sensors for Engineers and Scientists.* 2nd edn., John Wiley and Sons, New York, 2011.

Udd, E., *Fiber Optic Smart Structures*, Wiley Interscience Publication, New York, 1995.

Further Reading

Dills, R.R., Optical fiber thermometers, Online, available at, http://www.devileye.net/catalog/position_sensitive_friction_damper/optical_fiber_thermometer.html (accessed on August 2, 2012).

Introduction to fiber optics temperature measurement, Omega Engineering, http://www.omega.com/prodinfo/FiberOpticTemp.html (accessed on August 3, 2012).

Yoder, J., From glass thermometer to fiber optics, Flow Control, http://www.flowcontrolnetwork.com/measurement/temperature/article/from-the-glass-thermometer-to-fiber-optics (accessed on August 2, 2012).

Partial List of Vendors and Suppliers

Grainger Industrial Supply, 100 Grainger Pkwy., Lake Forest, IL 60045, Phone: 847-535-1000, 888-361-8649 (toll free), Fax: 847-535-9123, www.grainger.com/thomasnet

LumaSense Technologies, Inc., 3301 Leonard Ct., Santa Clara, CA 95054, Phone: 408-727-1600, Fax: 408-727-1677, www.lumasenseinc.com/EN/

Neoptic Inc., 1415 Frank-Carrel Ste. 220, Quebec, QC G1N4N7, Phone: 418-687-2500, Fax: 418-948-1173, www.neoptix.com

PICS INC, 81 Lancaster Ave., Suite 202, Malvern, PA 19355, Phone: 610-644-7701, Fax: 610-644-9297, www.picscontrols.com

Process Sensors Corp., 113 Cedar St., Milford, MA 01757, Phone: 508-473-9901, Fax: 508-473-0715, www.processsensors.com

Therm,-X, 1837 Whipple Rd., Hayward, CA 94544, Phone: 510-441-7566, Fax: 510-441-2414, www.therm-x.com



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

78

Thermal Imaging

78.1 Essential Components **78**-1
78.2 Thermal Imaging Wave Bands **78**-2
78.3 Emission from Source **78**-4
78.4 Atmospheric Transmission **78**-5
78.5 Detectors **78**-5
Photon Detectors • Thermal Detectors • Detector Performance Measures • Detector Cooling
78.6 Electronics **78**-9
78.7 Optics and Scanning **78**-10
78.8 Temperature References **78**-13
78.9 Imager Performance **78**-14
SNR and NETD • Minimum Resolvable Temperature Difference
78.10 Available Imagers **78**-15
78.11 Performance Trade-Offs **78**-16
78.12 Future Trends in Thermal Imaging **78**-18
References **78**-18
Further Information **78**-19

Herbert M. Runciman
Pilkington Optronics

All objects at temperatures above absolute zero emit electromagnetic radiation. Radiation thermometry makes use of this fact to estimate the temperatures of objects by measuring the radiated energy from selected regions. Thermal imaging takes the process one stage further and uses the emitted radiation to generate a picture of the object and its surroundings, usually on a TV display or computer monitor, in such a way that the desired temperature information is easily interpreted by the user.

Thermal imagers require no form of illumination to operate, and the military significance of this, together with their ability to penetrate most forms of smoke, has been largely responsible for driving thermal imager development. Although thermal imagers intended for military or security applications can be used for temperature measurement, they are not optimized for this purpose since the aim is to detect, recognize, or identify targets at long ranges by their shape; thus, resolution and sensitivity are favored over radiometric accuracy.

78.1 Essential Components

All thermal imagers must have a detector or array of detectors sensitive to radiation in the required wave band and optics to form an image of the object on the detector. In modern thermal imagers, the detector array might have a sufficient number of sensitive elements to cover the focal plane completely (a "staring array"), in the same way as a CCD television camera. Some of the most recent staring arrays can give good performance without cooling. In other imagers, the detector might take the form of a single row or column of elements, in which case a scanning mechanism is required to sweep the image across the detector array. If a single-element detector, or a very small detector array, is used, a means of

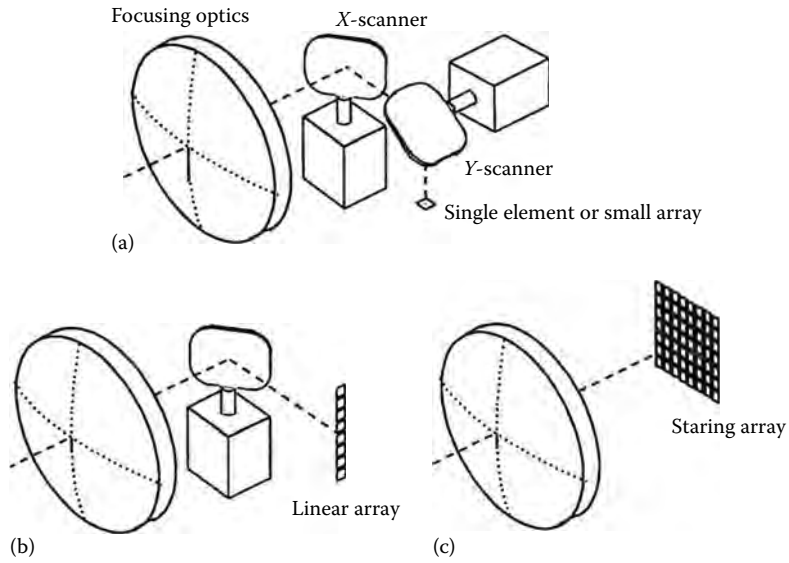


FIGURE 78.1 Thermal imaging options: (a) 2-D scanning for small detector array or single element, (b) 1-D scanning with linear detector array, and (c) staring array without scanning.

providing a 2-D scan is required. (Figure 78.1 shows these options schematically.) For scanning imagers, it is necessary to cool the detectors (usually to about 80–120 K) to achieve adequate performance.

Although in principle it would be possible to deduce target temperature from the absolute value of the detector signal, it is necessary in practice to estimate temperature by comparison with one or more reference bodies of known temperature. The temperature references are usually internal to the equipment and accessed by mechanical movement of the reference (which may take the form of a rotating chopper) or by deflecting the optical path using a mirror.

78.2 Thermal Imaging Wave Bands

The optimum wave band for thermal imaging is determined partly by the wavelength distribution of the emitted radiation, partly by the transmission of the atmosphere, and partly by the chosen detector technology.

The power radiated from a given area of an object depends only on its temperature and the nature of its surface. If the surface absorbs radiation of all wavelengths completely, it is referred to as a "blackbody." It then also emits the maximum amount possible, which can be calculated using the Planck equation (given later). Figure 78.2a shows the way in which blackbody emission varies with wavelength for several temperatures. It will be seen that for objects near normal ambient temperature, maximum output occurs at a wavelength of about 10 μm or about 20 times the wavelength of visible light. At wavelengths below about 3 μm, there is generally insufficient energy emitted to allow thermal imaging of room-temperature objects. The emissivity at any wavelength is defined as the ratio of the energy emitted at that wavelength to the energy that would be emitted by a blackbody at the same wavelength.

It is important that the atmosphere should have sufficient transparency to permit the target to be observed. There are two important "atmospheric windows"—one between 3 and 5 μm (with a notch at 4.2 μm due to carbon dioxide absorption) and one between 7.5 and 14 μm. These are commonly referred to as the medium-wave infrared (MWIR) and long-wave infrared (LWIR) windows, respectively. For thermal measurement over short ranges in the laboratory, it is possible to work outside these bands,

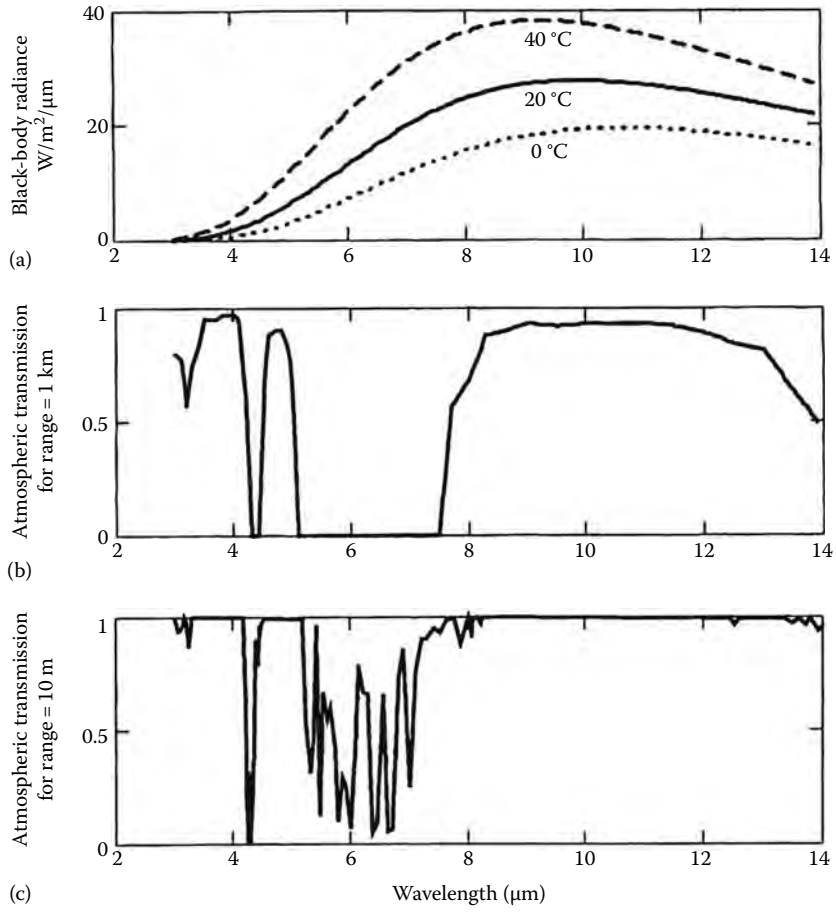


FIGURE 78.2 Factors determining thermal imaging wave bands. Imager must operate in regions where (a) radiance is sufficiently high and (b, c) atmospheric transmission is good.

but most instruments are optimized for either MWIR or LWIR. Typical transmissions through 1 km and 10 m of a clear U.S. Standard Atmosphere are shown in Figure 78.2b and c.

Emissivity for most naturally occurring objects and organic paints is high (>0.8) in the LWIR, but is lower and more variable in the MWIR. Metallic surfaces have low emissivity in both bands. Solar radiation in the MWIR is significant and can cause errors in measurements made outdoors. These considerations favor the LWIR for quantitative imaging, but the band chosen can also be influenced by detector technology, the latter frequently is being determined by its cost. Scanning imagers can be used in either band but are more sensitive for a given detector architecture in the LWIR. Cooled staring arrays give similar sensitivity in either band but are currently more readily available in the MWIR. Uncooled staring arrays work well only in the LWIR band.

For temperature measurement, the electronics can be used to encode signal level as false color, a color scale derived from the thermal references being injected into the display to allow the user to identify the temperature of the object under examination. For general surveillance, a conventional gray-scale image is usually preferred. Imagers for thermography can also include emissivity compensation. If accurate results are required for an object of low emissivity, it is important to ensure that the temperature of anything that might be reflected by the object is known and that the emissivity is accurately known. If the object is accessible, another object placed beside it with the same surface characteristics but known temperature can be used for calibration.

78.3 Emission from Source

The spectral radiance $W(\lambda,T)$ of a blackbody at temperature T and wavelength λ is given by the Planck equation [1]. For temperature differences between the target and the reference of a few degrees, it is frequently sufficiently accurate to assume a linear dependence of radiance on temperature difference, making the temperature derivative of the blackbody equation, $dW(\lambda,T)/dT$, more relevant. In the case of photon detectors, the detector output is proportional to the photon flux, which can be derived from the radiance using the fact that photon energy $E(\lambda) = hc/\lambda$, where h is the Planck constant. The total photon flux $N(\lambda)$ and its derivative with respect to temperature are thus relevant in this case. The equations are as follows:

$$W(\lambda,T) = \frac{c_1}{\lambda^5\left(e^{(c_2/\lambda T)}-1\right)}\ \mathrm{W\,m^{-2}\,\mu^{-1}} \tag{78.1}$$

$$N(\lambda,T) = \frac{c_3}{\lambda^4\left(e^{(c_2/\lambda T)}-1\right)}\ \mathrm{photons\ s^{-1}\,m^{-2}\,\mu m^{-1}} \tag{78.2}$$

$$\frac{dW(\lambda,T)}{dT} = \frac{c_1c_2e^{(c_2/\lambda T)}}{\lambda^6T^2\left(e^{(c_2/\lambda T)}-1\right)},\ \mathrm{W\,m^{-2}\,\mu m^{-1}\,K^{-1}} \tag{78.3}$$

$$\frac{dN(\lambda,T)}{dT} = \frac{c_3c_2e^{(c_2/\lambda T)}}{\lambda^5T^2\left(e^{(c_2/\lambda T)}-1\right)}\ \mathrm{photons\ s^{-1}\,m^{-2}\,\mu m^{-1}\,K^{-1}} \tag{78.4}$$

where

numerical values of the constants are

$c_1 = 3.742 \times 10^8$

$c_2 = 1.439 \times 10^4$

$c_3 = 1.884 \times 10^{27}$

The unit of wavelength is chosen for convenience to be the micrometer (μm).

The aforementioned values are for radiation into a hemisphere. The intensities (watts per steradian, photons per steradian, etc.) are obtained by dividing the aforementioned values by π. The actual radiances for real targets are obtained by multiplying by the spectral emissivity $\varepsilon(\lambda)$, but since target reflectivity $\rho(\lambda) = 1 - \varepsilon(\lambda)$, some caution is required. For example, a target at temperature T surrounded by a background of temperature T_b will appear to emit $W(\lambda,T)\varepsilon(\lambda) + W(\lambda,T_b)\rho(\lambda) = [W(\lambda,T) - W(\lambda,T_b)]\varepsilon(\lambda) + W(\lambda,T_b)$. Provided that the background surrounds the target and that the target is reasonably small, the background acts as an isothermal enclosure, which can be shown [2] to behave as an ideal blackbody (i.e., $\varepsilon(\lambda) = 1$). The differential spectral radiance against background $\Delta W(\lambda)$ is thus $[W(\lambda,T) - W(\lambda,T_b)]\varepsilon(\lambda)$, which for a small temperature difference ΔT is simply

$$\Delta W(\lambda) = \varepsilon(\lambda)\frac{dW(\lambda,T)}{dT}\Delta T \tag{78.5}$$

The spectral emissivity of a wide variety of natural and man-made objects is also given in [2].

A major difference between thermal imaging and visual imaging is the very low contrast. In the MWIR, the contrast calculated from Equation 78.1 due to 1 K at the target is about 4%, falling to about 2% in the LWIR.

78.4 Atmospheric Transmission

Provided that the absorption bands shown in Figure 78.2b and c are avoided, atmospheric transmission can frequently be ignored in the laboratory or industrial context. For longer ranges, an atmospheric transmission model must be used, or calibrating sources must be placed at the target range. The standard atmospheric transmission model is LOWTRAN [3], currently at version 7. The atmospheric transmission $T_a(\lambda)$ reduces the differential signal from the target proportionately, but has no effect on the background flux if the atmosphere is at background temperature. Atmospheric transmission in the LWIR is severely affected by high humidity, making the MWIR the band of choice for long-range operation in the Tropics. (Many gases and vapors such as methane or ammonia have very strong absorptions in the infrared, making thermal imaging a possible means of leak detection and location.)

78.5 Detectors

There are two main types of detector—photon (or quantum) and thermal. A more detailed discussion of photon detectors is given in this handbook, so only aspects unique to thermal imaging are discussed here.

78.5.1 Photon Detectors

In photon detectors, the response is caused by photons of radiation that generate free carriers in a semiconductor, which in turn increase the conductivity (for photoconductive detectors) or generate a potential difference across a junction (for photovoltaic detectors). Photovoltaic devices have the advantage of not requiring a bias current (important to reduce the heat load on the cooling system), and they have 40% lower noise because the electric field at the junction separates the carriers, thereby eliminating recombination noise. Whether or not the lower noise is achieved in practice depends on the readout electronics. The photon energy in the LWIR is only about l/20th of that of a photon in the visible region of the spectrum, so a semiconductor with a much smaller bandgap than silicon must be used. The most widely used material is a compound of mercury, cadmium, and tellurium (MCT or CMT) since not only is the quantum efficiency excellent (70% or more) but also the bandgap can be tuned to the desired wavelength (in either wave band) by altering the composition. Cooling of the detector to about 80 K is desirable for the LWIR, but about 120 K is acceptable for the MWIR. For the MWIR, indium antimonide (InSb) is also an excellent material, and since it is a true stoichiometric compound, it is easier to achieve good uniformity of response, but cooling to 80 K is required.

Detectors for use in scanning systems are frequently arranged so that several elements are scanned over the same part of the scene in rapid succession, the output of each element being delayed and added to the previous one to enhance the signal-to-noise ratio (SNR). This approach (Figure 78.3) is termed serial scanning or "time delay and integrate" (TDI) mode and gives a theoretical gain in the SNR equal to the square root of the number of elements in TDI. It is also possible to perform TDI in the detector material itself. In the signal processing in the element (SPRITE) detector, the sensitive element is an elongated strip of CMT. Photons incident on the device generate carriers that drift toward a readout electrode near one end. If the image is scanned along the detector at the same speed as the carrier drift, the signal builds up along the length of the device. The useful length is limited by carrier recombination, while diffusion of the carriers limits spatial resolution.

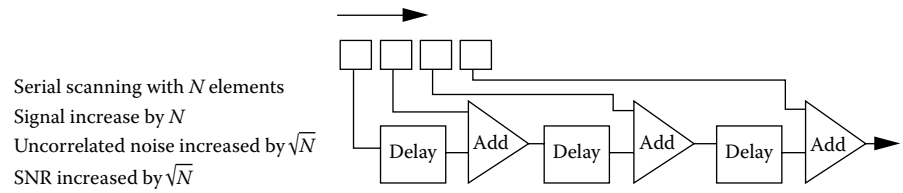


FIGURE 78.3 Use of serial scanning to enhance SNR. The delay times are chosen to match the speed at which the image is swept along the detector array. Serial scanning is usually combined with parallel scanning using a detector matrix.

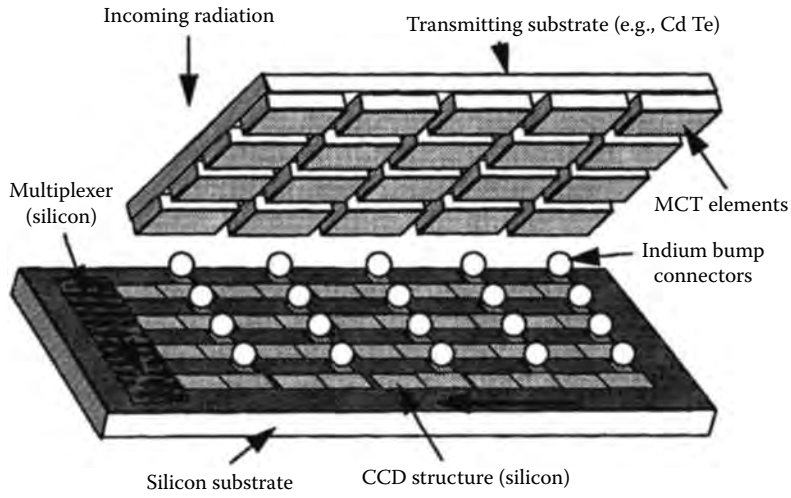


FIGURE 78.4 Typical hybrid detector construction. A typical element size is 30 μm. A large array of this type to match U.S. TV standard would have 640 × 480 elements.

Large arrays of photon detectors are generally of hybrid construction, the sensitive elements being bonded to a silicon CCD or CMOS addressing circuit using indium "bumps" (Figure 78.4). An exception is the Schottky barrier detector (e.g., platinum silicide), which can be manufactured by a monolithic process and thus tends to be lower in cost, but quantum efficiency is much lower, and operation is usually limited to the MWIR band. Detector arrays and readout architectures are discussed in depth in [2] (Vol. 3, p. 246–341) and [4].

The detector assembly is encapsulated in a Dewar as shown in Figure 78.5. In front of the detector is a "cold shield" that limits the acceptance angle of the radiation to match that of the optics.

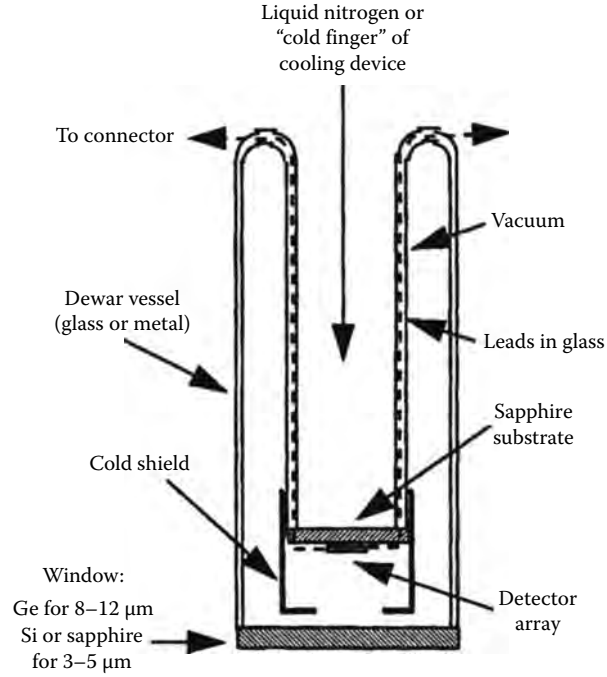


FIGURE 78.5 Construction of typical cooled detector. Cooling can be by liquid nitrogen, Joule–Thomson expansion of compressed gas, or cooling engine.

78.5.2 Thermal Detectors

Thermal detectors rely on the heating effect of the incoming radiation, the change in temperature causing a change in resistance, capacitance, or electrical polarization that might be detected electrically. They are generally fairly slow in response (several milliseconds) but have the advantage that cooling is not essential (although it can be of considerable benefit with some types). The detectivity of uncooled thermal detectors is typically 1/100 that of cooled photon detectors, so real-time imaging requires the use of staring arrays. The essentials of a pyroelectric array are shown in Figure 78.6. Incoming radiation is absorbed by the blackened electrode, and the heat generated is transferred to the pyroelectric layer, which comprises a dielectric material that has been polarized by means of a high electric field during manufacture. The change in electrical polarization with temperature gives the electric signal. One of the most important parameters is the thermal isolation of the sensitive elements, so some kind of insulating support structure is necessary, and for good performance, the device must be evacuated to prevent convection. In a variant of this approach, the dielectric bolometer, the rapid variation of the dielectric constant at temperatures near the Curie point, causes the capacitance of the sensitive element to change and hence the voltage for a constant charge. A detailed description of this approach is given in [5]. In both techniques, the detector responds only to change in temperature, so it is necessary to modulate the incoming radiation using a chopper. In the technique used initially by Wood [6] (now licensed to several manufacturers), the sensitive elements are vanadium dioxide coatings that undergo a large change in resistivity for a small temperature change. The elements are supported by silicon strips that are micromachined from the substrate and give excellent thermal insulation of the element. Changes in resistivity are read out by circuitry on the substrate, and no chopping is required, but the array must be maintained at a precise and uniform temperature.

78.5.3 Detector Performance Measures

The wavelength-dependent power responsivity of a detector $R(\lambda)$ is defined as the output potential or current that would result from 1 W of radiation at wavelength λ, assuming that linearity was maintained at such a high-flux level. The units are VW^{-1} or AW^{-1}. Photon responsivities in V photon^{-1} s^{-1} and A photon^{-1} s^{-1} are similarly defined.

A thermal detector has a power responsivity that is essentially independent of wavelength, the limits of response being determined by the transparency of the window and the absorption spectrum of the element or the material used to blacken it.

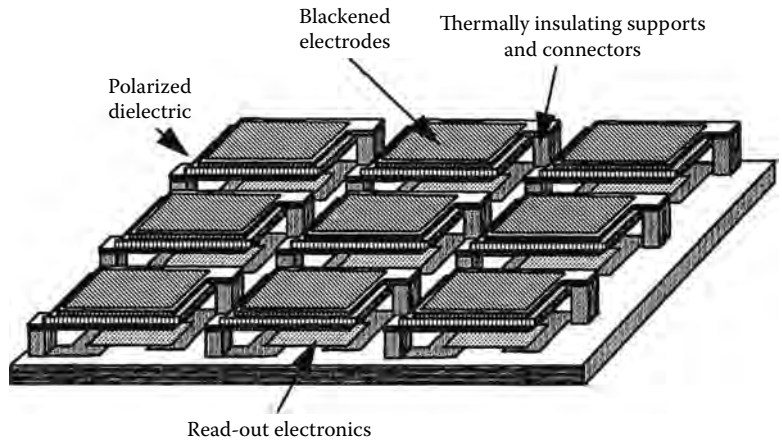


FIGURE 78.6 Essentials of an uncooled thermal detector array. The device is shown as pyroelectric or dielectric bolometer, but the essentials for a resistive bolometer are similar if the capacitors are replaced by resistors.

In an ideal photon detector, the quantum efficiency η (defined as the number of carriers generated per photon) would be constant at all wavelengths for which the photon energy is greater than the bandgap, that is, the photon responsivity is independent of wavelength up to the chosen cutoff wavelength. Since a given number of watts correspond to a number of photons proportional to the wavelength, the power responsivity (VW^{-1} or AW^{-1}) would increase linearly with wavelength until the cutoff. In practice, the cutoff is spread over about 0.5 μm, and the shortwave performance is modified by window transmission and antireflection coatings.

The sensitivity of a detector is limited by noise that may be due to the detector itself or due to the background radiation (as is discussed later). Noise-equivalent power (NEP) (λ) is defined as the power incident on the detector at wavelength λ, which gives a signal equal to the rms noise when the measurement is made with a 1 Hz bandwidth. The NEP depends also on the modulation frequency, the latter effect being large for thermal detectors but frequently negligible for quantum detectors. For many types of detector, the noise is proportional to the square root of the sensitive area A_d and the electrical bandwidth B so that $\sqrt{(A_dB)}/\text{NEP}$ is constant. A performance figure that is proportional to sensitivity can then be defined as specific detectivity $D^*(\lambda) = \sqrt{(A_dB)}/\text{NEP}(\lambda)$. For historical reasons, specific detectivity is usually given in units of $\text{cm}\sqrt{\text{Hz}}\text{W}^{-1}$, so it is important to remember to convert this to SI units or to measure detector area in square centimeters. Since noise is an electrical quantity particular to the detector under the conditions for which D^* is defined and is independent of wavelength, NEP(λ) is proportional to $1/R(\lambda)$, so if the value of detectivity D_p^* at the wavelength of peak responsivity R_p is known, for other wavelengths, $D^*(\lambda) = D_p^*(R(\lambda)/R_p)$. Sometimes, "blackbody $D^*(T)$" figures are quoted rather than D_p^*. If the blackbody temperature is T, the value of D_p^* is given by

$$D_p^* = \frac{D^*(T)R_p\int_0^\infty W(\lambda,T)\mathrm{d}\lambda}{\int_0^\infty W(\lambda,T)R(\lambda)\mathrm{d}\lambda} \tag{78.6}$$

The rms noise voltage V_n is simply the NEP multiplied by the responsivity. Since the ratio $D^*(\lambda)/R(\lambda)$ is independent of wavelength, one obtains

$$V_n = \sqrt{A_dB}\,\frac{R_p}{D_p^*} \tag{78.7}$$

The detector also affects the modulation transfer function (MTF) of the imager, defined as the ratio of the modulation depth of the signal due to a target with sinusoidally varying brightness of spatial frequency f cycles per milliradian to the modulation depth from a similar target at very low frequency. If there are no limitations due to time constant, the MTF of a detector is due to its instantaneous field of view (IFOV), which is given by IFOV = xlF where F is the focal length of the optics. For example, a lens of 500 mm focal length used with a 50 μm square detector would give an IFOV of 0.1 mrad. Then,

$$\text{MTF}_d = \frac{\sin(\pi \times f \times \text{IFOV})}{\pi \times f \times \text{IFOV}} \tag{78.8}$$

For staring arrays, frequencies above 1/2 cycle per element pitch (the Nyquist frequency) give aliasing, and it is undesirable to rely on performance above this frequency.

78.5.4 Detector Cooling

The method with the lowest cost is to use liquid nitrogen poured directly into the detector Dewar, and many detector manufacturers supply detectors in Dewars with sufficient capacity for many hours of use per filling. Outside the laboratory, this technique is seldom practical, and a cooling engine (usually based on the Stirling thermodynamic cycle) is commonly used. The cost of such an engine has now fallen to a level where it no longer dominates the cost of the instrument, and power consumption is only a few watts to give cooling to 80 K. Dimensions vary widely, but a typical low-power cooling engine is about 40 mm × 40 mm × 60 mm excluding the length of the cold finger that lies inside the detector Dewar. A disadvantage is the relatively slow cooldown time (several minutes typically), and where this is critical (as in military or security applications), Joule–Thomson cooling can be used. This method operates by expansion of air or nitrogen compressed to about 15 MPa through a nozzle and can give cooldown times of a few seconds. Thermoelectric cooling (using the Peltier effect) can be used for temperatures down to about 200 K. Some detectors (e.g., lead selenide) have been designed to operate at this temperature in the 3–5 μm band, but thermoelectric cooling is inadequate for most types of photon detectors. A survey of cooling methods and devices is given in [2] (Vol. 3, pp. 345–431).

78.6 Electronics

The electronics architecture depends on the type of detector and the application, but typically electronics are required to provide bias and clocking signals to the detector, to amplify the low-level signals from the detector, to equalize the responses of the outputs from different detector elements, to provide scan conversion to a form suitable for display, and to provide image processing suitable to the application.

In scanning imagers with a small number of detector elements, the output of each detector element can be amplified continuously, the amplifier outputs then being multiplexed to give the required display. The bandwidth (important for estimating sensitivity) must be at least sufficient to accommodate the data rate, but too wide a bandwidth gives excess high-frequency noise. An approach frequently used for the measurement of sensitivity (e.g., [1], p. 167) is to make the electronics response equivalent to that of a single-pole filter with the 3 dB point placed at a frequency corresponding to 1/(2 × dwell time), in which case the bandwidth is

$$B = \frac{\pi \times \mathrm{FOV_h} \times \mathrm{FOV_v} \times f_\mathrm{f}}{4E \times N \times \mathrm{IFOV}^2} \tag{78.9}$$

where

$\mathrm{FOV_h}$ and $\mathrm{FOV_v}$ is the horizontal and vertical fields of view, respectively
f_f is the field rate
E is the ratio of the active scan period to the total period
N is the number of parallel detector channels

Often high-frequency boost filters are used to compensate for optics and detector MTFs, in which case the noise bandwidth can be much increased. A commonly used criterion for the electronic filtering is to make the perceived noise independent of frequency up to the cutoff frequency of the detector.

To maintain good spatial resolution, it is desirable that the detector output be sampled at least twice during the dwell time so that the Nyquist frequency is not a major limitation [7]. Frequencies above the Nyquist frequency are changed by the sampling process into lower frequencies, a process known as aliasing. Thus, the fidelity of the image is affected so that noise that might be expected to be of sufficiently high frequency to be filtered out can appear within the passband. To avoid aliasing, a steep-cut filter is used to eliminate frequencies above the Nyquist frequency before sampling. In staring systems, the Nyquist sampling frequency is 1/2 cycle per element pitch, so a 256 × 256 element array would be limited

to a resolution of 128 cycles per line. This can be overcome by using "microscan" (or μscan), in which the image is collected over a number of fields with an image shift performed optically between each field, the commonest patterns being diagonal (low implementation cost) and 2×2 with four fields per frame.

For large detector arrays, charge is usually accumulated on a capacitor associated with each pixel for an integration time τ (which can often be controlled independently of the frame rate or dwell time to prevent saturation), the capacitor being discharged when the pixel is read out. The effective bandwidth of such an integrator is

$$B = \frac{1}{2\tau} \tag{78.10}$$

With state-of-the-art amplifiers, it is generally possible to make amplifier and readout noise less than the detector noise, the exceptions being systems of small aperture that are "photon-starved" and uncooled systems.

In multichannel systems, a crucial role of the electronics is to provide channel equalization. The importance of this is due to the very low contrast of the target against the background. If the full sensitivity of an imager with a typical noise-equivalent temperature difference (NETD) of 0.1 K is to be realized, the difference in response between adjacent detector channels must be less than 0.4% in the MWIR or 0.2% in the LWIR. In a real-time imager where the eye integrates over several frames, the requirement is 2–3 times more stringent, since nonuniformity, unlike SNR, is not improved by eye integration.

78.7 Optics and Scanning

The materials commonly used for visual optics are opaque in the thermal wave bands. In the LWIR, germanium is by far the most widely used material. It has a refractive index of 4, and chromatic dispersion is sufficiently low that it is frequently unnecessary to use a second material for color correction. These properties allow high performance to be achieved with few optical components, largely offsetting the relatively high cost of the material. In the MWIR, germanium has fairly high dispersion, but silicon/germanium doublets give highly achromatic performance. Zinc selenide and zinc sulfide are commonly used for color correction, some forms of the latter being transparent also in the visible band. The high refractive indices make antireflection coating essential to reduce surface losses—the transmission of a thin piece of uncoated germanium is only 40%, rising to better than 99% when coated. Front-surface mirrors coated with aluminum or gold perform well in the thermal bands, and optics based on parabolic mirrors can be used if the detector array is small. For most applications, it is necessary to seal the imager to prevent ingress of dust or moisture, in which case a window is required, making reflecting optics less attractive than it first appears, since the cost of a mirror and window can be a little less than that of a lens. Plastic materials in general have poor transmission, although a thin polythene or "cling-wrap" window might be acceptable for laboratory use.

For uncooled staring arrays, the sole function of the optics is to focus an image of the scene on the detector. Good performance in the 8–12 μm band can be achieved with a two-element Petzval lens [8] with aspheric surfaces (Figure 78.7a). With a cooled array, it is desirable that the cold shield inside the detector should form the aperture stop of the system, since any radiation from the interior of the instrument will add to the system noise and might give shading effects. If the field of view (FOV) is reasonably small, it is sometimes possible to use the same type of lens in the manner shown in Figure 78.7b, but it can be seen that the beam diameter that can be accepted is now much smaller than the lens diameter; thus, for long-range applications requiring large beam diameters, the lens becomes very costly. The solution is to use reimaging optics as shown in simplified form in Figure 78.7c. The relay stage not only reimages the scene on the detector but also images the cold shield on the objective lens so that the latter need be no larger than the input beam. The intermediate image can be useful to allow insertion of temperature references or microscanning.

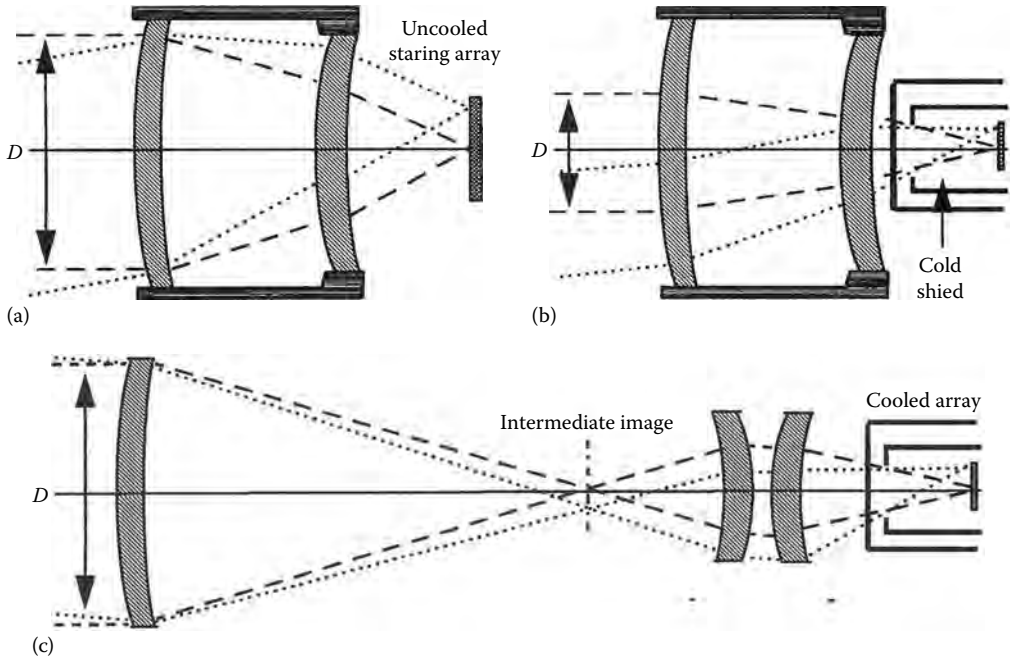


FIGURE 78.7 Optics for staring arrays. The same lens as is used for uncooled operation (a) can be used as in (b) for a cooled device, but reimaging optics (c) is needed for a cooled imager to use the full aperture. In practice, the optics in (c) would usually have more components than shown.

Scanning, when required, is most commonly performed by moving reflective surfaces since electro-optic and acousto-optic techniques either provide insufficient deflection angle or are highly wavelength dependent, causing smearing over the thermal wave bands. Rotating refractive polygons are now less used than previously. For fast line scanners, rotating reflective polygons (sometimes with curved facets) are used, in some cases using gas-bearing motors operating in helium to reduce windage. For frame scanning, a plane mirror driven by a powerful galvanometer is customary. This can give very linear sweeps at 60 Hz with scan efficiencies of about 80% when operated in a closed loop. For microscan or interlace, a small image movement can be achieved by tilting a mirror using piezoelectric actuators or by tilting a refractive plate using a galvanometer. For scanning systems, it is generally necessary to reimage the detector cold shield at each scan mechanism, since otherwise the optics must be enlarged to accommodate pupil movements. Figure 78.8 shows how this is done in a typical imager using a linear array. The movement of the interlace mirror is sufficiently small that pupil reimaging at this mirror is not required.

Figure 78.9 shows a typical arrangement for a 2-D scanner, a high-speed polygon rotor being used to generate the line scan.

The main optical parameters to be specified are transmission T_0, focal length F, and f/number $F_\#$ or numerical aperture (NA). Since IFOV = x/F, the focal length determines the spatial resolution for a given detector. The NA is defined as the sine of the semicone angle of the output beam from the optics. It can be shown that if a diffuse (Lambertian) source emits W W m^{-2} into a hemisphere, the radiance due to the source in the focal plane is $WT_0(\text{NA})^2$. $F_\#$ is defined as the ratio of focal length to diameter and for a distant object NA = $1/(2\ F_\#)$, so the ratio of the irradiance in the focal plane to the source radiance is

$$\frac{\text{Irradiance in focal plane}}{\text{Radiance from extended object}} = \frac{T_0(\lambda)T_a(\lambda)}{4F_\#^2} \tag{78.11}$$

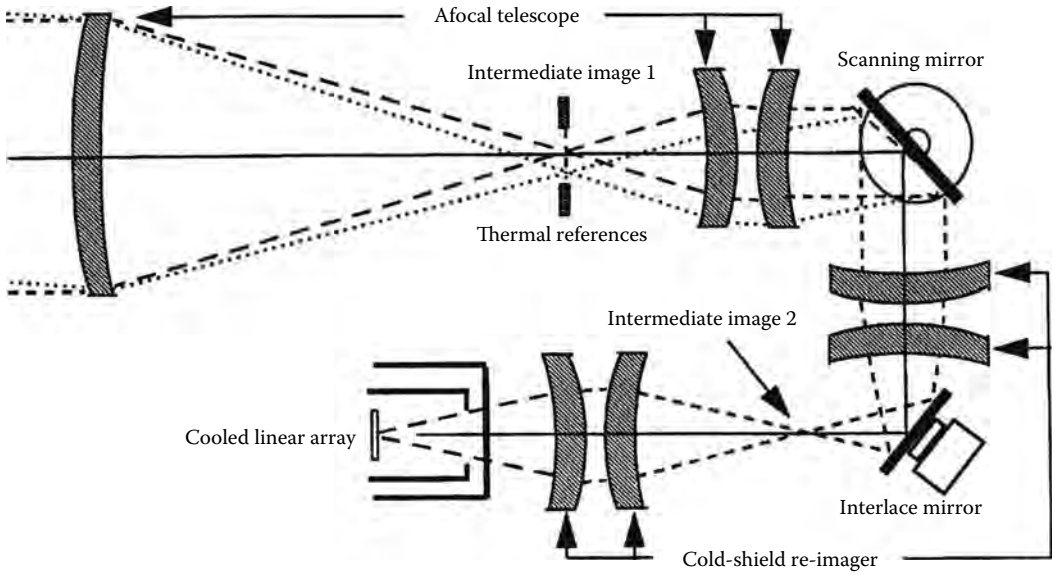


FIGURE 78.8 Optics for a linear array imager. The cold shield is imaged on the scanning mirror, and the scanning mirror is placed at the exit pupil of the telescope.

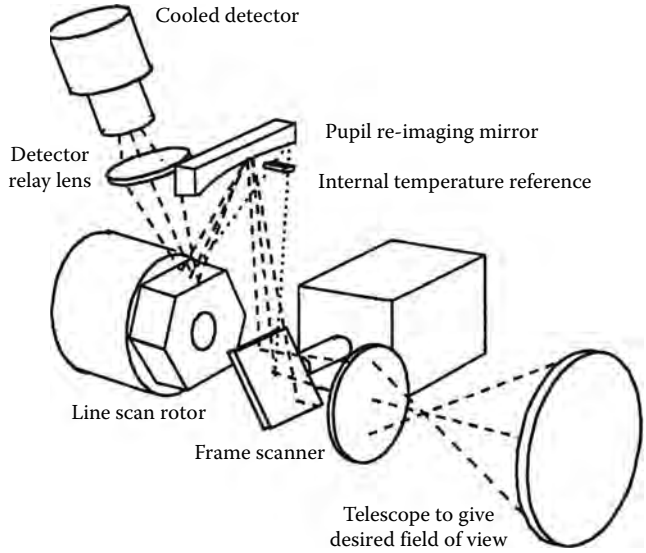


FIGURE 78.9 Simplified layout of imager using 2-D scanning (based on Pilkington Optronics HDTI). The line scan uses a high-speed polygon, and the concave strip mirror images the line scan pupil on the frame scanner.

The transmission of thermal imaging optics is typically 60%–90%, depending on complexity. The transmission of any optics between the temperature reference and the target must be known for quantitative measurement.

The imaging performance of a lens can be indicated by the size of the spot generated in the focal plane by a distant point object, and accurate results can be obtained if the intensity as a function of the angle α from the center of the image (point spread function, PSF) is known. A more usual approach is to use the MTF, defined as the ratio of the contrast of the image of a target with a sinusoidal spatial variation in intensity to the contrast of the target itself. The MTF depends on the spatial frequency f of the target

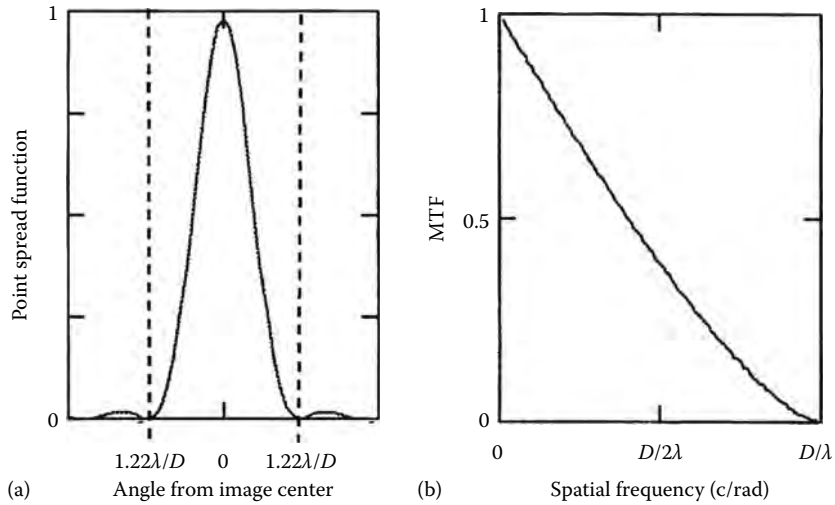


FIGURE 78.10 (a) PSF and (b) MTF for ideal optics of diameter D at wavelength λ. The first dark ring of the PSF has angular diameter $2.44\lambda/D$ and encircles 86% of the total energy. The MTF is zero for all frequencies above D/λ.

(expressed in cycles per radian) and on the wavelength of the radiation. For perfect optics of diameter D, diffraction gives the following values for PSF and MTF, which are also plotted in Figure 78.10:

$$\mathrm{PSF}(a) = 4\left[\left(\frac{\lambda}{\pi Da}\right)J_1\left(\frac{\pi Da}{\lambda}\right)\right]^2 \tag{78.12}$$

where J_1 is the Bessel function, and

$$\mathrm{MTF_o}(f) = \frac{2}{\pi}\left\{\mathrm{a}\cos\left(\frac{\lambda f}{D}\right) - \frac{\lambda f}{D}\sqrt{1-\left(\frac{\lambda f}{D}\right)^2}\right\} \tag{78.13}$$

In practice, optics is frequently far from the diffraction limit, so manufacturers' figures must be used. If a thermal imager is to be used for measurement, correction for MTF will be required unless the IFOV and the PSF are both smaller than the region over which the temperature is to be measured. The signal level for a subresolution point source can be obtained by integrating the PSF over the detector area. For diffraction-limited optics, a rule of thumb for the LWIR band is that the diffraction spot diameter in mrad is the reciprocal of the lens diameter in inches (or $25/D$ when D is in millimeters).

78.8 Temperature References

If a thermal imager is to be used for measurement, it is generally necessary to compare the signal from the target with that from one or more bodies at known temperature. The most precise results are obtained if two bodies at different known temperatures and having the same emissivity as the target (e.g., painted with the same paint type) are placed adjacent to the target and the target temperature is deduced by interpolation. There is then no dependence on emissivity, atmospheric transmission, and optical transmission. Outside the laboratory, this is seldom practical, however, and it is necessary to use thermal references within the imager itself. The fewer optical components there are between the reference and the scene, the better will be the accuracy. If the interior of a cavity of uniform temperature is viewed through a small

hole, the emission from the hole follows the Planck equation accurately, irrespective of the emissivity of the interior surface of the cavity. Although very accurate, such a blackbody cavity is usually inconveniently large, and it is more usual to use a blackened surface with deep grooves or pits to give high emissivity.

Unless the detector has only one element, thermal references are also desirable if not essential to allow the outputs of the different detector elements to be equalized.

78.9 Imager Performance

78.9.1 SNR and NETD

If the target is larger than the IFOV and the PSF, the signal level due to a small temperature difference is found by multiplying the source differential output, the atmospheric and optical transmissions, the geometric attenuation due to the *f*/number, the detector area, and the responsivity as derived or defined in the previous sections and integrating over the imager passband to give

$$V = \frac{A_d \Delta T}{4F_\#^2} \int R(\lambda) T_t(\lambda) T_a(\lambda) \varepsilon(\lambda) \frac{dW(\lambda,T)}{dT} d\lambda \tag{78.14}$$

The noise voltage V_n is given by Equation 78.7, so the SNR is obtained. If the number of serial detector elements is N_s, the signal level is increased by this factor, but since the noise is uncorrelated between the elements, it increases only as $\sqrt{N_s}$, so SNR improves as $\sqrt{N_s}$. The SNR also improves as the square root of the number of parallel channels since the dwell time is increased, giving a reduction in bandwidth B.

The sensitivity of a thermal imager for large targets is normally defined in terms of its NETD—the temperature difference between a large blackbody at zero range and its background—which gives a signal equal to the rms noise. This is given by V_n/V when $\Delta T = 1$, $T_a = 1$, and $\varepsilon(\lambda) = 1$ and is found to be

$$\text{NETD} = \frac{4F_\#^2}{\int_0^\infty D^*(\lambda) T_t(\lambda) (dW(\lambda,T)/dT) d\lambda} \sqrt{\frac{B}{A_d N_s}} \tag{78.15}$$

A good indication of SNR is given if it is assumed that T_a and ε are constant within the imager passband, in which case SNR = $\varepsilon\ T_a\ \Delta T$/NETD.

78.9.2 Minimum Resolvable Temperature Difference

The performance of a thermal imager is frequently defined by its minimum resolvable temperature difference (MRTD). MRTD(f) is the temperature difference between a four-bar square test pattern of frequency f c mrad^{-1} and its background required for an observer to count the imaged bars. The test is subjective but has the advantage of characterizing the complete system and display, including any effects of nonuniformity. MRTD is proportional to NETD and inversely proportional to the MTF and the square root of the number of frames presented within the integration time of the eye τ_e. The constant of proportionality depends on the degree of overlap between the scan lines. Discussion of the full MRTD model is beyond the scope of this chapter, but a simple expression (based on [1], p. 167)—which gives an indication of performance for an imager with square detector elements without overlap and at least two samples per IFOV and in which the electronics bandwidth is the same as that used for NETD calculation—is

$$\text{MRTD}(f) = \frac{3 \cdot \text{NETD} \cdot f \cdot \text{IFOV}}{\text{MTF}_o \text{MTF}_d \sqrt{\tau_e f_f}} \tag{78.16}$$

Calculation of MRTD is best performed using standard models, the most widely adopted being FLIR 92 [9].

78.10 Available Imagers

Recent advances in detector technology are only now beginning to be incorporated in commercially available systems. The result is that product ranges are currently changing rapidly and prices are very unstable. The prices for uncooled systems in particular are likely to drop significantly in the near future as the market size increases. An indication of cost at present is

Military high-performance imagers $100,000–300,000
Medium-performance imagers for measurement $30,000–100,000
Uncooled imagers $10,000–30,000 (but falling rapidly)

Prices reflect not only performance but also ruggedness, environmental survivability, and image processing software. Military imagers in many countries are based on "common modules" that are sometimes multisourced and that must be configured for specific applications. Table 78.1 lists some typical commercial imagers, which were selected to emphasize the wide variety of imager types, and the list must not be taken as a comprehensive survey. Some compact imagers weigh less than 2 kg, while some of the bench systems weigh over 100 kg. Accuracy of temperature measurement is not generally specified, but ±2 K or ±2% is a good figure for a calibrated imager. The information is based on published brochures, and there is a likelihood that improved models will be available by the time this book is published. It is strongly recommended that prospective purchasers should contact as many manufacturers as possible to obtain specifications and prices.

TABLE 78.1 Typical Commercially Available Thermal Imagers

Manufacturer	Model	Data	Description
AGEMA	880 LWB	LWIR, CMT	Thermal measurement system
		NETD = 0.07	175 pixels (50% MTF)
Amber	Radiancel	MWIR, InSb	Compact imager for measurement
		NETD = 0.025	256 × 256 pixels, InSb
Amber	Sentinel	LWIR	Uncooled compact imager
		NETD = 0.07 K	320 × 240 pixels
Cincinnati	IRRIS-160ST	MWIR, InSb	Compact imager
		NETD = 0.025	160 pixels/line
FLIR Systems	2000F	LWIR, CMT	Surveillance imager
		NETD = 0.1 K	>350 pixels/line
GEC Sensors	Sentry	LWIR	Uncooled low-cost imager 100 × 100 pixels
Hamamatsu	THEMOS 50	MWIR	Microscope, 4 μm resolution
		NETD = 0.2 K	256 × 256 pixels
Inframetrics	ThermaCAM	MWIR, PtSi	Thermal measurement system
		NETD < 0.1 K	256 × 256 pixels
Mitsubishi	IR-M600	MWIR, PtSi	High-definition imager
		NETD = 0.08 K	512 × 512 pixels
Nikon	LAIRD-3	MWIR, PtSi	High-definition imager
		NETD = 0.1 K	768 × 576 pixels
Pilkington Optronics	LITE	LWIR	Handheld surveillance imager
		NETD = 0.2 K	350 × 175 pixels
Quest	TAM200	MWIR	Microscope with probe facility
		NETD = 0.05 K	Bench system with 12.5 μm resolution

78.11 Performance Trade-Offs

The expression for NETD presented earlier is appropriate for evaluation of existing systems where D^* is known. To predict performance of future systems based on photon detectors and to carry out design trade-offs, it is important to appreciate that D^* is very dependent on conditions of use and on wave band. One reason is that the photons from an object of given temperature, although having a well-defined average flux, are emitted at random times. Statistical theory shows that if on average, N photons are collected within a given time interval, the standard deviation is $\sqrt{N}$ if N is large enough to make the distribution Gaussian (as is almost always the case in the infrared due to the high background flux). This "photon noise" frequently predominates, in which case the detector is referred to as background-limited photodetector (BLIP). The D^* is then determined more by the conditions of use than by the detector itself. If the efficiency of the cold shield η_c is defined as the ratio of the effective f/number for the background flux to that of the signal, the number of background electrons generated within an integration time τ is given by

$$N_e = \frac{\tau A_d}{4F_\#^2 \eta_c^2} \int_0^\infty N(\lambda,T)\eta(\lambda)d\lambda \tag{78.17}$$

For the reasons discussed earlier, the rms electron noise for a BLIP device within time τ is simply the square root of the aforementioned figure. If the detector also has readout noise of N_n electrons rms, this could be regarded as the noise that would be caused by a background flux that caused the generation of N_n^2 photoelectrons, so the noise can be written as

$$\text{Noise} = \sqrt{N_e + N_n^2} \text{ electrons rms} \tag{78.18}$$

The signal due to a temperature difference of 1 K between target and background expressed in electrons within the integration time is

$$\text{Signal} = \frac{\tau A_d}{4F_\#^2} \int_0^\infty T_t(\lambda)T_a(\lambda)\eta(\lambda)\frac{dN(\lambda,T)}{dT}d\lambda \tag{78.19}$$

Thus, the SNR for $\Delta T = 1$ K is found, NETD being simply 1/SNR when $T_a = 1$.

For staring systems, and some scanning systems, the integration time is limited by saturation. If the maximum number of electrons that can be stored is N_m, we find that the maximum integration time is

$$\tau_m = \frac{4F_\#^2 \eta_c^2 N_m}{A_d \int_0^\infty N(\lambda)\eta(\lambda)d\lambda} \tag{78.20}$$

When the integration time is controlled to restrict the number of electrons generated within the sampling interval to τ_m, the noise is given by letting $N_e = N_m$ in Equation 78.18 so is independent of f/number. If one also substitutes τ_m for τ in Equation 78.19 and divides Equation 78.19 by Equation 78.18, to obtain SNR, one obtains

$$\text{SNR} = \frac{\Delta T \eta_c^2 N_m \int T_o(\lambda)T_a(\lambda)\eta(\lambda)(dN(\lambda,T)/dT)d\lambda}{\sqrt{N_m + N_n^2}\int N(\lambda)\eta(\lambda)d\lambda} \tag{78.21}$$

The ratio of integrals is effectively the contrast of the scene behind the optics, which is further reduced by the inefficiency of the cold shield; thus, performance depends mainly on image contrast at the detector and the number of electrons stored. (Quantum efficiency essentially cancels out, and for a saturated device, $\sqrt{N_m}$ is usually greater than N_n.)

The aforementioned formulae are sufficiently general to allow trade-offs for photon detectors to be carried out against aperture, wave band, range, and frame rate, the main uncertainty being the readout noise. For staring array detectors with processing on the focal plane, a typical figure is $N_n = 1000$, but much better (and worse) values are possible. The SNR and NETD values obtained in this way are for a single pixel. If the samples overlap spatially, the SNR is improved by the square root of the overlap factor. In practice, nonuniformity of the detector array will introduce spatial noise that will affect MRTD, and even after electronic correction, this is frequently the main limitation on performance. Nonuniformity effects are discussed in [10,11].

The resulting trade-offs are discussed at some length in [12]. It is observed that if the integration time is controlled to prevent saturation, NETD is independent of *f*/number and quantum efficiency but depends strongly on image contrast. This tends to favor the MWIR band and makes high cold-shield efficiency crucial. The problem can be overcome in the LWIR by operating at a high frame rate, which has no effect on NETD but improves MRTD due to eye integration. For long-range applications, it is frequently necessary to use long focal lengths, giving large *f*/numbers if the optics diameter is to be kept within reasonable bounds. Under these conditions, saturation is less likely, and the high photon flux in the LWIR is necessary to overwhelm the readout noise. If high spatial resolution rather than sensitivity is required, the MWIR band has a much better diffraction limit. Figure 78.11 indicates the ways in which the various limitations combine to define the combination of IFOV and wave band that gives optimum performance when it is necessary to resolve a given spatial frequency; the positions of the lines on this chart are of course peculiar to the system under investigation.

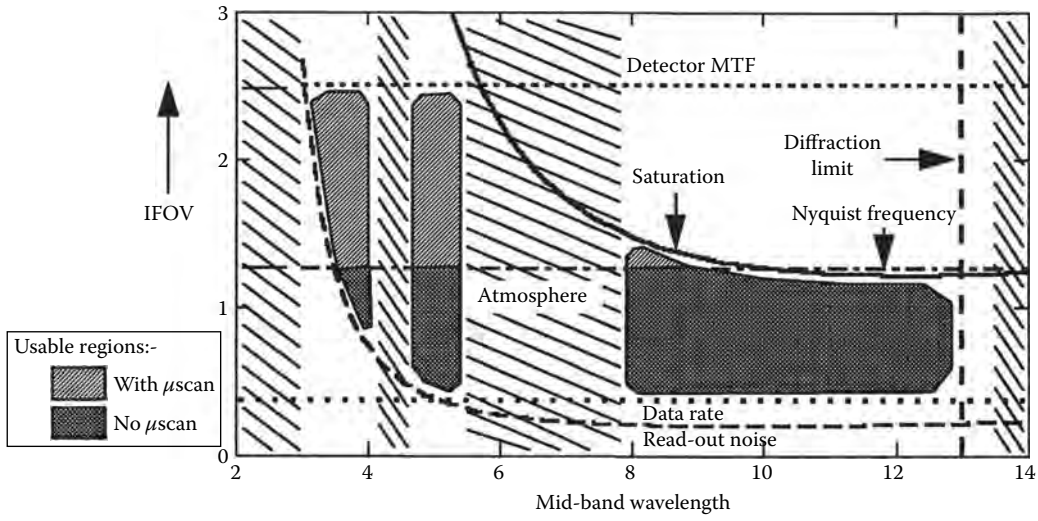


FIGURE 78.11 Physical limitations to thermal imaging when a specified spatial frequency must be resolved. The shaded regions indicate the combinations of wavelength and IFOV for which the task can be achieved. All the lines on the graph depend on system parameters such as aperture, quantum efficiency, readout noise, and electron storage but can be calculated from the equations given in the text.

TABLE 78.2 Companies That Make Commercial Thermal Imagers

AGEMA Infrared Systems Inc., 550 County Avenue, Secaucus, NJ 07094, Tel.: (201) 867-5390
Amber Engineering Inc., 57566 Thornwood Drive, Goleta, CA 93117-3802, Tel.: (800) 232-6237, Fax: (805) 964-2185
Cincinnati Electronics Corporation, Detector and Microcircuit Devices Laboratories, 7500 Innovation Way, Mason, OH 45040-9699, Tel.: (513) 573-6275, Fax: (513) 573-6290
Hamamatsu Photonic Systems, 360 Foothill Rd., Bridgewater, NJ 08807-0910, Tel.: (908) 231-1116, Fax: (908) 231-0852, http://www.hamamatsu.com/
Inframetrics Imaging Radiometer Group, 12 Oak Park Drive, Bedford, MA 01730, Tel.: (617) 275-8990, TWX: (710) 326-0659, www.inframet.pl/
Mitsubishi Electronics America Inc., 5665 Plaza Drive, P.O. Box 6007, Cypress, CA 90630-0007, Tel.: (800) 843-2515, http://www.mitsubishi-imaging.com/
Nikos Corporation (Nikon), 1502 West Campo Bello Drive, Phoenix, AZ 85023, Tel.: (602) 863-6182, http://www.nikon.com/
Pilkington Optronics Inc., 7550 Chapman Ave., Garden Grove, CA 92841, Tel.: (714) 373-6061, Fax: (714) 373-6074, http://www.pilkington.com/
Quest Integrated Inc., 21414-68th Avenue South, Kent, WA 98032, Tel.: (206) 872-9500, Fax: (206) 872-8967, http://www.qi2.com/

78.12 Future Trends in Thermal Imaging

At the high-performance end of the market, the developments are most likely to be driven initially by military requirements. Many countries already have thermal imaging common module programs, and many developments will be aimed at productionizing these modules to reduce cost and enhance performance. Multispectral instruments to aid in camouflage penetration and to broaden the conditions of operability through the atmosphere can be taken out of the laboratory into service. Currently, these are mainly scanning devices using adjacent long linear arrays of LWIR and MWIR detectors. For spectral agility, detector arrays based on multi-quantum-well (MQW) technology can be used, since they can be tuned to some extent by varying the electrical bias. Multispectral refracting optics is complex and has poor transmission, so reflecting optics is generally used. Multispectral systems will allow more accurate compensation for emissivity, but their high cost might limit their use except for high-value applications such as earth resources surveys in aircraft or satellites. Techniques employed include frame-sequential filtering and imaging Fourier transform spectroscopy (Table 78.2).

For commercial thermal imaging, the main thrust is likely to be cost reduction via the use of uncooled imagers and in packaging to provide increased ease of use and functionality using in-built software. Optics cost can be reduced or performance improved by the use of hybrid aspheric components, which are now practical due to improvements in diamond turning. A diffractive surface is generated on one surface of a refracting lens. Because the power of a diffractive component is proportional to wavelength, a very low diffractive power can compensate for the chromatic aberration of the lens. This allows the use of materials such as zinc sulfide, which would otherwise be unacceptable due to chromatic aberration but which have other desirable properties such as low cost or a low thermal coefficient of refractive index.

References

1. J.M. Lloyd, *Thermal Imaging Systems*, New York: Plenum Press, 1975, pp. 20–21.
2. J.S. Acetta and D.L. Shumaker (Eds.), *The Infrared & Electro-Optical Systems Handbook*, Bellingham, WA: SPIE Optical Engineering Press, 1993, Vol. 1, 52, pp. 251–254.
3. F.X. Kneizys, E.P. Shettle, L.W. Abreu, G.P. Anderson, J.H. Chetwynd, W.O. Gallery, J.E.A. Selby, and S.A. Clough, *Users' Guide to LOWTRAN 7*, Report No. AFGL-TR-88–0177 Hanscom, Air Force Geophysics Laboratory, 1988.
4. M. Bass (Ed.), *Handbook of Optics*, New York: McGraw-Hill, 1995, Vol. 1, Chapter 23.

5. C.M. Hanson, Uncooled ferroelectric thermal imaging, *Proc. SPIE 2020*, *Infrared Technology XIX*, 1993.
6. R.A. Wood, Uncooled thermal imaging with monolithic silicon focal plane arrays, *Proceedings of SPIE 2020*, *Infrared Technology XIX*, San Diego, CA, p. 329, 1993.
7. G.C. Holst, *Testing and Evaluation of Infrared Imaging Systems*, Winter Park, FL: JCD Publishing, 1993, pp. 36–42.
8. M.J. Riedl, *Optical Design Fundamentals for Infrared Systems*, Bellingham, WA: SPIE Optical Engineering Press, 1995, p. 55.
9. L. Scott and J. D'Agostino, NVEOD FLIR92 thermal imaging systems performance model, *Proceedings of SPIE*, *Vol. 1689*, *Infrared Imaging*, pp. 194–203, 1992.
10. A.F. Milton, F.R. Barone, and M.R. Kruer, Influence of nonuniformity on infrared focal plane performance, *Opt. Eng*., 24, 855–862, 1985.
11. H.M. Runciman, Impact of FLIR parameters on waveband selection, *Infrared Phys. Technol*., 37, 581–593, 1996.

Further Information

Catalina Bixler, *Types of Thermal Sensors, Fiber Optic Transducers*, http://www.ehow.com/list_6767046_types-thermal-sensors.html (accessed on June 21, 2012).

Compix, Thermal imaging: more than just pretty pictures, http://www.compix.com/articles.quality.html (accessed on June 28, 2012).

Gaussorgues, G., *Infrared Thermography*, London, U.K.: Chapman and Hall, 1994.

Grum, F. and Becherer, R.J. (Eds.), *Optical Radiation Measurements*, New York: Academic Press, 1979, Vol. 1.

Holst, G.C., *Electro-optical Imaging System Performance*, Winter Park, FL: JCD Publishing, 1993, pp. 36–42.

Wyatt, C.L., *Radiometric Calibration: Theory and Methods*, New York: Academic Press, 1978.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

79

Calorimetry Measurement

79.1 Introduction **79**-1
Thermal Energy, Heat, and Temperature • Work and Enthalpy • Signals Measured with Calorimeters
79.2 Calorimeters: Criteria for Their Classification..... **79**-3
Isothermal Calorimeters • Phase-Transition or Thermoelectric Compensation of the Heat • Adiabatic Calorimeters Often Measure Time-Dependent Temperature Differences • Isoperibol and Heat-Flux Calorimeters • Chip Calorimeters
79.3 Scanning Calorimetry Sweeps the Experiment Temperature ... **79**-6
Differential Scanning Calorimeters Techniques • Converting the Measured Curve to the Actual Progress of the Experiment
79.4 Calibration of Calorimeters **79**-9
Calibration of Chip Calorimeters
79.5 Typical Applications of Calorimeters..... **79**-10
Specific Heat, Transition Heat, and Temperatures • Analysis of Chemical and Physical Reactions
79.6 Thermal Analysis of Materials and Their Behavior with Temperature..... **79**-10
Biological Analysis from Cells to Entire Human Beings • Summary
79.7 Choosing the Proper Calorimeter for an Application **79**-13
What Do You Want to Measure? • Budget and the Need of Pre- and After-Sales Service
79.8 Can the Instrument of Choice Measure the Signals Desired?.... **79**-13
Instrument Control and Data Management
79.9 Commercially Available Calorimeters **79**-15
References..... **79**-15
Further Information..... **79**-16

Sander van Herwaarden
Xensor Integration

Elina Iervolino
Xensor Integration

79.1 Introduction

This chapter is a revision of the homonymous chapter contained in the previous edition of this book. While the overview of the most important calorimeter showed in the previous chapter has been left, more space is given to the latest innovation in the field of chip calorimeter.

This chapter starts again with a small introduction on thermodynamics. An overview of the most important types of calorimeters and chip calorimeters will then be presented. The typical application of calorimeters and chip calorimeters follows. For those who want to do measurements without having to make their own devices, an overview of the most important calorimeters and their vendors and approximate prices will be given. The chapter concludes with some hints on present and future developments and on further reading, since there is a large amount of literature on calorimeters.

79.1.1 Thermal Energy, Heat, and Temperature

One of the forms in which energy can be present in a system is the random, internal kinetic energy of the particles (molecules or atoms) of a system, which can intuitively be called "thermal energy." Thermal energy is the portion of internal energy U that is responsible for a system temperature and has to be distinguished from the average, external movement of a system of particles as a whole, which can be called the "mechanical energy" of the system. For gases, thermal energy is closely related to the random velocity of the molecules and, in the case of multiatom molecules, the rotations and vibrations of the atoms within the molecules. The thermal energy is an extensive property as it is dependent by the amount of matter the object contains. The temperature is instead an intensive property, and its empirical definition comes out from the zeroth law of thermodynamics. The zeroth law of thermodynamics states that if two systems are each in equilibrium with a third system, they will be in equilibrium with each other. They are said to have the same temperature. For gases, statistical mechanics shows the direct relation between the thermal energy stored in the system and the temperature [1]. This law, however, also applies to liquids and solids, although the quantitative relation between thermal energy (agitation of the particles) and temperature is less straightforward and not so easy to calculate as for gases. Temperature and thermal energy distribution can be viewed as the result of statistical processes, such as diffusion. Diffusion ensures that if ever there is a surplus of thermal energy (i.e., of fast molecules or electrons or a higher density of phonons) in some area, some of it will flow toward areas with a lower thermal energy density until thermal equilibrium has been established. This flow of energy P (in W = J s^{-1}) is called heat flow. Heat flow is therefore the transfer of (thermal) energy from one (part of a) system to another. Note that heat is not conserved, because a change in thermal energy of a system may be achieved by heat exchange with the environment but also by mechanical interaction, and heat can be used to do mechanical work (motor) just as well as heating up a system. Heat can be viewed as "thermal energy on the move." Heat is energy transferred from one body to another due to thermal contact when the systems are at different temperatures, but whether it ends up as thermal, mechanical, or another form of energy depends on the circumstances.

79.1.2 Work and Enthalpy

The first law of thermodynamics is an expression of the principle of the conservation of energy. In a simple form, in the absence of other forms of energy exchange, it states that the change in internal energy ΔU of a system is equal to the heat Q supplied by the ambient to the system minus the work (mechanical energy) $p\Delta V$ done by the system on the ambient:

$$\Delta U = Q - p\Delta V \tag{79.1}$$

This is of importance when measuring the heat capacity C (in J K^{-1}) of a substance, defined as the amount of heat required to change the temperature of the substance by a given amount:

$$C = \frac{\Delta Q}{\Delta T} \tag{79.2}$$

where

ΔQ is the heat exchanged

ΔT is the temperature change of the substance

This measurement can be performed under two conditions: at constant volume V and at constant pressure p. At constant volume, no work will be done by the system, since $p\Delta V = 0$. So, the heat capacity

at constant volume c_v is related to the change in internal energy ($\Delta Q = \Delta U$)0. The specific heat capacity of a system at constant pressure c_p is (usually) higher, since additional energy (heat) is needed to perform work on the ambient: $p\Delta V$. The quantity combining these contributions to the energy at constant pressure is called enthalpy H, and the change in enthalpy versus temperature (in J K^{-1}) at constant pressure is the heat capacity C_p of a sample at constant pressure: $C_p = (dH/dT)_p$.

79.1.3 Signals Measured with Calorimeters

In measurement science, physical quantities can be distinguished by six different so-called signal domains: thermal, electrical, chemical, mechanical, magnetic, and radiant [2]. Signals can be transduced from one domain to the other. For a thermal signal, the change of the ambient temperature can be transduced into an electrical signal thanks to a resistor, reading its temperature-dependent resistance value with a multimeter. All signals in calorimeters are either thermal signals or other signals that are transduced into thermal signals. So, calorimetric measurement of signals can occur from all domains [3]. The thermal signal (e.g., heat) is then transduced into an electrical signal. Calorimeters usually measure the thermal effects of (bio)chemical or mechanical processes, or they measure the thermal effect of temperature changes on matter.

79.2 Calorimeters: Criteria for Their Classification

In essence, a calorimeter performs three functions: It encloses a chamber in which a thermal experiment is carried out; it measures the heat exchange between the sample under test and the calorimeter (and often other quantities are being measured as well, such as temperature and amount of substance); and it thermally separates the experimental chamber from its surroundings. The same three functions are realized also in a chip calorimeter. In silicon (Si) technology, a chip calorimeter usually schematically consists of a heater and a temperature sensor on a thin-film membrane connected to the Si frame that gives mechanical stability to the calorimeter, and it acts as thermal heat sink; see Figure 79.1. For a chip calorimeter, the three aforementioned functions are realized by the hot spot that delimits the sample area where the experiment is performed, the temperature sensor that measures the temperature, and the thin membrane that thermally isolates the sample area from the surrounding. There exist many types of calorimeters. Four essential criteria can be used to classify calorimeters [4,5]:

1. *Heat measurement*. What does the calorimeter do with the heat that is generated (or absorbed) by the experiment?
2. *Relation with surrounding*. How does the calorimeter relate to its surroundings?
3. *Single or twin design*. Does the calorimeter measure a single experimental sample, or is it a twin design with a reference compensating for common-mode errors?

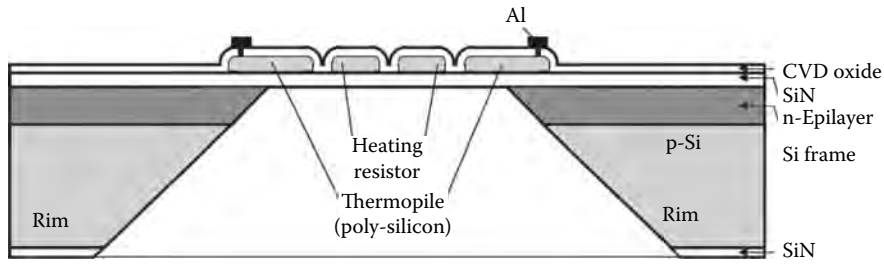


FIGURE 79.1 Schematic of a chip calorimeter: on a thin-film membrane, the heater, which makes up the sample area, and the hot junctions of the thermopile (temperature sensor) are located. The cold junctions of the thermopile are on the Si frame that acts as thermal heat sink, and it gives also mechanical stability to the device. (Picture loaded from the website www.xensor.nl)

TABLE 79.1 Overview of the Classification Criteria for Calorimeters

Relation to Surroundings Heat Measurement	Heat Measurement
Isothermal	Phase-transition compensation of heat Thermoelectric compensation of heat
Adiabatic	Measurement of temporal temperature difference
Isoperibol	Measurement of spatial temperature difference

4. *Fixed temperature or scan of a temperature range*. Does the calorimeter function at a fixed temperature, or can it scan a temperature range?
   a. A fifth criterion can be added to include calorimeters that are based on a calorimeter chip.
5. *Conventional design or chip design*. Is the calorimeter based on a "conventional design," or is it based on a chip calorimeter?

Table 79.1 lists the various possibilities for the relation to the surroundings and for the heat measurement. In principle, almost every combination is possible. In practice, some combinations naturally go together because they compensate for their strengths and weaknesses. For example, calorimeters in which the experimental temperature is scanned often use a twin configuration to eliminate the common-mode errors arising from the continually changing temperature. In the following paragraphs, the main types of calorimeters are described.

79.2.1 Isothermal Calorimeters

In the isothermal calorimeter, the experiment is always kept at a fixed temperature. This is attained by instantly removing (or supplying) any heat that the experiment releases (or absorbs). The isothermal calorimeter was the first to be developed. In 1780, Lavoisier and Laplace made the "ice calorimeter" in which the heat generated by the experiment is used to melt ice into water. If enough ice is available, the calorimeter will remain at 0 °C, regardless of the progress of the experiment. The experimentally generated heat Q (J) is calculated by weighing the mass m (kg) of the melt water and multiplying by the heat of the ice-water transition q_{fus} (J kg^{-1}):

$$Q = m q_{\text{fus}} \tag{79.3}$$

The experiment with the melting ice is enveloped within a thermostat, which consists of a double-walled vessel with melting ice between the walls, which is always at 0 °C as well. Thus, the isothermal calorimeter has a perfect thermal isolation between experiment and surroundings in the form of a second guard vessel that buffers all the heat coming from the surroundings.

79.2.2 Phase-Transition or Thermoelectric Compensation of the Heat

In the isothermal calorimeter, the heat generated by the experiment is immediately absorbed by the calorimeter. This can be accomplished by phase-transition compensation of the heat, for example, by melting of solids or evaporation of liquids. But nowadays, compensation by electrical means is preferred because it can be measured so easily. Thus, heat Q (J) absorbed by the experiment is replaced using Joule heating (dissipation of heat by a current I (A) through an electric heater with resistance R_{h} (Ω)), while heat generated by the experiment is absorbed by Peltier coolers:

$$Q = \int I^2 R_{\text{h}} \mathrm{d}t \tag{79.4}$$

79.2.3 Adiabatic Calorimeters Often Measure Time-Dependent Temperature Differences

In the adiabatic calorimeter, no heat exchange with the surroundings is allowed, and all the heat generated by the experiment is used to increase the temperature of the calorimeter. The amount of heat Q (J) generated follows from the temperature increase ΔT (K) divided by the heat capacity of the calorimeter C_c (K J^{-1}):

$$Q = \frac{\Delta T}{C_c} \tag{79.5}$$

The absence of heat exchange with the surroundings of the calorimeter is obtained by immersing the experimental chamber of the adiabatic calorimeter in an outer vessel. The temperature of the outer vessel is kept at the same (increasing) temperature as the experimental chamber by means of electronic feedback, heating the outer vessel to maintain a practically zero temperature difference. In the adiabatic calorimeter, one must wait a few minutes after the experiment has finished to allow the heat to spread itself uniformly over the chamber and to obtain the final temperature. An example of such calorimeter is a so-called calorimetric bomb. In its experimental chamber, fuel is fully burned to measure its heat of reaction. The bomb is immersed in a vessel filled with water—the inner vessel. A stirrer assures fast heat exchange between the bomb and the inner vessel; thermometers measure the temperature of the inner guard vessel and the outer guard vessel. The outer vessel is regulated to the inner-vessel temperature by means of electrical heaters and refrigerator coolers.

79.2.4 Isoperibol and Heat-Flux Calorimeters

The term *isoperibol* was devised to indicate a calorimeter having "uniform surroundings." In this calorimeter, the outer shell provides a reference temperature, and customarily, the experiment starts at the same temperature. The experimental chamber is connected to the outer shell by a well-defined thermal conductance. Any heat generated by the experiment will cause a well-defined temperature difference ΔT (K) across the thermal conductance G_{th} (W K^{-1}), and this temperature difference is subsequently measured as a "local temperature difference." Calorimeters utilizing this way of measuring the heat are referred to as "heat-flux calorimeters" and often measure the power P (J s^{-1} or W) generated by the experiment, rather than the energy:

$$P = \Delta T G_{th} \tag{79.6}$$

Chip calorimeters are an example of isoperibol calorimeters.

79.2.5 Chip Calorimeters

With respect to "conventional calorimeters," chip calorimeters bring in the calorimeter field two important improvements: the scale-down of the sample to nanogram amounts of material [6,7] and the possibility to increase the temperature scan rate to 10^6 K/s (scanning calorimeter is presented in the next paragraph). Ultrafast-scanning calorimeters are commercially available, and extensive characterizations are present in literature [8,9]. Calorimeter chips generally consist of a heater resistor and thermopile (temperature sensor) on a thin-film dielectric material. Different designs of calorimeter chips have been proposed with the intent to come to a device capable of fast temperature scan rates [9–11]. So far, the use of these sensors was restricted to isoperibol operation around room



FIGURE 79.2 Flash DSC 1 of Mettler-Toledo with a close-up of the device UFS 1. The device is made of two identical cells, the reference and the sample cell, respectively. Both cells consist of two heaters covered with an aluminum layer that makes up the sample area, in the center of the membrane; the hot junctions of the eight thermocouples are inside the sample area.

temperature [12], using aluminum heat sinks to provide the reference temperature. The main drawback of the chip calorimeters was the absence of a commercial available apparatus and software for the signal analysis. The problem has been overcome with the recent introduction on the market of the Flash DSC 1 of Mettler-Toledo. Figure 79.2 shows the Flash DSC 1 with a close-up of the chip designed for the instrument. The chip is made of two identical cells, the reference and the sample cell, respectively. Both cells consist of a 1.6×1.6 mm^2 large and about 2 μm thick membranes. The membranes of the two cells are thermally separated by a thick Si frame of 300 μm that acts as heat sink. Two heaters are integrated in the center of the membrane. The main heater, with a resistance of about 5 kΩ, is used for the general temperature scan program. The other heater, of about 4 kΩ, is active only in power compensation mode and has to compensate for temperature differences between the reference and the sample cell. The two heaters, covered by an aluminum (Al) layer, make up a circular area with a diameter of 0.5 mm, where the sample can be loaded. The Al layer is used to get temperature uniformity in the sample area. The temperature of the heated area is measured with an integrated thermopile.

79.3 Scanning Calorimetry Sweeps the Experiment Temperature

In the scanning calorimeter, the experimental chamber is not kept at (approximately) one temperature, but it is swept over a temperature range. The temperature is increased at given rate (e.g., 10 K min^{-1}) by electric heating or decreased by, for example, cooling with liquid nitrogen. To achieve the temperature sweep, the experiment is placed inside a computer-controlled oven. To compensate for common-mode errors (i.e., backlogging or deviation of the sample temperature from the oven temperature and imperfections in the oven temperature profile with respect to time and location), scanning calorimeters are often built with twin experimental sites for which the difference is measured. One site is for the sample under test; the other site serves as reference, which either is left empty or contains material resembling the sample under test as much as possible, except for the phenomena to be measured. Such calorimeters are called differential scanning calorimeters, DSCs (see "Further Information" for a general book on DSC). There are three principal types of DSCs design: heat flux, power compensation, and (the new introduced) dynamic power compensation [13]. In heat-flux measurements, the reference and the sample have a common heating source, or more in general, they are subjected to the same heating/cooling function; see Figure 79.3. The output signal is the temperature difference between the reference and the sample site. In the second type of DSC, the power compensation DSC, sample and reference each have an individual heating source, which is electronically controlled to maintain sample and reference at the same required temperature course. The difference between reference and sample

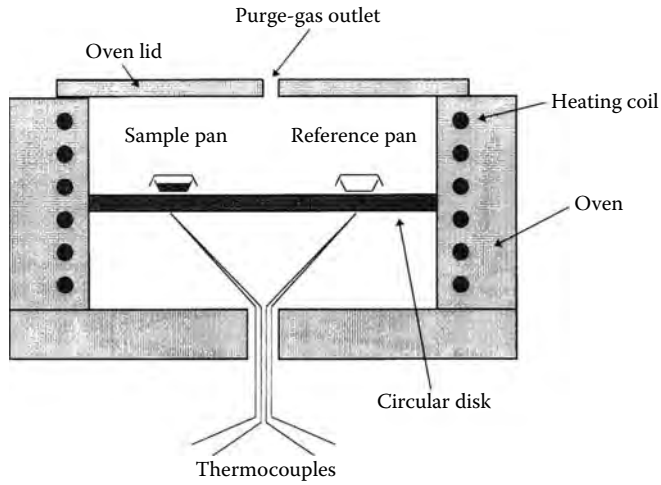


FIGURE 79.3 The heat-flux DSC consists of two experimental sites on a heat-conducting disk. Heat generation in the sample pan results in a temperature increase of the sample with respect to the reference, measured by thermocouple (the disk being one part of the couple).

heating power immediately gives the experimental heat. In dynamic compensation DSC, sample and reference still have each an individual heating source, but the sign of the applied compensation power is always positive. This can be made by dynamic switching the cell to which the excess power is applied. The additive nature of this compensation principle enables implementation by dedicated compensation heating units, physically distinct from the heaters empowering the common temperature program. The separation of compensation power from bulk power offers advantages in terms of noise, signal resolution, and response time.

There is also a fourth (less accurate) type of DSC, where only the temperature at which phenomena occur is registered. This type can be used for bigger samples (2–3 mL against 0.1 mL for the heat-flux DSC) or at very high temperatures up to 2000 °C.

79.3.1 Differential Scanning Calorimeters Techniques

In DSCs, three different techniques have recently become popular: temperature-modulated DSC (TM-DSC), high-sensitivity DSC (HS-DSC), and fast-scan DSC (FS-DSC). Each technique has some advantages, and the choice of one over the other technique has to be done in dependence of the sample to be studied. The TM-DSC and the FS-DSC techniques are treated in more details in the following subparagraph.

79.3.1.1 Temperature-Modulated DSCs Improve Insight into the Measurement

One of the developments in DSC is the use of a nonuniform temperature scan. On the standard linear temperature increase, mostly slow (e.g., 1 K min^{-1}), an alternating temperature change is superimposed, which is fast compared with the linear scan rate. The signal can be a sinusoidal signal, or alternatively, a stepped temperature profile can be used. Here, the temperature is increased during, for example, 0.5 min with 2 °C, and then stabilized for 0.5 min, resulting in an overall scan rate of 2°C min^{-1}. In fact, the scan is now modulated with a block wave of amplitude 2 °C min^{-1}; see Figure 79.4. For an instrument with small time constant, this means that the heat flow for increasing the experiment temperature in the stabilization time will be completely stopped. Any heat flow remaining results from the processes in the sample itself; see Figure 79.4.

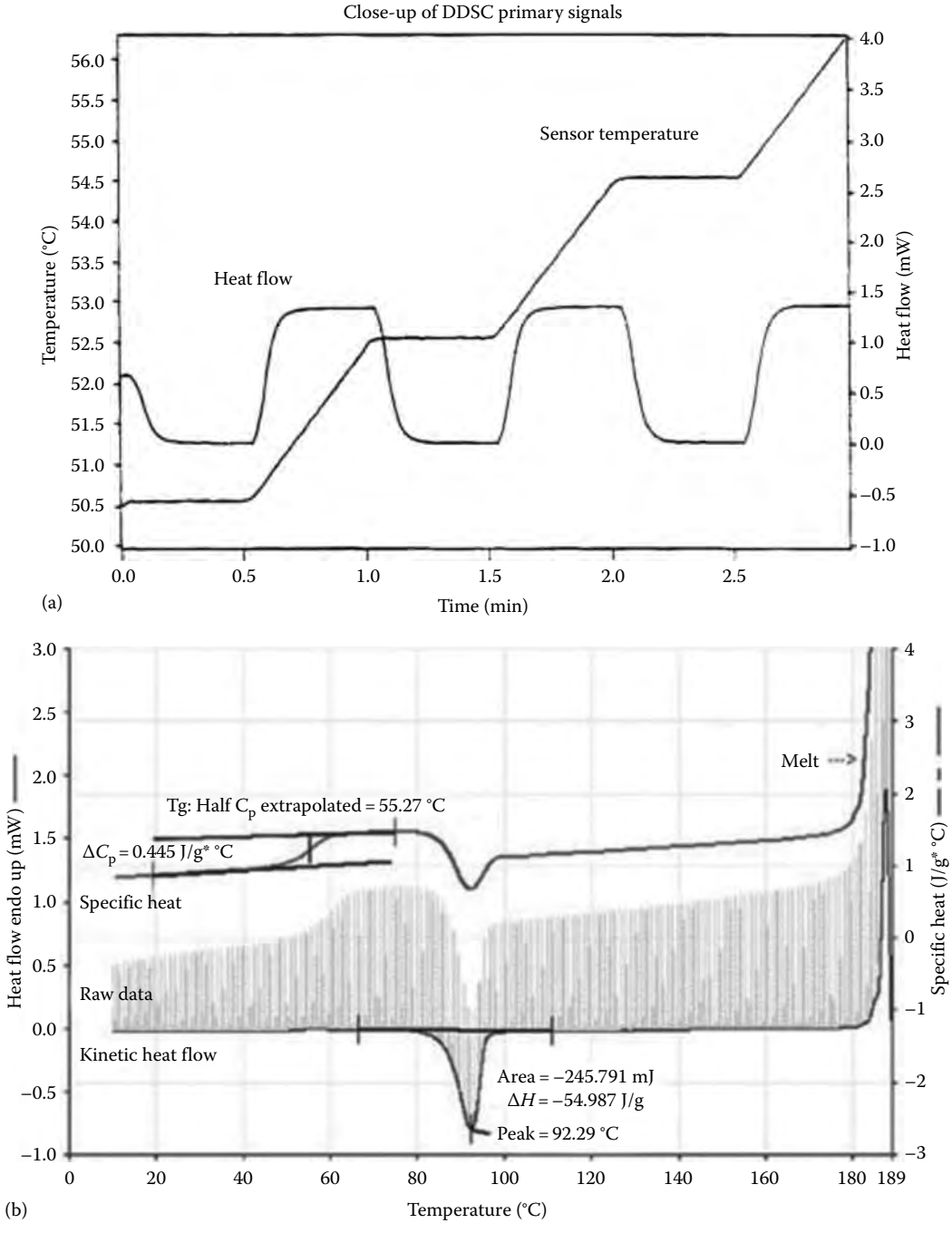


FIGURE 79.4 In modulated or dynamic DSCs, the temperature is not increased at a given rate, but modulated with a sinusoidal or (a) stepwise deviation to separate heat transfer due to temperature increase (specific heat) and heat transfer due to phase changes in the material (glass transition, crystallization, melting). (b) TM-DSC of carbamazepine shock cooled from the melt. This figure shows the absence of heat flow in regions without material change, while the curve indicates the glass transition of the amorphous form and the crystallization into a crystalline polymorphic form, which then melts. (Courtesy of Perkin Elmer, Waltham, MA.)

79.3.1.2 FS-DSCs and FS-DSCs Based on Chip Calorimeter

FS-DSCs gain popularity, thanks to the new possibilities of fast scan rates that brought

1. The study of materials in nonequilibrium states. Fast scan rates prevent reorganization during the scan.
2. The study of microgram–gram samples. Fast scan rates yield to heat flows that can be detected by FS-DSC.
3. Short run time.
4. Higher sensitivity than conventional DSCs with slow scan rates. The gain in sensitivity (the magnitude of each thermal event) corresponds to a loss in resolution (separation between thermal events) [14].

FS-DSCs can be realized either modifying existing machines or using chip calorimeters. In the first case, the drawback is that only a few hundreds of K/s can be realized, while with chip calorimeters, 10^6 K/s have been demonstrated to be possible. The recent introduction on the market of the Flash DSC 1 of Mettler-Toledo combined the advantages of the chip calorimeters with the comfort of a standard DSC.

79.3.2 Converting the Measured Curve to the Actual Progress of the Experiment

Due to the thermal resistance and the heat capacitance inside the experimental chamber of a calorimeter, there is a time constant associated with distributing the experimental heat over the chamber and reaching the final temperature. A heat pulse does not create a temperature pulse but rather a smeared-out curve. In a DSC, this smearing out is usually not significant because of the small time constant (1–3 s). For some other instruments, the so-called curve desmearing has to be carried out using convolution integrals to extract all the information from the measurement curve. With commercially available instruments, software is often supplied that does the job for you, but the accuracy of this software is not satisfactory in all cases, so one might need to do some further study on desmearing [4].

79.4 Calibration of Calorimeters

The inaccuracy of the calorimeter can be reduced by *calibration*. Heat production can easily be simulated electrically. Unfortunately, the thermal leakage of the electric leads reduces the accuracy of this method. More common, therefore, is the use of primary standards with a known heat of transition or reaction for calibration. For bomb calorimeters, the response can be calibrated using the heat of reaction of benzoic acid, which has been carefully determined to be 26,457 J g^{-1} [15]. For a DSC, a range of materials can be used for calibration. This is due to the wide range of applications and also to the wide range of temperatures used. Generally, one or more metals are used for calibration of both the power and the temperature scale, for example, indium, which has a melting point of 156.6 °C, and an enthalpy of melting of 28.6 J g^{-1}. The uncertainty of the value for the heat of fusion (in this case, of melting) is about 0.5%. This is the limit when calibrating in this manner. Basically, DSC inaccuracy is around a few percent. For up-to-date details on calibration and the calibration materials, please consult [16,17], the references given there, and the most recent data published, since calibration procedures for the DSC are still improving. Specific heat capacity measurements are often performed in three steps. First, the baseline offset of the DSC is measured with empty reference and empty sample sites (one measures the difference between sample and reference baseline, i.e., the systematic asymmetry of the instrument). Then, a measurement of a primary standard with accurately determined specific heat capacity is made; for this, sapphire (crystalline Al_2O_3) is often used. Finally, a measurement of the unknown sample is carried out. By correcting for the baseline offset and division of the measurements of the known and the unknown samples, the specific heat capacity of the unknown sample can be determined. Currently, these corrections are all made by the computer controlling the DSC [18].

79.4.1 Calibration of Chip Calorimeters

Chip calorimeters need also to be calibrated. Like for "conventional" calorimeters, different primary standards can be chosen to cover the entire temperature range of interest. The drawback is that for each primary standard, so for each measured temperature, a different chip has to be used [19] since the sample has to be deposited directly on the chip. Moreover, since the chips used for the calibration are contaminated, calorimetric measurements can be performed only with not calibrated devices. In [20], a calibration method that allows the use of calibrated devices for calorimetric measurements is proposed. First, the chip is inserted into an oven and the heater integrated on the calorimeter chip is calibrated in the temperature range of interest. Then, the obtained calibration curve is used for the calibration of the temperature sensor, that is, the thermopile. In order to determine the accuracy of the calibration, the extrapolated onset temperature (T_e) of primary standards is measured with chips having uncalibrated heater resistance.

79.5 Typical Applications of Calorimeters

79.5.1 Specific Heat, Transition Heat, and Temperatures

Specific heat capacity can be measured by accurate point-wise measurements, but measurement with a DSC is much more efficient. With the DSC, one also obtains the temperature and heat of phase transitions by enthalpy measurement at the transition temperature. Many materials also exhibit changes in crystallization at some given temperature (e.g., glass transition points). Also, these are transition points with their specific heat and temperature, and these can be measured with DSC as well.

79.5.2 Analysis of Chemical and Physical Reactions

Calorimetry is very well suited for analysis of chemical reactions. In particular, the heat of exothermic and endothermic reactions can be determined. An important example is the calorific value of fuels. Households and factories buying fuels are primarily interested in the calorimetric value of their purchases. The so-called oxygen bomb calorimeter is indispensable for primary calibration in this application. The oxygen bomb is usually an adiabatic calorimeter, in which a sample of the fuel (such as solid coal, liquid oil, or gaseous methane) is brought together with an excessive amount of oxygen (e.g., at a pressure of 30 bar). Then, the mixture is ignited, and the heat of burning will spread over the bomb, until equilibrium is reached. The temperature increase of the bomb, divided by the heat capacity of the device, gives the heat of reaction. In general, corrections have to be made for volume and pressure changes (work) as liquids or solids are burned and converted to gases and liquids and also for the additional heat capacity and for the transition heat of the reaction products. Of course, chemical reactions other than burning of fuel can be analyzed in this way as well. In a slightly modified version, the adiabatic calorimeter can also serve to analyze, for example, weak bases and weak acids that do not easily respond to other analysis methods. Similarly, the heat of mixing two solutions, of dissolving a solid in a solvent, and of diluting a mixture and even the heat of wetting can be determined in this so-called solution calorimeter.

79.6 Thermal Analysis of Materials and Their Behavior with Temperature

Apart from directly measuring chemical reactions, one can also analyze materials by exposing them to a temperature sweep, either in an inert atmosphere (nitrogen or helium) or in an oxidizing atmosphere. Then, all effects—such as glass transition, crystallization, melting, evaporation, decomposition, and even

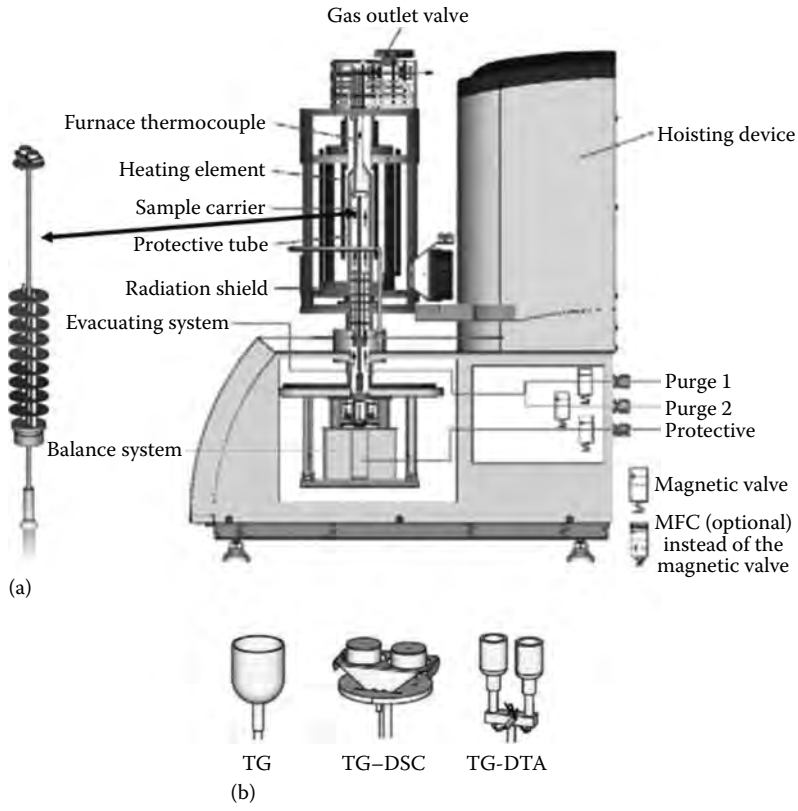


FIGURE 79.5 (a) Schematic diagram of STA 449 Jupiter. In a TG-DSC instrument, heat is transferred to the sample from the oven by the enveloping purge gas, with radiation shields to diminish heat loss by infrared radiation. The entire sample carrier, in (b) showed three different types of sample carriers for the STA 449 Jupiter, is placed on a balance to enable thermogravimetric analysis. (Courtesy of Netzsch GmbH, Selb, Germany.)

oxidation—can be detected. For this, a DSC is again utilized. In the more expensive analysis systems, DSC instruments often offer the possibility of incorporating other techniques. Mass changes due to oxidation and evaporation are detected by thermogravimetry (TG). This is accomplished by placing the entire experimental chamber on a balance that continuously measures the sample mass. These data are available as a function of temperature as well, parallel to the heat data. The DSC inner gas atmosphere is usually refreshed continuously using purge gas. This makes it possible to collect the gases coming from the calorimeter furnace and subject them to further analysis, such as mass spectrometry (MS), gas chromatography (GC), Fourier transform infrared analysis (FTIR), and other analysis methods. With sensitive DSC, it is thus possible to detect almost any structural change in matter as a function of temperature. Figure 79.5 shows a drawing of a combined DSC and TG instrument, capable of analyses up to 2400°C. Radiation shields are required to reduce heat losses by infrared radiation, which is significant at high temperatures.

79.6.1 Biological Analysis from Cells to Entire Human Beings

Calorimetry is not just the measurement of thermal effects in 1 mL samples. Calorimeters receiving entire human beings are available as well. In practice, there is a wide range of application of calorimeters to biology. This is not surprising, since all forms of life produce heat in activity and often also in rest. The

first calorimeters were already used to measure the heat produced by animals. But also, the study of cells (with their heat production of about 1 pW [24]) and the efficiency of enzymatic conversions can be studied using chip calorimetry. Calorimeter chip has also been used for the measurement of thermal conductivity and diffusivity of liquid samples [21–23]; see [24] for an overview and many references. Applications of calorimetry are also found in the food industry, for routine analyses, ecology, plants, etc.

79.6.2 Summary

The applications of thermal analysis are overwhelming in number. In many cases, when wanting to learn more about materials, thermal analysis can add insight. Some applications were mentioned earlier. In Table 79.2, some of the often encountered applications for various disciplines are listed.

TABLE 79.2 Applications of Calorimeters

Area of Interest	Parameter That Can Be Measured with Calorimeters
Material characterization and all other areas listed as follows	Specific heat
	Melting and boiling temperature and behavior
	Fusion and reaction heat and kinetics
	Heat of solution, dilution, mixing, wetting
	Thermal safety
	Glass transition
	Rate and degree of cure
	Crystallization time, temperature, and percentage
	Purity and solid–liquid ratios
	Thermal and oxidative stability
	Identification of multicomponent systems
	Dehydration
Polymers	Effect of the thermomechanical history
	Compatibility tests
	Effects of additives
Pharmaceuticals/cosmetics	Purity and compatibility of active ingredients
	Polymorphism
	Effects of storage and hydrolysis
	Tablet-compression characteristics
	Influence of emulsifiers
	Concentration of medicines in polymers
	Melting and crystallization behavior of waxes
Foods	Melting, solidification behavior, and specific heat of fats and oils
	Polymorphism
	Denaturation of proteins
	Gelatinization of starch
	Freezing–thawing behavior
Biology	Metabolism of cells, organs, animals, and human beings
	Influence of nutrition, toxins, and others on organisms
	Enzymatic efficiency and selectivity
	Concentration of solutions using enzymes or organisms
	Environmental monitoring

79.7 Choosing the Proper Calorimeter for an Application

When faced with a practical analysis problem, one can choose between many instruments. The three main categories are the DSC, the calorimetric bomb or solution calorimeter, and the large reaction or fermentation calorimeter.

79.7.1 What Do You Want to Measure?

If one wants to measure specific heats of reaction, oxidation, solution, etc., at room temperature, a calorimetric bomb or a solution calorimeter might be the first choice. They are accurate and economical. If one wants to optimize a chemical or biological process, reaction calorimeters or fermentation calorimeters might be the choice. In almost all other cases, a DSC will be the most effective instrument. A DSC is especially useful when one wants to obtain the thermal behavior of materials as a function of temperature. In turn, the temperature-dependent behavior of materials can tell a lot about their structure, their properties, and even their thermomechanical history. Important is the scanning range and rate of the DSC. For a sample in the sub-microgram range, chip calorimeters can be used. The chip calorimeters give also new possibilities especially for studying the phase transitions of polymers. Scan rates as fast as 1 kK/s prevent recrystallization of the sample during the scan, and thermal properties of polymers can be studied under the same conditions as during injection molding [25,26]. In some cases, the measurement problem cannot be solved by a standard available calorimeter, and one is forced to either use an experimental calorimeter or develop a special-purpose calorimeter.

79.7.2 Budget and the Need of Pre- and After-Sales Service

A cost-effective DSC will cost about $40,000; the most expensive models can cost up to $150,000 and include thermogravimetric measurement. Calorimetric bombs are somewhat cheaper, starting at $20,000. Finally, one can buy experimental instruments or make one's own system. An adiabatic system consisting of a Dewar in a polystyrene housing, using a commercially available stirrer and platinum resistance (measured with any 5½ digit DMM), will put you in business, although at reduced accuracy. However, the budget for this will not need to exceed much more than about $1000 (excluding the DMM). It is less advisable to make one's own DSC (for cost reasons) or calorimetric bomb (for cost and safety reasons).

79.8 Can the Instrument of Choice Measure the Signals Desired?

Here, various parameters must be considered. Accuracy, the degree to which an instrument reading approaches the true value, lies around a few percent for DSCs and parts of percent for adiabatic calorimeters. Repeatability of the major calorimeter measurement results—such as heat peak area and peak starting temperature—depends, among other things, on such matters as baseline noise/drift, influence of sample preparation, and positioning in the DSC sample site. Resolution should be considered for two parameters: heat and temperature. First, there is the sensitivity of the instrument for generated heat or power, related to its noise. For DSCs, resolution for power is usually around 0.1–10 μW; for bomb calorimeters, resolution for heat is around 0.1–10 J. But there is also the resolution of a DSC for separating two heat pulses at two nearby temperatures. This can be designated as the temperature resolution and depends on time constant and heating rate. Finally, there is the point of linearity and time constant. In case the time constant of the instrument is much larger than that of the process, good linearity is required if one wants to "desmear" the measured curve and obtain the actually produced heat as a function of time. But a current DSC is a very fast instrument, and desmearing is not really necessary anymore. So, linearity is less important for fast DSCs.

79.8.1 Instrument Control and Data Management

Software is becoming more and more important. Since it will control the measurement, its user-friendliness determines how easy and how error-free one can operate the instrument. It also can take care of analysis of the measurement results and graphical and numerical presentation of the results. The software can also take care of quality-control aspects of analysis, such as writing to file all the measurement details (not only what was measured but also how it was measured). Presently, top models of all three types described earlier will perform these functions. With DSC instruments, software can often be used to control all kinds of thermal analysis instruments apart from the DSC, such as TG and DMA, and merge results obtained with DSCs, TG, DMA, and other analysis techniques. This facilitates interpretation of measurements.

TABLE 79.3 Some Commercially Available Calorimeters and Their Specifications

Type	Instrument	Vendor	Cost[a] ($1000)	Range[b] (°C)	Scan Rate[c] (K min^{-1})
Bomb	C 200	IKA	20	25	—
	C 5000	IKA	50	Ambient	—
	C 7000	IKA	46	Ambient	—
Calvet	C 80	Setaram	70	Ambient to 300	2
DSC	Flash DSC	Mettler	75	−95 to 420	2400 000
	DSC 1	Mettler	30	−150 to 700	300
	DSC 200F3	Netzsch	27	−170 to 600	100
	DSC 204F1	Netzsch	40	−180 to 700	200
	DSC 8500	Perkin Elmer	95	−180 to 750	750
	DSC 8000	Perkin Elmer	80	−180 to 750	300
	DSC 6000	Perkin Elmer	45	−180 to 450	100
	DSC 4000	Perkin Elmer	40	−100 to 450	100
	μDSC7evo	Setaram	70	−45 to 120	2
MMC[d]	MMC 274Nexus	Netzsch	55	25 to 500	2

[a] Approximate price of the simplest system without intracooler.
[b] Addition of the widest ranges available.
[c] The highest controlled scan rate (usually for heating).
[d] Multi-module calorimeter.

TABLE 79.4 Companies That Sell Calorimeters

SETARAM Instrumentation, 7, Rue de l'Oratoire, 69300 Caluire, Tel.: +33 (0)4 72 10 25 25, http://www.setaram.com

IKA® Werke GmbH & Co. KG, Janke & Kunkel-Str. 10, D-79219 Staufen, Germany, Tel.: +49 7633 831-0, http://www.ika.net

Mettler-Toledo (Schweiz) GmbH, Im Langacher, Greifensee, CH-8606, Switzerland, Tel.: (41)044 944 45 45, http://www.ch.mt.com

Netzsch Gerätebau GmbH, Wittelsbacherstrasse 42, 95100 Selb, Germany, Tel.: (+49) 9287 881-0, http://www.netzsch-thermal-analysis.com

Parr Instrument Company, 211 53rd St., Moline, IL 61265, Tel.: (309)762-7716, http://www.parrinst.com

Perkin Elmer, 940 Winter Street, Waltham, MA 02451, Tel.: +1 (781) 663-6900, http://www.perkinelmer.com

SII Nano Technology USA Inc., 19865 Nordhoff Street, Northridge, CA 91324, Tel.: +1-818-280-0745, http://www.siint.com

Shimadzu Corp., 1 Nishinokyo-Kuwabaracho, Nakagyo-ku, Kyoto 604-8511, Japan, Tel.: (81)075 823-1111 http://www.shimadzu.co

TA Instruments Inc., 159 Lukens Dr., New Castle, DE 19720, Tel.: (302)427-4005, (302)427-4001 fax, http://www.tainstruments.com

Xensor Integration, 2645 EJ Delfgauw, the Netherlands, Tel.: + 31-15-2578040, http://www.xensor.nl

79.9 Commercially Available Calorimeters

Tables 79.3 and 79.4 describe the different instruments and vendors. Table 79.3 gives an overview of instruments and some characteristics, while Table 79.4 gives vendor information. Tables 79.3 and 79.4 are (necessarily) incomplete since only some of the major vendors have been listed; thus, if looking for a calorimeter, please complete the list with local (and up-to-date) information.

References

1. H.B. Callen, *Thermodynamics and an Introduction to Thermostatistics*, New Delhi, India: Wiley India Pvt. Ltd., 2006.
2. S. Middelhoek and S.A. Audet, *Silicon Sensors*, London, U.K.: Academic Press, 1989.
3. A.W. van Herwaarden, Physical principles of thermal sensors, *Sens. Mater*., 8, 373–387, 1996.
4. W. Hemminger and G.W.H. Höhne, *Calorimetry—Fundamentals and Practice*, Weinheim, Germany: Verlag Chemie, 1984.
5. W. Hemminger, Calorimetric methods. In: V.B.F. Mathot (Ed.), *Calorimetry and Thermal Analysis of Polymers*, Munich, Germany: Hanser Publishers, 1994.
6. K. Allen, F. Hellman, Specific heat of C60 and K3C60 thin films for T = 6–400 K, *Phys. Rev. B* 60, 11765–11772, 1999.
7. K. Allen, F. Hellman, Specific heat of endohedral and higher fullerene thin films, *J. Chem. Phys*. 111, 5291, 1999.
8. A. Minakov, J. Morikawa, T. Hashimoto, H. Huth, C. Schick, Temperature distribution in a thin-film chip utilized for advanced nanocalorimetry, *Meas. Sci. Technol*. 17, 199–207, 2006.
9. A.A. Minakov, S.A. Adamovsky, C. Schick, Non-adiabatic thin-film (chip) nanocalorimetry, *Thermochim. Acta*, 432, 177–185, 2005.
10. S.L. Lai, G. Ramanath, L.H. Allen, P. Infante, and Z. Ma, High-speed (104 OC/S) scanning microcalorimetry with monolayer sensitivity (J/m^2), *Appl. Phys. Lett*., 67, 1229–1231, 1995.
11. S. Zhang, Y. Rabin, Y. Yang, and M. Asheghi, Nanoscale calorimetry using a suspended bridge configuration, *J. Microelectromech. Systems*, 16, 861–871, 2007.
12. A.W. van Herwaarden, P.M. Sarro, J.W. Gardner, and P. Bataillard, Liquid and gas micro-calorimeters for (bio)chemical measurements, *Sens. Actuat.*, A43, 24–30, 1994.
13. E. Van De Kerkhof, Thermoanalytical instrument. Pub. No.: US 2009/0310644 A1, Appl. No.: 12/466,068, Filed: May 14, 2009. Greifensee, CH: Mettler-Toledo AG, December 17, 2009.
14. S. Gaisford, Fast-scan differential scanning calorimetry, *Eur. Pharm. Rev*., 13(4), 2008.
15. K.N. Marsh (Ed.), *Recommended Reference Materials for Realization of Physicochemical Properties*, Oxford, U.K.: Blackwell Scientific, 1987.
16. E. Gmelin and St.M. Sarge, Calibration of differential scanning calorimeters, *Pure Appl. Chem*., 67, 1789–1800, 1995.
17. G.V. Poel, Vincent B.F. Mathot, High-speed/high performance differential scanning calorimetry (HPer DSC): Temperature calibration in the heating and cooling mode and minimization of thermal lag, *Thermochim. Acta* 446, 41–54, 2006.
18. T.M.V.R. de Barros, R.C. Santos, A.C. Fernandes, and M.E. Minas da Piedade, Accuracy and precision of heat capacity measurements using a heat flux differential scanning calorimeter, *Thermochim. Acta*, 269/272, 51–60, 1995.
19. W. Winter, G.W.H. Höhne, Chip-calorimeter for small samples, *Thermochim. Acta*, 403, 43–53, 2003.
20. E. Iervolino, A.W. van Herwaarden, F.G. van Herwaarden, E. van de Kerkhof, P. van Grinsven, A. Leenaers,V.B.F. Mathot, P.M. Sarro, Temperature calibration and electrical characterization of the differential scanning calorimeter chip UFS1 for the Mettler-Toledo Flash DSC 1, *Themochim. Acta*, 512(1), 55–58, 2011.

21. E. Iervolino, A.W. van Herwaarden, P.M. Sarro, Calorimeter chip calibration for thermal characterization of liquid samples, *Thermochim. Acta* 492, 95–100, 2009.
22. T. Adrega, E. Iervolino, A. W. van Herwaarden, P.M. Sarro, Dual-inlet microfluidic calorimeter for thermal characterization of bio-chemical solutions, *Proceedings of the Eurosensors XXIV*, September 5–8, 2010, Linz, Austria.
23. Y. Zhang, S. Tadigadapa, Thermal characterization of liquids and polymer films using a microcalorimeter, *Appl. Phys. Lett*. 86, 034101/1–034101/3, 2005.
24. P. Bataillard, Calorimetric sensing in bioanalytical chemistry: Principles, applications and trends, *Trends Anal. Chem*., 12, 387–394, 1993.
25. V. Brucato, F.G. Crippa, S. Piccarolo, G. Titomanlio, Crystallization of polymer melts under fast cooling I: Nucleated polyamide 6, *Polym. Eng. Sci*., 31(19), 1411–1416, 1991.
26. V. Brucato, F. De Santis, A. Giannattasio, G. Lamberti, G. Titomanlio, Crystallization during fast cooling experiments, a novel apparatus for real time monitoring, *Macromol. Symp*., 185, 181–196, 2002.

Further Information

Much is being published on calorimetry each year. Some very good books to obtain a basic understanding of calorimetry are

Brown, M., *Introduction to Thermal Analysis*, 2nd edn., Dordrecht, the Netherlands: Kluwer, 2001.

Brown, M., Gallagher, P., Handbook of Thermal Analysis and Calorimetry. Vol. 5. Recent Advances, Techniques and Applications, Amsterdam, the Netherlands: Elsevier, 2008.

Haines, P. J., *Principles of Thermal Analysis and Calorimetry*, London, U.K.: Royal Society of Chemistry, 2002.

Hill, J.O. (Ed.), *For Better Thermal Analysis and Calorimetry*, 3rd edn., ICTAC, Vancouver, Canada, 1991.

On DSC, an up-to-date book is: Höhne, G.W.H., Hemminger, W., and Flammersheim, H.-J., *Differential Scanning Calorimetry*, 2nd edn. Berlin, Germany: Springer-Verlag, 2003.

Mathot, V.B.F. (Ed.), *Calorimetry and Thermal Analysis of Polymers*, Munich, Germany: Hanser Publishers, 1994.

Wunderlich, B., *Thermal Analysis of Polymeric Materials*, Berlin, Germany: Springer, 2005.

Apart from these, many relevant papers appear in: *Thermochimica Acta*, Elsevier Science Publishers; *Journal of Thermal Analysis and Calorimetry*, Springer.

Leafing through the latest volumes will bring one up to date on calorimetry research. Regular conferences on calorimetry and thermal analysis are being held, organized by local and global institutes such as ICTAC (International Confederation on Thermal Analysis and Calorimetry), ESTAC (European Symposium on Thermal Analysis and Calorimetry), and NATAS (North American Thermal Analysis Society). Proceedings of these conferences are often published in *Thermochimica Acta* or the *Journal of Thermal Analysis*.

Very interesting is the publication of ICTA from 1991, which contains significant information on nomenclature, literature, suppliers of instrumentation, etc.

VIII

Radiation

80 **Radioactivity Measurement** *Bert M. Coursey*..... **80**-1
Radioactivity • Particle Emission Radioactivity • Electromagnetic Emission Radioactivity • Interactions with Matter • Neutron Interactions • Radioactivity Measurements • References • Further Readings • Partial List of Manufacturers and Suppliers

81 **Radioactivity Detectors** *Larry A. Franks, Ralph B. James, and Larry S. Darken* **81**-1
Gaseous Detectors • Germanium Detectors • Silicon Detectors • Room-Temperature Semiconductors • Prices and Availability • References • Further Readings • Partial List of Vendors and Suppliers

82 **Charged-Particle Measurement** *John C. Armitage, Madhu S. Dixit, Jacques Dubeau, Hans Mes, and F. Gerald Oakham*..... **82**-1
Introduction • Nuclear Emulsions • Gas-Filled Charged-Particle Counters • Scintillation Counters • Solid-State Detectors • References

83 **Neutron Measurement** *Steven M. Grimes* **83**-1
Introduction • Detector Types • Efficiency Calculations and Measurements • Summary • Defining Terms • References

84 **Dosimetry Measurement** *Brian L. Justus, Mark A. Miller, and Alan L. Huston*..... **84**-1
Radiation Dosimetry Quantities and Units • Thermoluminescence Dosimetry • Ionization Chamber Dosimeters • Film Dosimetry • Track-Etch Dosimetry • Bubble Dosimetry • Electronic Personal Dosimeters • Defining Terms • References • Further Information



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

80

Radioactivity Measurement

80.1 Radioactivity **80**-1
80.2 Particle Emission Radioactivity **80**-3
Alpha Particles • Beta Particles • Neutrons • Protons • Fission Fragments
80.3 Electromagnetic Emission Radioactivity **80**-7
Gamma Rays • X-Rays • Conversion Electrons
80.4 Interactions with Matter **80**-8
Annihilation Radiation • Bremsstrahlung
80.5 Neutron Interactions **80**-13
80.6 Radioactivity Measurements **80**-14
References **80**-17
Further Readings **80**-18
Partial List of Manufacturers and Suppliers **80**-18

Bert M. Coursey
National Institute of Standards and Technology

80.1 Radioactivity

Radioactivity is the phenomenon of emissions of neutral or charged particles or electromagnetic radiations from unstable atomic nuclei. The more common types of radiations are listed in Table 80.1. Several naturally occurring (primordial) radionuclides and some produced by cosmic rays in the Earth's atmosphere are given in Table 80.2. Methods of producing other unstable nuclei are discussed as follows. Radioactive isotopes from man-made sources and from naturally occurring nuclides are widely used in health sciences, industry, and academic research.

Radioactivity was first discovered by Henri Becquerel in 1896. His student Marie Curie and her husband Pierre Curie were the first to chemically separate the radioactive elements polonium and radium. Becquerel and the Curies received the Nobel Prize in physics in 1903 for their pioneering work in nuclear and radiochemistry. The historical development of the field of radioactivity is covered in a number of standard texts, including *The Atomic Nucleus* [1] and *Radioactivity and Its Measurement* [2] and in *Radioactivity Measurements: Principles and Practice* [3]. Other excellent texts are available for students in nuclear medicine [4], in radiochemistry and nuclear chemistry [5,6], in nuclear and particle physics [7,8], and a popular general textbook *Radiation Detection and Measurement* by Knoll [9] for nuclear engineering and health physics.

Radioactivity is defined by the International Commission on Radiation Units and Measurements [10] in terms of the *activity*, A, of an amount of a radionuclide in a particular energy state at a given time. Mathematically, it is defined as the quotient of dN by dt, where dN is the number of spontaneous nuclear transformations from that energy state in the time interval dt; thus,

$$A = \frac{dN}{dt} \tag{80.1}$$

TABLE 80.1 Characteristics of Nuclear Radiations

Type	Origin	Process	Charge	Mass (MeV)	Spectrum (Energy)
α-Particles	Nucleus	Nuclear decay or reaction	+2	3727.33	Discrete (MeV)
β^--Rays (negatrons)	Nucleus	Nuclear decay	−1	0.511	Continuous (keV-MeV)
β^+-Rays (positrons)	Nuclear	Nuclear decay	+1	0.511	Continuous (keV-MeV)
γ-Rays	Nucleus	Nuclear de-excitation	0	0	Discrete [keV-MeV]
X-rays	Orbital electrons	Atomic de-excitation	0	0	Discrete (eV-keV)
Internal conversion electrons	Orbital electrons	Nuclear de-excitation	−1	0.511	Discrete (high keV)
Auger electrons	Orbital electrons	Atomic de-excitation	−1	0.511	Discrete (eV-keV)
Neutrons	Nucleus	Nuclear reaction	0	939.57	Continuous or discrete (keV-MeV)
Fission fragments	Nucleus	Fission	≅20	80–160	Continuous (bimodal) 30–150 MeV

Source: Leo, W.R., *Techniques for Nuclear and Particle Physics Experiments*, Springer-Verlag, Berlin, Germany, 1992.

TABLE 80.2 Some Cosmogenic and Naturally Occurring Radionuclides

Radionuclide	Half-Life
^{3}H	12.35 years
^{7}Be	53.3 days
^{14}C	5730 years
^{40}K	1.28×10^{9} years
^{232}Th	1.405×10^{10} years
^{235}U	7.04×10^{8} years
^{238}U	4.47×10^{9} years

The unit of activity in the International System (SI) of units is the becquerel (Bq), which is equal to the unit reciprocal second (s^{-1}). In many fields, the older unit, the curie (Ci), is still in use, where 1 Ci = 3.7×10^{10} Bq (exactly).

The activity of an amount of radionuclide is given by the product of the decay constant, λ, and the number of nuclei present at time t; thus,

$$A = \lambda N \tag{80.2}$$

The reciprocal of λ is the mean life of the nuclide. More commonly, one computes the time necessary for one-half of the nuclei to decay, and since this is an exponential process, the half-life is given by

$$T_{1/2} = \frac{\text{In2}}{\lambda} = \frac{0.69315}{\lambda} \tag{80.3}$$

The activity at any time t can be computed using the initial activity A_0 and the decay time t according to

$$A = A_0 \exp\left(-0.69315 \frac{t}{T_{1/2}}\right) \tag{80.4}$$

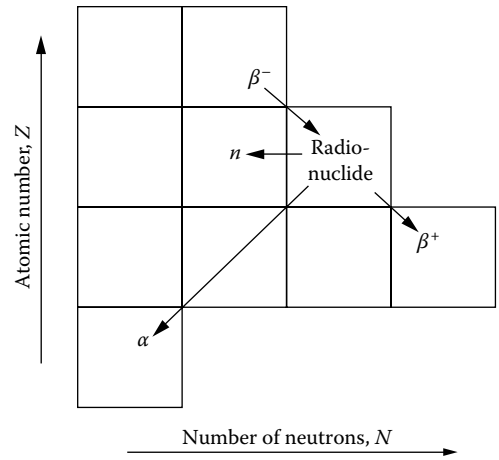


FIGURE 80.1 Characteristics of nuclear radiations. (From Leo, W.R., *Techniques for Nuclear and Particle Physics Experiments*, Springer-Verlag, Berlin, Germany, 1992.)

The term *specific activity* is in use in several applied fields, and one should be careful to look at the units in any particular application. It is normally defined in terms of the amount of activity per unit mass of the element. For example, the nominal activity of ^{14}C in modern carbon (which is found in the biosphere) is 0.3 Bq g^{-1}. In nuclear medicine, however, specific activity is widely used to express the amount of activity per unit quantity of the radioactively labeled compound. For example, one might produce ^{11}C in a cyclotron and oxidize it to carbon monoxide CO with a specific activity of 1 MBq ^{11}C μmol^{-1} (CO) (megabecquerel of activity per micromole of substance).

Stable nuclei are those nuclei in their ground state that have a proper number of neutrons and protons to balance the nuclear forces between constituents and thus do not decay spontaneously. Radioactive nuclei may be divided into three categories:

1. Those that have an excess of neutrons over protons
2. Those that are neutron deficient
3. Those that are in excited nuclear states

A schematic showing the most prominent simple transformations between elements of atomic number Z and neutron number N is given in Figure 80.1. The atomic number Z is the number of protons, and the isotopic mass number A is the total number of nucleons (protons plus neutrons). The notation for an isotope—for example, one of krypton—is ${}^{A}_{Z}\mathrm{Kr}$. In this chapter, isotopes will be identified only by the mass number and the element symbol, for example, ^{241}Am.

80.2 Particle Emission Radioactivity

80.2.1 Alpha Particles

Alpha particles are ^{4}He nuclei (two protons plus two neutrons) and are usually emitted by heavy nuclei ($A > 150$). Alpha particles are monoenergetic and usually have energies between 4 and 6 MeV. Table 80.3 lists several alpha emitters, including some at lower and higher energies that can be used for the energy calibration of instruments. Alpha particles carry a +2 charge, and their detection with a variety of instruments is described in detail in Chapter 82. Alpha-particle pulse-height distributions for several radionuclides of interest in the nuclear fuel cycle are shown in Figure 80.2. The alpha-particle pulse-height distribution for an ^{241}Am source obtained with a high-resolution detector is given in Figure 81.16.

TABLE 80.3 Alpha-Particle-Emitting Radionuclides

Radionuclide	Half-Life	Energies (MeV)
^{148}Gd	90 years	3.183
^{241}Am	432.2 years	5.486, 5.443
^{210}Po	138 days	5.305
^{242}Cm	163 days	6.113, 6.070

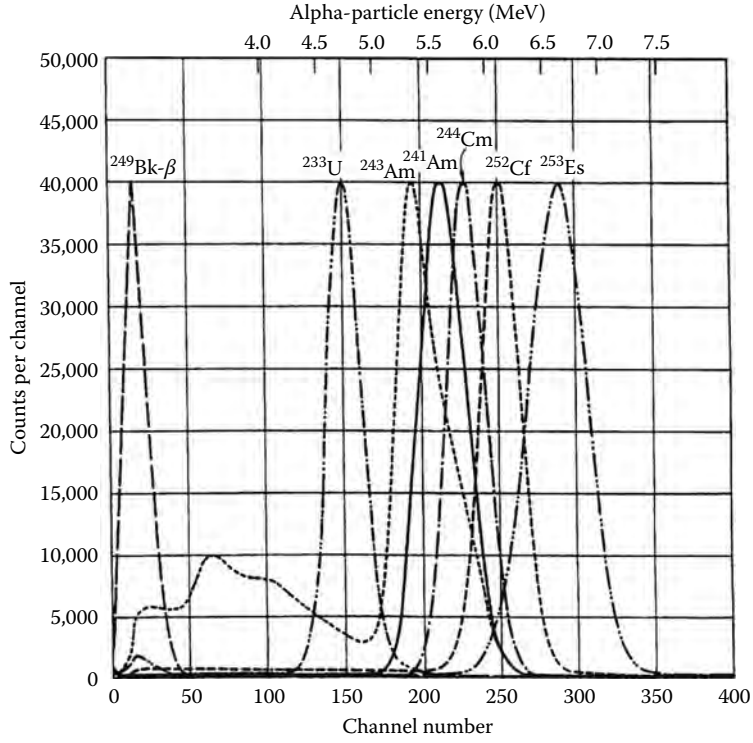


FIGURE 80.2 Liquid-scintillation spectra of selected actinides using a single-phototube detector with the peak channel counts for each major peak preset to 40,000 counts. (From McDowell, W.J., Liquid scintillation counting techniques for the higher actinides, in *Organic Scintillators and Liquid Scintillation Counting*, Horrocks, D.L. and Peng, C.T. (eds.), p. 937, Academic Press, New York, 1971.)

80.2.2 Beta Particles

Beta particles are high-energy electrons emitted by the nucleus as a result of the "weak force" interactions in nuclei that contain too many or too few neutrons. For a system with excess neutrons, the neutron is transformed into a proton, and in the process, an electron and an antineutrino $\overline{\nu}$ are emitted:

$$n \rightarrow p + e^{-} + \overline{\nu} \tag{80.5}$$

The proton remains behind in the nucleus. The antineutrino is required to maintain the conservation of momentum, energy, and "weak-interaction" quantum numbers with respect to the emitted electron [5]. The emitted negatively charged beta particles, usually denoted as β^{-}, have a continuum of energies from zero to a maximum energy denoted by $E_{\beta\max}$. $E_{\beta\max}$ is the total energy available for the kinetic energy of the electron. The antineutrino carries away that portion not transferred to the electron. An example of

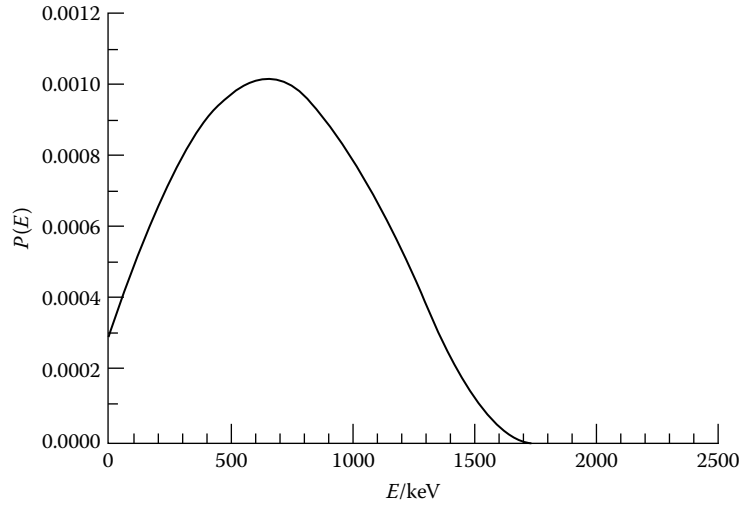


FIGURE 80.3 Theoretical distribution of beta particles as a function of energy for ^{32}P. (From Collé, R., NIST, private communication, 1997.)

an emission spectrum for ^{32}P is given in Figure 80.3. The mode of decay for neutron-deficient nuclei is quite similar, except that the proton is converted to a neutron with the emission of a positive electron or *positron* (β^+) and a neutrino, which may be written as

$$p \rightarrow n + e^+ + \nu \tag{80.6}$$

Although it is unusual, some radioisotopes decay by both negatron β^- and positron β^+ emission. For example, in the decay of the 13 day half-life ^{126}I, one has

$$^{126}\mathrm{I} \rightarrow {}^{126}\mathrm{Te} + \mathrm{e}^+ + \nu \tag{80.7}$$

$$^{126}\mathrm{I} \rightarrow {}^{126}\mathrm{Xe} + \mathrm{e}^- + \bar{\nu} \tag{80.8}$$

Table 80.4 gives a short list of pure-beta-particle emitters.

80.2.3 Neutrons

Neutrons are constituents of the nucleus that have a mass number of unity and carry no charge. In normal radioactive decay processes, neutrons are not emitted from the nucleus. With energies exceeding a few million electron volts (MeVs), however, one can add sufficient energy to cause a neutron to be ejected from the nucleus. These energies can be attained by bombardment with alpha particles from radioactive decay, which is the basis for the common neutron source americium–beryllium (AmBe) in which the alpha particles from the ^{241}Am induce neutron emission from the ^{9}Be nucleus:

$$^{241}\mathrm{Am} \rightarrow {}^{237}\mathrm{Np} + \alpha \tag{80.9}$$

$$\alpha + {}^{9}\mathrm{Be} \rightarrow {}^{13}\mathrm{C}^{*} \text{ (excited state)} \tag{80.10}$$

$$^{13}\mathrm{C}^{*} \rightarrow \begin{cases} {}^{12}\mathrm{C}^{*} + n \\ {}^{8}\mathrm{Be} + \alpha + n \\ 3\alpha + n \end{cases} \tag{80.11}$$

TABLE 80.4 Pure-Beta-Particle-Emitting Radionuclides

Radionuclide	Half-Life	Maximum Energy (keV)	Average Energy (keV)
Negatron (β^-) Emitters			
^{3}H	12.35 years	18.6	5.69
^{14}C	5730.0 years	156.5	49.5
^{32}p	14.29 days	1710	695
^{33}P	25.34 days	249	77
^{35}s	87.44 days	167	49
^{89}Sr	50.5 days	1492	583
^{90}Sr	28.5 years	546	196
^{90}Y	64.0 h	2283	934
Positron (β^+) Emitters[a]			
^{11}C	20.38 min	960	386
^{18}F	109.71 min	633	250
^{22}Na	2.602 years	545	216

[a] These nuclides exhibit a small amount of EC decay.

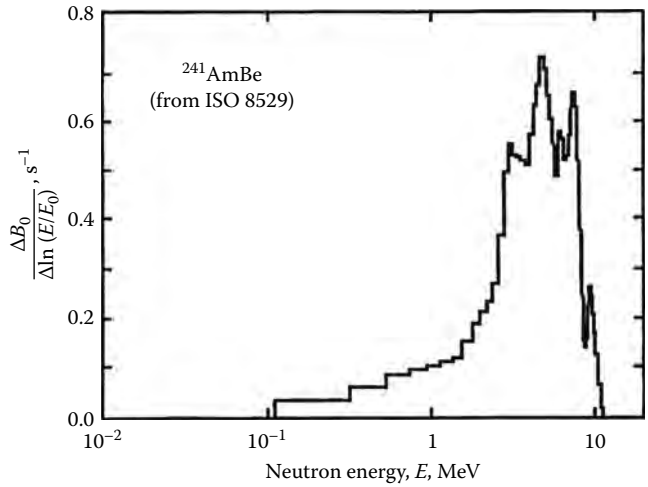


FIGURE 80.4 Neutron spectrum of an ^{241}AmBe radioisotopic source as given in the American National Standard [13] and the International Standard [14]. The units used on the *y*-axis are explained in Ref. [13].

In this case, the ^{13}C nucleus is formed in a highly unstable excited state and rapidly disintegrates by several processes, all of which lead to neutron emission. The neutron spectrum from a ^{241}AmBe source is shown in Figure 80.4. It is very difficult to measure neutron spectra directly because it is necessary to deconvolve a complicated detector response. Figure 80.4 is the "standard" spectral shape recommended in recent national and international standards [13,14]. The units on the *y*-axis are explained in Ref. [13].

Neutrons are also readily produced in high-energy particle accelerators. Various combinations of accelerated particles and target materials are used to produce neutron beams with selected energies (e.g., 2 and 14 MeV). Neutrons from isotopic sources, accelerators, and nuclear reactors can then be used in such applications as boron neutron capture therapy to treat cancer or in metal fatigue analysis. The neutrons can also be used to produce radionuclides for other applications.

80.2.4 Protons

Protons with unit positive charge and unit mass number are simply ^{1}H nuclei stripped of the single orbital electron. Spontaneous emission of protons by nuclei is an exotic decay mode that has been observed in only a few cases involving extremely neutron-deficient nuclei [15]. However, beams of protons can be made in positive-ion, charged particle accelerators by injecting high-purity hydrogen gas, stripping the electrons, and then accelerating the protons to high energies. When these protons are incident on stable target nuclei, a variety of nuclear reactions are possible. The product nuclides are normally radioactive and neutron deficient.

80.2.5 Fission Fragments

Fission fragments are the large-mass debris formed when a high-Z nuclide such as ^{235}U spontaneously (or by absorption of a slow neutron) breaks up into two or more smaller nuclides in a process known as nuclear fission. The best-known fissionable nuclei are ^{235}U and ^{239}Pu. Their fission is a result of capturing neutrons. For a given nucleus, the fission fragments are distributed according to a mass distribution that has two maxima. For ^{235}U fission, the maxima occur at $A = 90$ and 130, resulting in fission-product radionuclides such as ^{90}Sr, ^{95}Zr, ^{131}I, and ^{137}Cs. A small number of high-Z radionuclides decay by spontaneous fission (e.g., ^{248}Cm). This decay process usually competes with alpha-particle decay and is mostly limited to high-Z nuclides.

80.3 Electromagnetic Emission Radioactivity

80.3.1 Gamma Rays

Gamma rays are photons emitted during nuclear de-excitation processes. These gamma-ray transitions may be from a metastable excited state or between levels in a daughter nucleus. Two examples are shown in Figure 80.5 for the 6 h half-life ^{99m}Tc and the 2 day ^{111}In. The large majority of gamma rays from fission products and man-made radionuclides have energies between 20 keV and 2 MeV. A list of gamma-ray emitters used for instrument calibrations is given in Table 80.5. The probability of gamma-ray emission in a particular radionuclide decay (P_γ) is also given in the table. The rate of gamma-ray emission is given by the product $P_\gamma A_0$.

80.3.2 X-Rays

Characteristic x-rays are photons emitted during atomic relaxation processes. X-rays are emitted from radionuclides when electrons are involved in the high-energy processes in the nucleus. In the *electron-capture* (EC) process, for example, the nucleus captures an electron (usually a K-shell electron, since it is closest to the nucleus), and a proton and an electron form a neutron. This process leaves a K-shell vacancy, and a characteristic x-ray from the daughter nucleus can be emitted as orbital electrons from higher shells fill the vacancy. An example is the decay of the 2.6 year half-life ^{55}Fe

$$^{55}\mathrm{Fe} + e^{-} \rightarrow {}^{55}\mathrm{Mn} + \mathrm{X} \tag{80.12}$$

The rate of Mn K x-ray emission from an ^{55}Fe source of activity A_0 is

$$N_{\mathrm{K}} = P_{\mathrm{KX}} A_0 \tag{80.13}$$

where P_{KX} = probability of K x-ray emission.

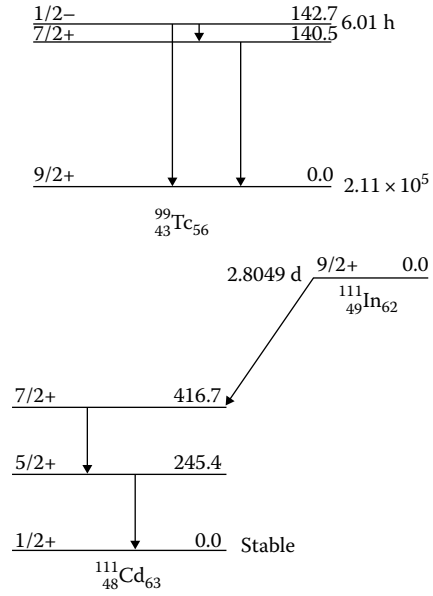


FIGURE 80.5 Simplified nuclear decay schemes for the radionuclides ^{99m}Tc and ^{111}In. The vertical arrows indicate transitions between nuclear excited states. Gamma-ray emission competes with internal conversion processes for each transition. (From Leo, W.R., *Techniques for Nuclear and Particle Physics Experiments*, Springer-Verlag, Berlin, Germany, 1992.)

TABLE 80.5 Gamma-Ray-Emitting Radionuclides Used for Detector Efficiency Calibrations

Radionuclide	Half-Life (Days)	Energy (keV)	Gamma Rays per Decay, P_γ
^{109}Cd	462.6	88	0.0363 ± 0.0002
^{57}Co	271.79	122	0.8593 ± 0.0027
^{139}Ce	137.64	166	0.7987 ± 0.0006
^{203}Hg	46.595	279	0.8148 ± 0.0008
^{113}Sn	115.09	392	0.6489 ± 0.0013
^{85}Sr	64.849	514	0.984 ± 0.004
^{137}Cs	1.102×10^4	662	0.851 ± 0.002
^{88}Y	106.63	898	0.940 ± 0.003
^{60}Co	1925.5	1173	0.9990 ± 0.0002
^{60}Co	1925.5	1332	0.99983 ± 0.00001
^{88}Y	106.63	1836	0.9936 ± 0.0003

80.3.3 Conversion Electrons

Conversion electrons are monoenergetic electrons emitted from the nucleus in isomeric transitions between nuclear levels. Conversion-electron emission competes with gamma-ray emission as a mode of nuclear de-excitation.

80.4 Interactions with Matter

Before examining in detail the manner in which ionizing radiations can interact, consider two sources of photon radiations, annihilation radiation and bremsstrahlung, which are associated with radioactive decay but depend on the material in which the decay process occurs.

80.4.1 Annihilation Radiation

Annihilation radiation is a form of photon radiation associated with a class of radioactive decays. Positrons, since they are the antiparticles of ordinary electrons, cannot survive long in normal matter. Thus, positrons emitted during radioactive decay will slow down in matter until they reach thermal equilibrium. (Depending on the material, they may exist for some microseconds in a state of matter known as positronium.) Ultimately, they combine with an electron in an annihilation event in which their combined mass (1.022 MeV) is converted to energy. This takes the form of two annihilation quanta of 0.511 MeV each, which are oppositely directed (to ensure the conservation of momentum). Depending on the positron energy and the density and thickness of the stopping material, there is a slight but calculable probability of *annihilation in flight* [1]. These decays do not result in characteristic 0.511 keV quanta.

80.4.2 Bremsstrahlung

Bremsstrahlung is the photon radiation emitted by the deceleration of an electron in the Coulomb field of an atom. Thus, bremsstrahlung radiation is present during all beta decay processes as the emitted β-particles (both negatrons and positrons) slow down in matter; it has a continuum of photon energies extending up to the maximum beta-particle energy. The shape of this continuum depends on the nature of the stopping material, as is described in the following section. Bremsstrahlung radiation represents a particular safety hazard for high-energy pure-beta-particle emitters such as ^{32}P (14.4 day, 1.710 MeV β^- max). At higher energies, bremsstrahlung production is almost linear in electron energy and proportional to Z^2 of the stopping material. Thus, to minimize bremsstrahlung, one uses low-Z shielding materials such as plastic.

Now, consider the more general ways in which ionizing radiations from radioactive decay interact with matter. They interact by collisions with the electrons and nuclei along their path. The nature of these interactions is characterized by the charge, mass, and energy of the incident ionizing photon or particle, as well as the charge and mass of the traversed matter. In the energy region of interest, electromagnetic radiation (gamma-ray and x-ray photons) interacts with matter principally by three processes: photoelectric absorption, Compton scattering, and pair production. In the photoelectric process, which dominates at lower energies, the photon transfers its energy to an atomic electron, which is then ejected from the atom with an energy equal to that of the incident photon minus the electron's binding energy. In Compton scattering (incoherent scattering), the photon loses a fraction of its energy to an atomic electron, and the scattered photon emerges generally in a direction different from that of the incident photon. Higher-energy photons will undergo multiple Compton scatter events until the process is finally terminated by a photoelectric absorption. For photons with energies exceeding 1.022 MeV, the process of pair production can occur whereby an electron–positron pair is formed. The positron produced will ultimately annihilate with the production of two 0.511 MeV photons (or in flight). The three interaction processes compete as a function of photon energy, electron density, and nuclear charge of the stopping material. Above 2.044 MeV, triplet production can occur [16], but this is a low-probability process, and few radionuclides emit gamma rays above 2 MeV.

Quantitative measures of the photon interactions in matter are attenuation coefficients based on cross sections for specific interactions [8,9]. The total narrow-beam attenuation coefficient μ is given by the sum

$$\mu = \mu_{\text{photoelectric}} + \mu_{\text{Compton}} + \mu_{\text{pair production}}$$

The attenuation process for a beam of photons traversing a slab of matter is an exponential function of the form

$$I = I_0 \exp(-\mu l) \tag{80.14}$$

where

I and I_0 are the intensities of the transmitted beam and the incident beam, respectively
l is the distance traveled in matter
μ is the linear attenuation coefficient

A useful procedure is to express distances in terms of the mass thickness—the product of density ρ and thickness l. The beam transmission equation can then be rewritten as

$$I = I_0 \exp\left[\frac{-\mu}{\rho}\right](\rho l) \tag{80.15}$$

where μ/ρ is the *mass attenuation coefficient*. The mass attenuation coefficient, in contrast to the linear attenuation coefficient, does not depend on the density of the absorber, but only on its composition. The mass attenuation coefficient as a function of energy for water and lead is shown in Figures 80.6 and 80.7. This discussion of photon interactions has been limited to narrow-beam (or ideal) geometries.

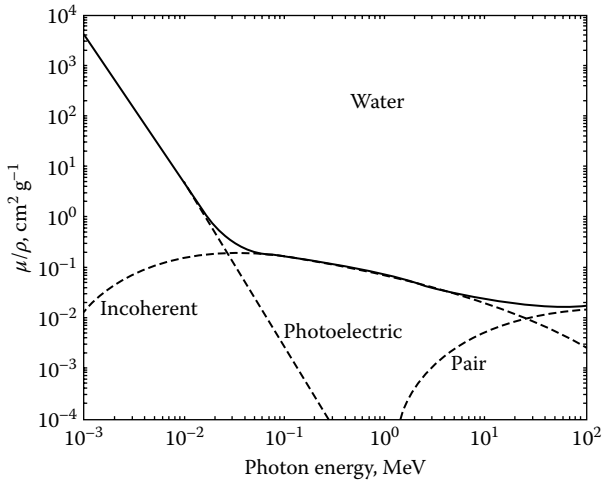


FIGURE 80.6 Narrow-beam mass attenuation coefficients for water as a function of photon energy. (From Seltzer, S.M., *Radiat. Res*., 136, 147, 1993.)

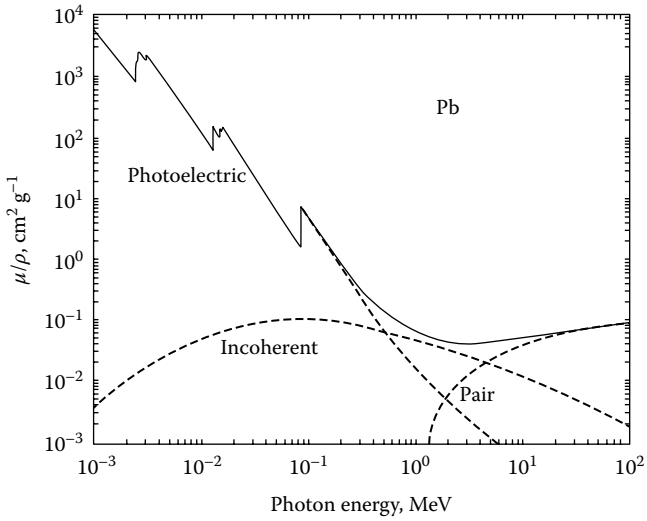


FIGURE 80.7 Narrow-beam mass attenuation coefficients for lead as a function of photon energy. (From Seltzer, S.M., *Radiat. Res*., 136, 147, 1993.)

In practice, most radioactive sources emit a broad beam of radiation such that the detector registers not only events from the incident beam but also those scattered through large angles, and these effects must be included in computing detector response [9].

Charged particles interact with matter in many different ways. These include

1. Inelastic collisions involving ionization and excitation of atomic electrons of the material
2. Inelastic collisions involving bremsstrahlung production in the field of the atom
3. Elastic scattering in the field of the atom
4. Nuclear reactions
5. Emission of Cherenkov radiation (photons in the visible and ultraviolet regions)

Electrons and positrons, due to their much smaller mass, behave differently from protons, alpha particles, and low-Z charged particles. Two quantitative parameters used to describe inelastic interactions of charged particles as they slow down in matter are *stopping power* and *range*. The total mass stopping power, S/ρ, is defined by the ICRU [10] as the quotient of dE by ρdl, where dE is the energy lost by a charged particle in traversing a distance dl in the material of density ρ. For energies for which the nuclear interactions can be neglected, the total mass stopping power can be represented as the sum of an electronic (ionization and excitation) term and a radiative (bremsstrahlung) term as follows:

$$\frac{S}{\rho} = \frac{1}{\rho}\left(\frac{dE}{dl}\right)_{\text{elec}} + \frac{1}{\rho}\left(\frac{dE}{dl}\right)_{\text{rad}} \tag{80.16}$$

Figure 80.8 illustrates the contributions of the electronic term and the radiative term to the total mass stopping power for electrons in lead and water over the energy range from 10^{-2}–10^{3} MeV. For alpha particles, the radiative term can be neglected, but the nuclear interactions cannot. The nuclear interactions are pronounced at lower energies, while at higher energies, the inelastic collision process dominates. This is illustrated in Figure 80.9 for alpha-particle interactions in lead and water.

The range of a charged particle of a given energy in a material is an important parameter in designing detectors and shielding materials. In general, a charged particle does not travel in a

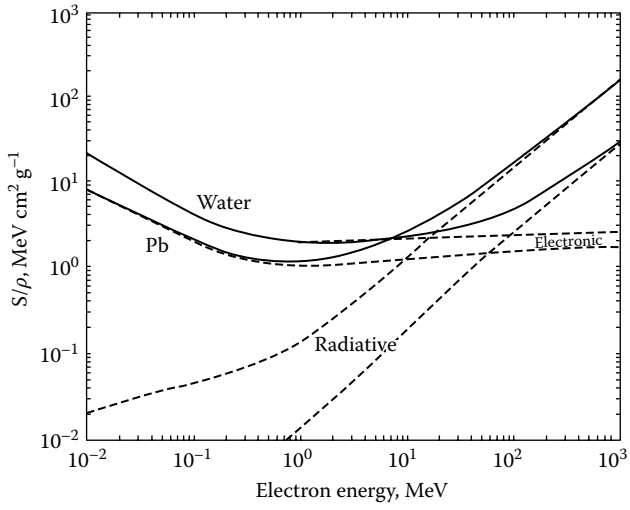


FIGURE 80.8 Stopping power for electrons in water and lead as a function of electron energy. (From ICRU, Stopping powers for electrons and positrons, Report 37, International Commission on Radiation Units and Measurements, ICRU/NCRP Publications, Bethesda, MD, 1984.)

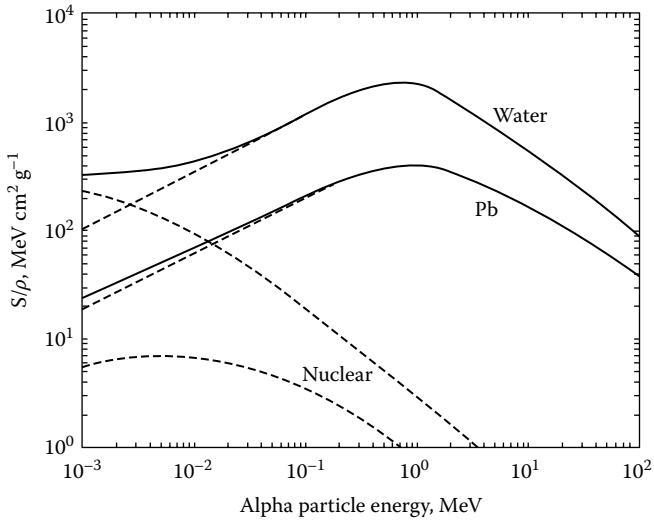


FIGURE 80.9 Stopping power for alpha particles in water and lead as a function of alpha-particle energy. (From ICRU, Stopping powers and ranges for alpha particles and protons, Report 49, International Commission on Radiation Units and Measurements, ICRU/NCRP Publications, Bethesda, MD, 1993.)

straight line as it slows down. However, the total rectified path length traveled by the charged particle, called the mean range, is given by

$$r_0 = \int_0^E \frac{\rho}{S} dE \tag{80.17}$$

where E is the incident electron energy. The range as given in Equation 80.17 has units of g cm^{-2}. Range in water and lead as a function of incident particle energy for electrons and for alpha particles is given in Figures 80.10 and 80.11, respectively.

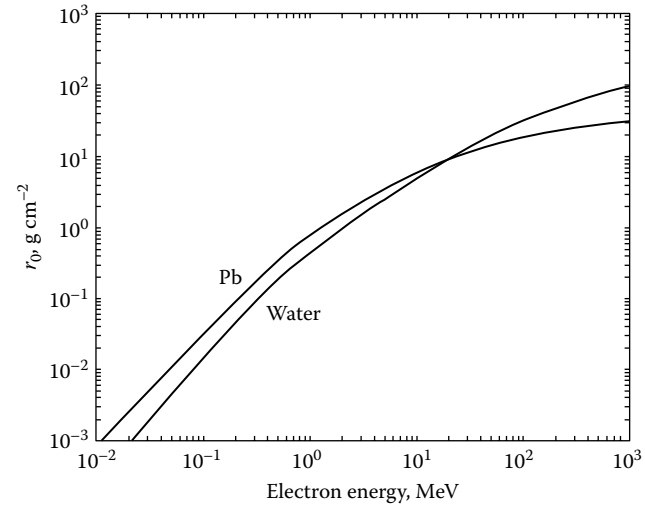


FIGURE 80.10 Range–energy curves for electrons in water and lead. (From ICRU, Stopping powers for electrons and positrons, Report 37, International Commission on Radiation Units and Measurements, ICRU/NCRP Publications, Bethesda, MD, 1984.)

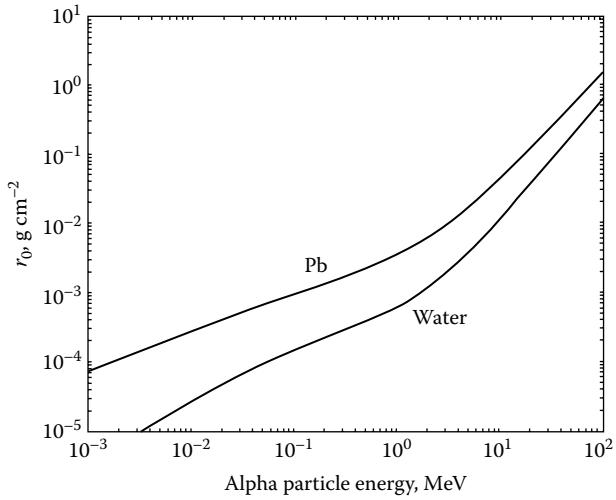


FIGURE 80.11 Range–energy curve for alpha particles in water and lead. (From ICRU, Stopping powers and ranges for alpha particles and protons, Report 49, International Commission on Radiation Units and Measurements, ICRU/NCRP Publications, Bethesda, MD, 1993.)

80.5 Neutron Interactions

Since neutrons are uncharged, they can travel some distance in matter before they interact. This interaction is normally a "nuclear reaction" in which the neutron is absorbed or scattered or produces a nuclear reaction. The probability of each of these interactions occurring is described by the capture, scatter, and nuclear reaction cross sections. Secondary particles produced by neutrons are usually high-energy charged particles that are stopped in short distances in the stopping material. Neutrons from nuclear reactions normally start with energies of a few MeV. They are moderated by collisions with atomic nuclei until they experience a capture process. Capture cross sections for a given elemental stopping material exhibit resonances characteristic of specific nuclear excited states. The total neutron cross section of carbon as a function of neutron energy is shown in Figure 80.12 [20]. The structure in the cross section

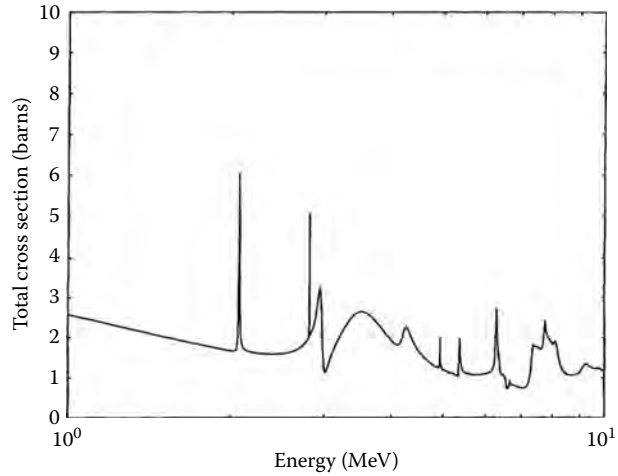


FIGURE 80.12 Total neutron cross section for carbon as a function of neutron energy. The *barn* is the traditional unit of area used to describe neutron cross sections (1 *barn* = 10^{-28}m^2). (From Rose, P.F., ENDF-201, ENDF/B-VI summary documentation, BNL-NCS-17541, 4th edn., Brookhaven National Laboratory, Upton, NY, 1991.)

is the result of resonances in the compound nucleus formed by neutron interaction with ^{12}C. The intensity of a neutron beam in an absorber will decrease exponentially (analogous to a photon beam) with absorber thickness. The attenuation coefficient will include the scatter and capture components.

80.6 Radioactivity Measurements

Measurements of electromagnetic radiations and charged particles are discussed in various chapters, particularly in Chapter 81. However, there are certain classes of instruments that are widely used for the measurement of radioactivity. Three examples will be dealt with here: alpha- and beta-particle liquid-scintillation spectrometers, high-purity germanium (HPGe) gamma-ray spectrometers, and gamma-ray dose calibrators.

Most liquid-scintillation spectrometers operate with two phototubes that view a cylindrical vial containing the radionuclide intimately mixed with an organic liquid scintillator [21,22]. The radiations emitted during radioactive decay interact in the scintillator to produce light pulses that strike the photocathodes of the phototubes. The magnitude of the light pulse depends on the energy deposited by the radiation and on chemical and optical quenching processes [21]. The scintillation yield for high-energy electrons (low *dE/dx*) for the best organic scintillators does not exceed 5%. That is, 5% of the deposited energy is emitted as photons in the near ultraviolet (~400 nm wavelength), and the remaining 95% is lost to radiationless transitions in the organic molecules. In real samples, this photon yield is reduced further through chemical processes (electron scavengers) and optical losses (colored samples, mismatches at vial interfaces, etc.). The counting efficiency for the two phototubes in coincidence is defined as the ratio of the counting rate to the activity. The counting efficiency approaches unity for beta particles with energies exceeding 20 keV [23]. For alpha particles, the efficiency is essentially unity (although the magnitude of the light pulse for alpha particles is less than that for beta particles due to the nature of the interactions of the particles with the organic scintillator). The counting systems are calibrated using standard solutions of known activity [22]. Since the pulse height is largely proportional to the energy of the incident particle, the signal from the counting system can be sorted according to pulse height in a multichannel analyzer (MCA). The pulse-height spectra for a mixture of alpha-particle emitters are shown in Figure 80.13 [24]. In practice, one obtains the area

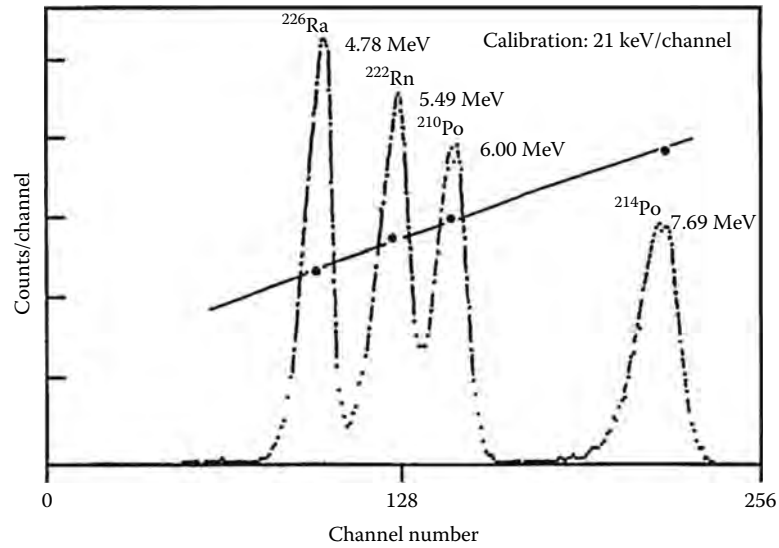


FIGURE 80.13 Liquid-scintillation spectrum of mixture of alpha-particle-emitting radionuclides. (From McDowell, W.J., *Radioact. Radiochem.*, 3(2), 26, 1992.)

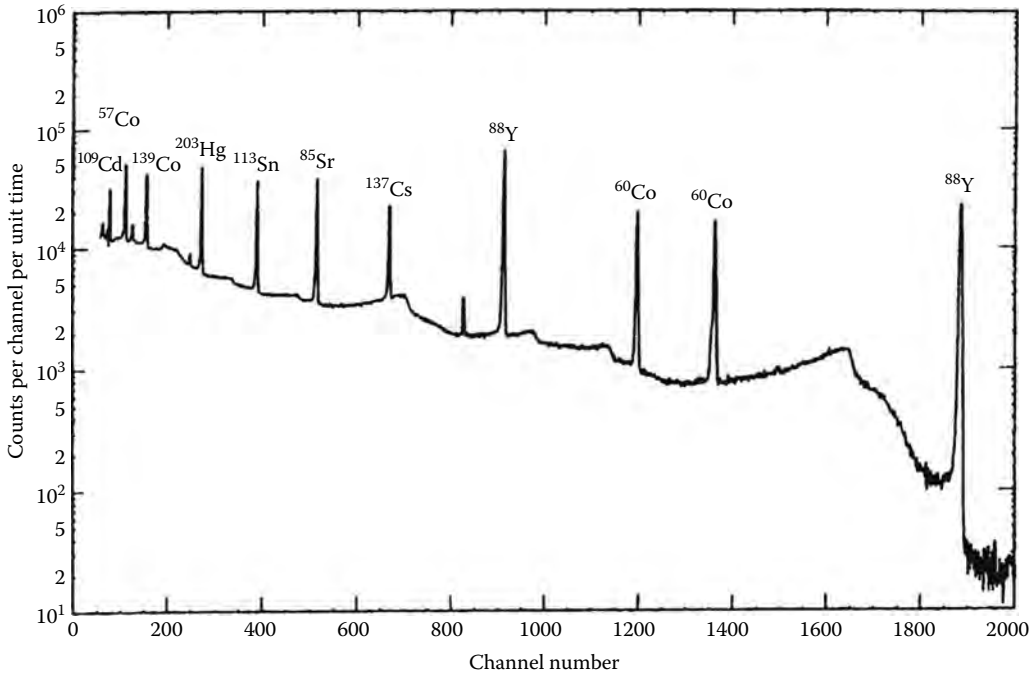


FIGURE 80.14 Gamma-ray spectrum for a solution standard of mixed radionuclides obtained with a Ge(Li) semiconductor detector. (From Mann, W.B. (ed.), *A Handbook of Radioactivity Measurements Procedures*, Report No. 58, 2nd edn., National Council on Radiation Protection and Measurements, NCRP Publications, Bethesda, MD, 1985.)

under the peak for a given radionuclide and divides by time to get the counting rate N_α. The activity A_0 is given by N_α/ε_c, where ε_c is the coincidence counting efficiency.

The activity of mixtures of gamma-ray emitters is needed at many stages in the nuclear fuel cycle. For example, in measurements of environmental radioactivity, a gamma-ray spectrometer allows rapid determination of many fission and activation product nuclides in the presence of naturally occurring radioactivity. Both NaI(Tl) scintillation detectors and HPGe semiconductors are used for this purpose, and their design and operation are covered in Chapter 81. Gamma-ray spectra for a mixture of radionuclides with a semiconductor detector are shown in Figure 80.14 [22]. The detectors are calibrated using radioactivity standard sources prepared by the National Institute of Standards and Technology or one of the commercial secondary standards suppliers. A photopeak efficiency curve for a HPGe detector is shown in Figure 80.15.

Single radionuclides are widely used in many applications in health sciences. In nuclear medicine, most samples are assayed in reentrant dose calibrators, which are essentially ionization chambers. In this counting geometry, the sample is placed inside a well detector that is normally a sealed, pressurized argon gas, ionization chamber (see, e.g., Figure 80.16). The photons emitted from the sample traverse the inner wall of the chamber and cause ionization events in the counting gas. The chambers typically operate at a few atmospheres pressure at a bias voltage of 600–1000 V. The charge created is subsequently collected at the anode. These chambers are designed to assay large amounts of activity such that the current from the device in picoamperes is proportional to the activity in megabecquerels. For a given radionuclide, the manufacturer can establish calibration factors for specific radionuclides in order to convert the current response of the chamber to a unit activity (pA MBq^{-1}). A dose-calibrator response curve for gamma-ray emitters is shown in Figure 80.17 [26].

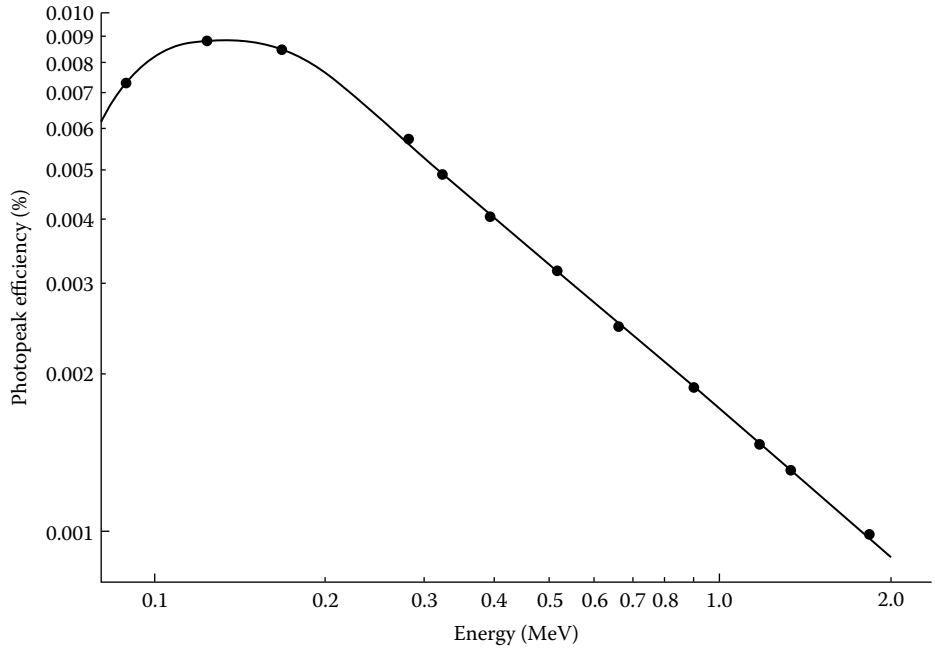


FIGURE 80.15 Photopeak efficiency as a function of energy for a 90 cm^3 HPGe detector for a mixed radionuclide gamma-ray standard of the type listed in Table 80.5. A 5 mL solution source was positioned at 10 cm from the side of a *p*-type semiconductor. (From Cessna, J., NIST, private communication, 1997.)



FIGURE 80.16 A well-type ionization chamber used for the assay of single radionuclides in nuclear medicine. (From Dell, M.A., Capintec, instruments, private communication, 1997.)

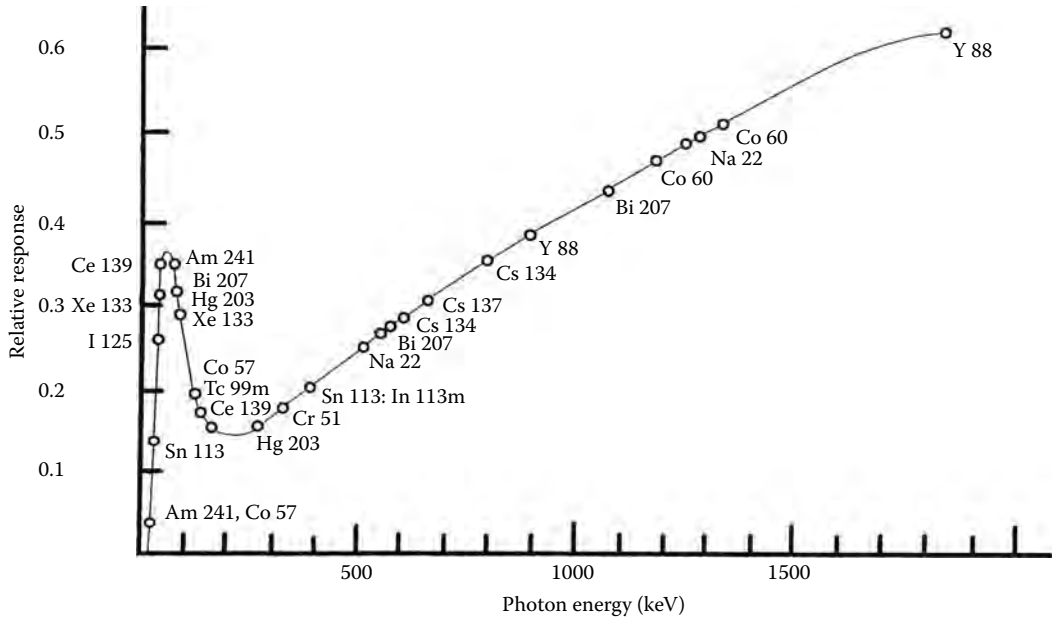


FIGURE 80.17 Photon response as a function of energy for a pressurized reentrant ionization chamber. (From Dell, M.A., Capintec, instruments, private communication, 1997.)

References

1. R.D. Evans, *The Atomic Nucleus*, New York: McGraw-Hill, 1955.
2. W.B. Mann, R.L. Ayres, and S.B. Garfinkel, *Radioactivity and Its Measurement*, Oxford, U.K.: Pergamon Press, 1980.
3. W.B. Mann, A. Rytz, and A. Spernol, *Radioactivity Measurements: Principles and Practice*, Oxford, U.K.: Pergamon Press, 1991.
4. J.C. Harbert, W.C. Eckelman, and R.D. Neumann, *Nuclear Medicine: Diagnosis and Therapy*, New York: Thieme, 1996.
5. G. Friedlander, J.W. Kennedy, E.S. Macias, and J.M. Miller, *Nuclear and Radiochemistry*, New York: Wiley-Interscience, 1981.
6. G. Choppin, J. Rydberg, and J.O. Liljenzin, *Radiochemistry and Nuclear Chemistry*, 2nd edn., Oxford, U.K.: Butterworth-Heinemann, 1995.
7. K.S. Krane, *Introductory Nuclear Physics*, New York: John Wiley & Sons, 1988.
8. W.R. Leo, *Techniques for Nuclear and Particle Physics Experiments*, Berlin, Germany: Springer-Verlag, 1992.
9. G.F. Knoll, *Radiation Detection and Measurement*, 2nd edn., New York: John Wiley & Sons, 1989.
10. ICRU, Radiation quantities and units, Report 33, International Commission on Radiation Units and Measurements, Bethesda, MD: ICRU/NCRP Publications, 1980.
11. W.J. McDowell, Liquid scintillation counting techniques for the higher actinides, in *Organic Scintillators and Liquid Scintillation Counting*, Horrocks, D.L. and Peng, C.T. (eds.), New York: Academic Press, 1971, p. 937.
12. R. Collé (NIST), private communication, 1997.
13. Personnel Performance Criteria for Testing, An American National Standard, HPS N13. 11, Health Physics Society, McLean, VA, 1993.

14. International Standard ISO 8529, Neutron reference radiations for calibrating neutron-measuring devices used for radiation protection purposes and for determining their response as a function of neutron energy, International Organization for Standards, Geneva, Switzerland, 1989.
15. C.N. Davids, P.J. Woods, J.C. Batchelder, C.R. Bingham, D.J. Blumenthal, L.T. Brown, B.C. Busse et al., New proton radioactivities, ^{165,166,167}Ir and ^{171}Au, *Physical Review*, C55, 2255, 1997.
16. J.W. Motz, H.K. Olsen, and H.W. Koch, Pair production by photons, *Reviews of Modern Physics*, 41, 581–639, 1969.
17. S.M. Seltzer, Calculation of photon mass energy-transfer and mass energy-absorption coefficients, *Radiation Research*, 136, 147–170, 1993.
18. ICRU, Stopping powers for electrons and positrons, Report 37, International Commission on Radiation Units and Measurements, Bethesda, MD: ICRU/NCRP Publications, 1984.
19. ICRU, Stopping powers and ranges for alpha particles and protons, Report 49, International Commission on Radiation Units and Measurements, Bethesda, MD: ICRU/NCRP Publications, 1993.
20. P.F. Rose, ENDF-201, ENDF/B-VI summary documentation, Upton, NY: Brookhaven National Laboratory, BNL-NCS-17541, 4th edn., October 1991.
21. D.L. Horrocks, *Applications of Liquid Scintillation Counting*, New York: Academic Press, 1974.
22. W.B. Mann (ed.), *A Handbook of Radioactivity Measurements Procedures*, Report No. 58, 2nd edn., National Council on Radiation Protection and Measurements, Bethesda, MD: NCRP Publications, 1985.
23. B.M. Coursey, W.B. Mann, A. Grau Malonda, E. Garcia-Torano, J.M. Los Arcos, J.A.B. Gibson, and D. Reher, Standardization of carbon-14 by $4\pi\beta$ liquid-scintillation efficiency tracing with hydrogen-3, *International Journal of Applied Radiation and Isotopes*, 37, 403–408, 1986.
24. W.J. McDowell, Photon/electron rejecting alpha liquid scintillation spectrometry, *Radioactivity and Radiochemistry*, 3(2), 26–54, 1992.
25. J. Cessna (NIST), private communication, 1997.
26. M.A. Dell, Capintec, instruments, private communication, 1997.

Further Readings

Frank Herbert Attix, *Introduction to Radiological Physics and Radiation Dosimetry*, Morlenbach, Germany: Wiley-VCH, 2004.

Glenn F. Knoll, *Radiation Detection and Measurement*, 4th edn., New York: Hamilton Printing Company, 2010.

James E. Turner, *Atoms, Radiation and Radiation Protection,* 3rd edn., Weinheim, Germany: Wiley-VCH, 2007.

Partial List of Manufacturers and Suppliers

Global Dosimetry Solutions, 2652 McGaw Ave., Irvine, CA 92614, Phone: 800-251-3331 (toll free), Fax: 949-296-1130, www.mirion.com/index.php?p=dosimetry_div…

Mirion Technologies, 3000 Executive Pkwy., Suite 222, San Ramon, CA 94583, Phone: 925-543-0800, www.mirion.com

MP Biomedical, Inc., 29525 Fountain Pkwy, Solon, OH 44139, Phone: 440-337-1200, 800-848-1163 (toll free), Fax: 714-668-3149, 800-334-6999 (toll free), www.mpbio.com/?country=223

Radiation Detection Company, 8095 Camino Arroyo, Gilroy, CA 95020, Phone: 408-842-2700, Fax: 800-250-3314 (toll free), www.radetco.com/

RAE Systems, Inc., 3775 N. First St., San Jose, CA 95134, Phone: 408-952-8200, 877-723-2878 (toll free), Fax: 408-952-8480, www.raesystems.com

Sierra Radiation Dosimetry Solutions, 2312 Park Ave. # 9b, Tustin, CA 92782, Phone: 949-242-0469, 866-897-8707 (toll free), Fax: 714-249-9001, www.sierradosimetry.com

81

Radioactivity Detectors

81.1 Gaseous Detectors **81**-2
General Operating Principles • *W* Values • Charge Transport • Electric Field Effects • Ionization Chambers • Proportional Counters • Geiger–Müller Counters • Availability • Scintillation Detectors • Scintillation Process • Scintillator Types

81.2 Germanium Detectors **81**-10
Leakage Current • Coaxial Detectors • X-Ray Detection • Current Status of HPGe Detector Technology

81.3 Silicon Detectors **81**-17
Diffused Junction Detector • Surface Barrier Detector • Ion-Implanted Detectors • Energy Resolution • Spatial Resolution • Strip Detectors • Drift Detectors • CCD Detectors • Present Status of Silicon Detector Technology

81.4 Room-Temperature Semiconductors **81**-22
Theory of Operation • Operational Considerations

81.5 Prices and Availability **81**-25

References **81**-25

Further Readings **81**-27

Partial List of Vendors and Suppliers **81**-27

Larry A. Franks
Sandia National Laboratories

Ralph B. James
Sandia National Laboratories

Larry S. Darken
Oxford Instruments, Inc.

Radioactivity was first discovered by Henry Becquerel in 1896 when he noticed that photographic plates became fogged after exposure to uranium atoms. In addition to uranium nuclei, many other naturally occurring and man-made isotopes are known to be radioactive and decay by emitting uncharged particles (gamma rays and neutrons) and charged particles (alpha or beta particles) from their nuclei. Over the past few decades, there has been a growing need for monitoring, locating, and imaging radioactive sources in a wide variety of medical, environmental, industrial, space, scientific, and national security applications. Many of the applications rely on the use of commercially available radiation detectors, whereas others require the development of new detectors to meet system requirements related to sensitivity, power, size, portability, ruggedness, maintenance, radiation hardness, and energy resolution.

The signals generated by a radiation-sensing system depend on (1) the mechanisms in which the incident radiation interacts with the detector material and (2) the scheme used to read out the interaction. In general, there are three types of radiation sensors in use today: gas-filled detectors, scintillation devices, and semiconductor detectors. The types of radiation sensor can also be divided into two groups according to whether they can measure the energy of the emitted gamma rays, x-rays, or charged particles. Whenever an energy-resolving capability is desired, a large number of information carriers must be generated for each absorbed pulse to reduce the statistical limit on energy resolution. This is best accomplished by the use of semiconductor detectors, although some spectral information can be achieved through the use of scintillators and gas-filled detectors.

The typical unit of measure of radioactivity is the becquerel (Bq), which is defined as the rate of one disintegration per second of a decaying isotope. Another common measure of radioactivity is the curie, which can be obtained by multiplying the becquerel by 3.7×10^{10}.

This chapter is divided into sections on gaseous, scintillation, and semiconductor detectors. Separate sections describing semiconductor detectors that must be cooled to cryogenic temperatures for operation and those capable of operation at ambient temperatures are presented. There are several texts devoted to the subject of radiation detection [1–7], and the reader is directed to these books for more detailed information on the principles of detector operation, device performance, problems limiting detector performance, areas of current research and development, and applications.

81.1 Gaseous Detectors

Ionization in a gas resulting from the interaction of x-radiation and gamma radiation is the basis for a wide variety of radiation detectors. The versatility of gas-phase detection arises from the great flexibility in detector sizes and shapes that can be produced, their relatively low production cost, and particularly, the ability to perform a multitude of measurement tasks. Gas detectors can be separated into three distinct types: ionization chambers, Geiger–Müller counters (more commonly known as G-M counters or tubes), and proportional counters, with numerous variations in each type. All utilize the ions and electrons created by the interaction of an incident photon with the detector gas, directly or indirectly, to produce an output signal. They differ in the characteristics of the electric field and nature of the output signal.

Ionization chambers can be operated in either a current mode or pulsed mode, although the current mode is more common. They operate by collecting (with an applied electric field) all the charge generated by the original ionizing event and differ in this way from both proportional and G-M counters. The latter, normally operated in a pulsed mode, have output signals that result from amplification of the original ion pairs by gas multiplication processes.

81.1.1 General Operating Principles

Fundamental to the operation of gas detectors is the generation of electron–ion pairs and their movement through the surrounding gas under the influence of an applied electric field. It is convenient to summarize basic features of these processes before turning to specific detector designs.

81.1.2 *W* Values

The energy (W) required for producing an electron–ion pair in a gas depends on the gas and the type of radiation (and its energy). W values for fast electrons in common filling gases range from 26.4 eV per ion pair in argon to 41.3 eV per ion pair in helium. The presence of nonionizing energy loss processes accounts for the W values greatly in excess of the ionization energy. Fluctuations in the number of pairs produced from photons of the same energy are of significance in pulse-mode operation. The variance is generally less than expected based on Poisson statistics and is accounted for by an empirical constant, the Fano factor (typically 0.1–0.2).

81.1.3 Charge Transport

The motion of free electrons and ions in the gas under the influence of the electric field (E) is quite different. The drift velocity (v) of the more massive ions is a linear function of E/p where p is the gas pressure. It can be expressed as

$$v = \frac{\mu E}{p} \tag{81.1}$$

The proportionally constant μ is the mobility, which depends on the type of gas and the charge of the ion. Values are typically in the region of 1000 $cm^2/V \cdot s/mm$ Hg. The electron mobility is normally about 1000 times ion values; the electron drift velocity is not linear in E/p [8].

81.1.4 Electric Field Effects

The amplitude of a pulse resulting from the interaction of a photon with the wall or fill gas depends strongly on the voltage applied to the detector and serves to distinguish the three detector types. A plot of pulse amplitude as a function of applied voltage is shown in Figure 81.1. The plateau following the initial steep segment is the region of ion chamber operation. It is a region where the electric field is sufficient to reduce recombination of the original pairs to an acceptable value and further voltage increases yield no more charge (as it has all been collected) and ion saturation is established. Assuming complete charge collection, the output current accurately represents the rate of ion pairs that are being produced. This is the basis for ionization chamber operation.

The rapidly rising portion following the plateau marks the onset of gas multiplication (the initial electrons can acquire enough energy between collisions to generate further ionization). In the initial segment, the multiplication process is linear; that is, the charge collected is proportional to the number of original ion pairs. This is the defining characteristic of proportional counter operation. The proportionality eventually is ended at higher voltages by space charge effects caused by positive ions.

At yet higher voltages, the space charge becomes sufficient to reduce the electric field below the multiplication threshold, and no further multiplication takes place. Thus, a condition is reached where the same number of positive ions is produced for all initial ion-pair populations, and the pulse amplitude is independent of the initial conditions. This characterizes G-M counter operation.

Further details on the operation of these devices may be found in texts by Knoll [1], Price [9], Attix and Roesch [10], and Tait [11]. The monograph by Rossi and Staub [2] contains a lengthy discussion of ionization chambers together with data on the physics of the transport process.

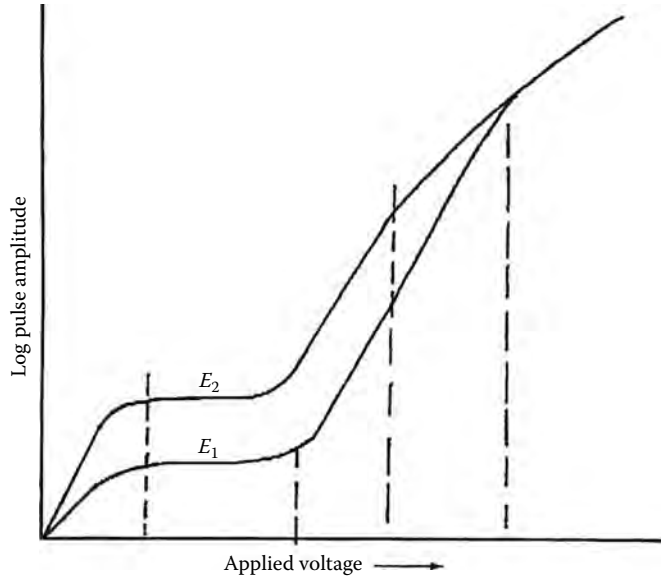


FIGURE 81.1 Distinct operation regions of gas-filled detectors. E_1 and E_2 depict pulses from photons of two energies.

81.1.5 Ionization Chambers

Ionization chambers have been designed for numerous x- and gamma-ray measurement applications. They are frequently employed in radiation survey instruments. They are found in parallel-plate, cylindrical, and spherical geometries. Essential features of their design can be found in the parallel-plate chamber shown schematically in Figure 81.2. The design shown contains an optional guard ring that helps to define the active volume of the chamber. The ring is normally maintained near the collector electrode potential. (Guard rings are also employed in very-low-current designs to reduce leakage current.) Because the current for the ionization chamber is low, typically 1 nA or less, considerable care must be taken with the insulators to minimize leakage currents.

Because of the low output current, special care must be taken with the readout system. Both DC and dynamic-capacitor types are used. Only electrometers of the highest quality should be used. Dynamic-capacitor (or vibrating reed) varieties provide more stable operation. They are normally the choice for very-low-current applications.

Several special-purpose ion chambers are notable. The "free-air" chamber, a parallel-plate variant, is valued for accurate gamma-ray exposure measurements, particularly when absolute measurements are required. This is achieved by collimation of the incoming gamma flux and by an internal design that ensures compensation for ionization produced outside the sensitive volume by secondary electrons. These chambers are limited to energies below about 100 keV, however. Cavity ionization chambers are widely used for dosimetry purpose. To function in this manner, it is necessary that they be designed to meet the requirement of the Bragg–Gray principle [12]. This states that the absorbed dose in a medium can be determined from the ionization produced in a small gas-filled cavity in the medium. The cavity dimension must be small compared to the range of the ionizing particles generated in the material so that the particle flux is essentially unperturbed. Of particular interest for health physics applications are tissue-equivalent chambers where the wall material is a plastic that closely simulates the absorption properties of tissue.

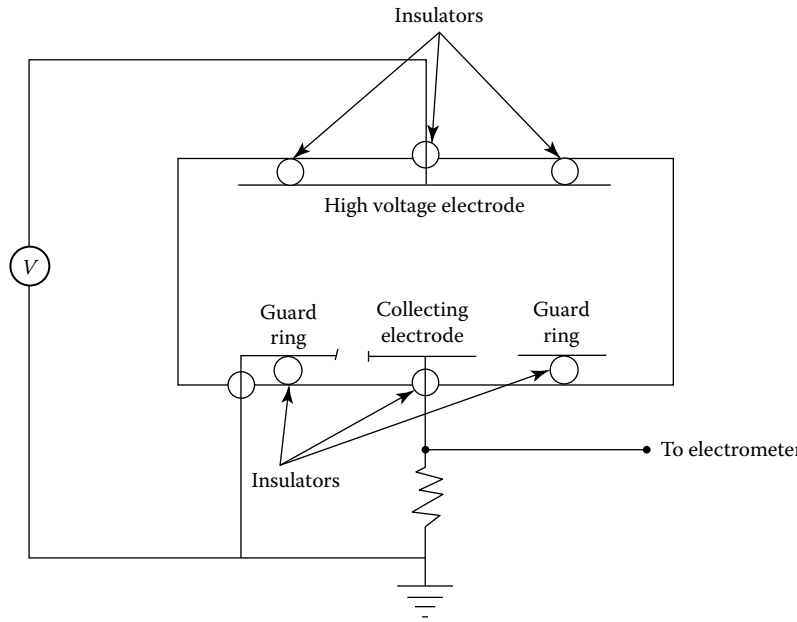


FIGURE 81.2 Schematic diagram of parallel-plate ionization chamber with guard ring defining active volume.

81.1.6 Proportional Counters

Proportional counters utilize gas multiplication to amplify the charge of the initial ion-pair population and have the important characteristic that the charge associated with the amplified pulse is proportional to the ion pairs produced initially. They are normally operated in a pulse mode. They can be operated sealed or in a gas flow mode. The sealed style is most common for x- and gamma-ray applications. While they are found in a number of geometries, cylindrical is the most common shape. A typical cylindrical design is shown in Figure 81.3. This relatively simple design features a central wire that is maintained at high voltage and a surrounding metal container that serves as the cathode. The electric field in this geometry is given by

$$E(r) = \frac{V}{r}\ln\left(\frac{b}{a}\right) \tag{81.2}$$

where

V is the anode–cathode voltage
a is the anode wire radius
b is the inner radius of the cathode

It should be noted that $E(r)$ depends on the diameter of the anode wire, allowing the threshold electric field needed for multiplication to be obtained at relatively modest voltages. The multiplication process is confined to a small volume surrounding the anode where the field exceeds the multiplication threshold. The characteristics of several common counting gases are given in Table 81.1.

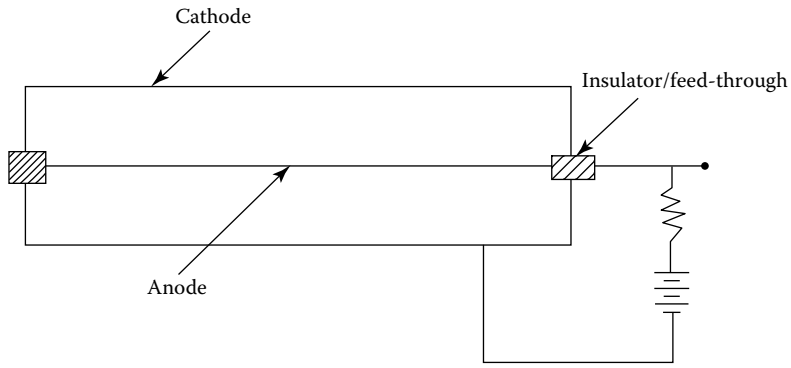


FIGURE 81.3 Schematic diagram of cylindrical proportional counter.

TABLE 81.1 Characteristics of Common Proportional Counter Gases

Gas	W (eV per Ion Pair)	Fano Factor	Resolution (%) at 5.9 keV	
			Calculated	Measured
Ne + 0.5% Ar	25.3	0.05	10.1	11.6
Ar + 0.5% C_2H_2	20.3	0.075	9.8	12.2
Ar + 10% CH_4	26.0		12.8	13.2

Source: Adapted from Knoll, G.F., *Radiation Detection and Measurement*, John Wiley & Sons, New York, 1979.

Proportional counters can be used for both photon and particle measurements. In photon applications, they are particularly valued for spectroscopy in the low-energy x-ray region. Their energy resolution and detection efficiency is generally inferior to semiconductor detectors such as lithium-drifted silicon or mercuric iodide, but they offer large surface areas, reliable operation, and low cost. The energy resolution ΔE for given gas mixture for a specific photon energy can be estimated from the statistical limit given by

$$\Delta E = 2.35\left[W\frac{(F+\beta)}{E}\right]^{1/2} \tag{81.3}$$

where

F is the Fano factor
β is the variance factor (typically about 0.5) [1]
E is the energy in eV
W is the energy in eV required to produce an ion pair

Energy resolution for a proportional counter at 5.9 keV can be expected to be on the order of 10% (Table 81.1).

81.1.7 Geiger–Müller Counters

G-M counters remain among the most widely used means of detecting x-radiation and gamma radiation. They have high sensitivity, are rugged, and offer low cost. Also of importance is the large amplitude of the output pulse (several volts, typically), which greatly simplifies the readout. Because of these factors, they are the detectors of choice for a variety of commercial gamma survey instruments. Like proportional counters, they utilize gas amplification to generate an output pulse, although here the output is independent of the initial number of ion pairs. They are counters only and not applicable to spectroscopy. A G-M counter with typical readout is shown schematically in Figure 81.4.

G-M counters typically use a noble gas filling, most frequently argon or helium. The gas pressure is normally in the region of a few tenths of an atmosphere. At this pressure, a typical counter would require about 1000 V.

The process that terminates multiplication in a G-M counter has important counting implications. The positive ions' space charge and resultant subthreshold electric field (for gas multiplication) persist

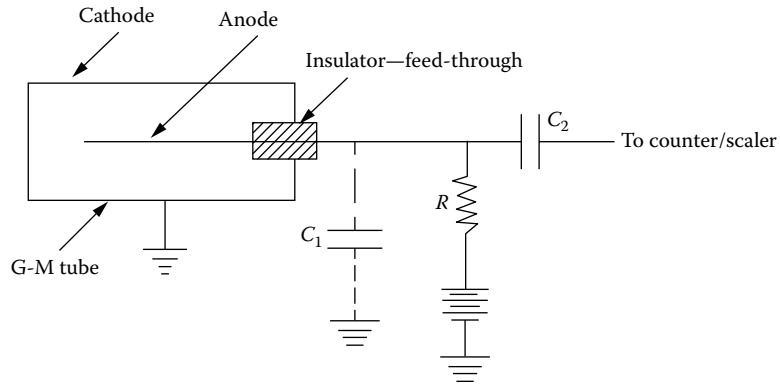


FIGURE 81.4 Typical counting circuit using a G-M tube. C_1 represents the combined tube and wiring capacitance; C_2 is a blocking capacitor that isolates the high voltage from the counter/sealer. The time constant of the circuit is RC_1.

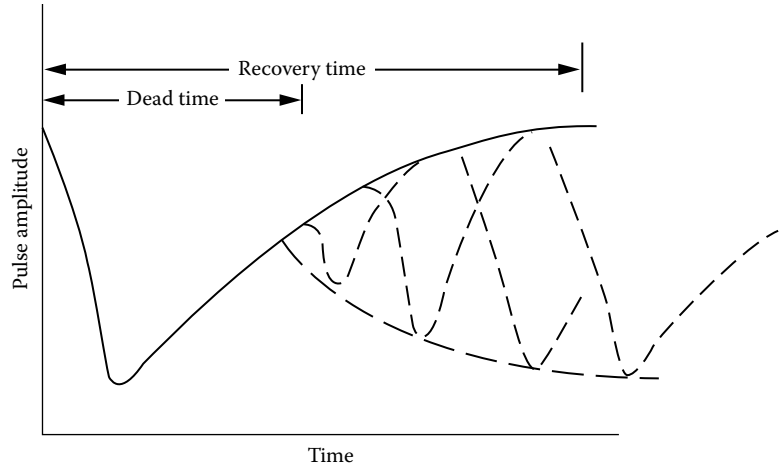


FIGURE 81.5 Illustration of dead time (T_d) in a typical G-M counter.

for some time after the discharge is terminated. As a result, there is an interval following each pulse when pulses from subsequent input gamma rays are not produced or have less than full amplitude. This is illustrated in Figure 81.5. The period between a full-amplitude pulse and the next discharge of any size is the dead time of the G-M counter (T_d). The time from the initial full-amplitude pulse until another full-amplitude pulse produced is the recovery time. Dead times in G-M counters are on the order of 100 μs, with recovery time several times that. If the dead time is known and independent of count rate, the true count rate R_t may be determined in most cases from the observed rate R_o by the expression

$$R_t = \frac{R_o}{1 - R_o R_d} \tag{81.4}$$

81.1.8 Availability

Ionization chambers, proportional counters, and G-M counters are standard commercial products. They are available as stand-alone detectors requiring the user to supply ancillary electronics or as part of an instrument. Custom design devices are also available.

81.1.9 Scintillation Detectors

Scintillators, one of the oldest means of detecting gamma radiation, remain the method of choice for a multitude of counting and spectroscopy applications, particularly when counting efficiency rather than high-energy resolution is the primary objective. They are also widely used to record transient radiation events and in timing applications. While available in solid, liquid, and gas phases, solids and liquids are preferred for gamma-ray applications. Scintillators have the common property of converting energy absorbed from the incident gamma ray into visible or near-visible light. The scintillation detector thus consists of a scintillator element, in which gamma-ray energy is converted to optical photons and some form of a photocell, normally a photomultiplier tube, to convert the optical photons into an electrical signal for processing by ancillary electronics.

The desired properties of a scintillator include high transparency to its own optical emission, efficient conversion of the absorbed gamma energy into optical output, short duration of the output, and output

that is proportional to the amount of energy absorbed. Other useful properties include the ability to be produced in large sizes, ability to be machined, stable output over a wide range of environmental conditions, and efficient gamma-ray absorption.

81.1.10 Scintillation Process

The excess energy contained by the scintillator as a result of the gamma-ray absorption is dissipated largely through nonradiative processes and appears as heat. A small fraction decays radiatively and appears as visible and near-visible light. That fraction of the absorbed energy converted to optical photons is referred to as the scintillation efficiency, which varies from <1% to ≈10% for the more efficient scintillators.

The optical emission originates from electronically excited atomic and molecular states. The emission is broadband, frequently exceeding 50 nm at the half-intensity points. Temporally, it is characterized by a very fast rise, followed by exponential decay. For most purposes, the light pulse can be adequately represented by

$$I(t) \propto N\left[e^{-t/\tau} - e^{-t/\tau_R}\right] \tag{81.5}$$

where

$I(t)$ is intensity at time t
N is related to the total number of scintillation photons
τ_R is the time constant associated with the pulse rise time
τ is the time constant for decay

The time required for the intensity to fall to $1/e$ of the maximum value is the pulse decay time. Optical pulses from organic scintillators are often described by the full width of the pulse at the half maximum of intensity. Further details of the scintillation process can be found in the text by Knoll [1] and, in particular, the treatise by Birks [13].

81.1.11 Scintillator Types

As previously noted, the scintillation process is observed in solid, liquid, and gas phases, but gamma applications are limited mainly to solids and liquids. This brief discussion is confined to the most common solid- and liquid-phase scintillants that are applicable to gamma-ray detection. A number of scintillators, including pure organic crystals and gas scintillators used primarily for charged particles, are not considered. Those remaining ones fall into two groups: inorganic and organic.

81.1.11.1 Inorganics

The most important of this group is the crystals of inorganic salts containing trace quantities of activators to enhance the emission. Because of their high atomic number, high density, and light output, they are valued for gamma spectroscopy. They are available commercially in a variety of standard forms, including right cylinders and parallelepipeds, and are readily machined to custom shapes. Ruggedness varies considerably among crystal types; care in packaging and use is required. Encapsulated crystals are available, but integrated units comprising a scintillator and photomultiplier are also available and often preferable.

The most widely used of this group is sodium iodide activated with thallium, NaI(Tl). NaI(Tl) has high light output and is generally the choice for scintillator-based spectroscopy systems. This material is hydroscopic and must be encapsulated. A typical pulse-height spectrum from a NaI(Tl) photomultiplier combination is shown in Figure 81.6.

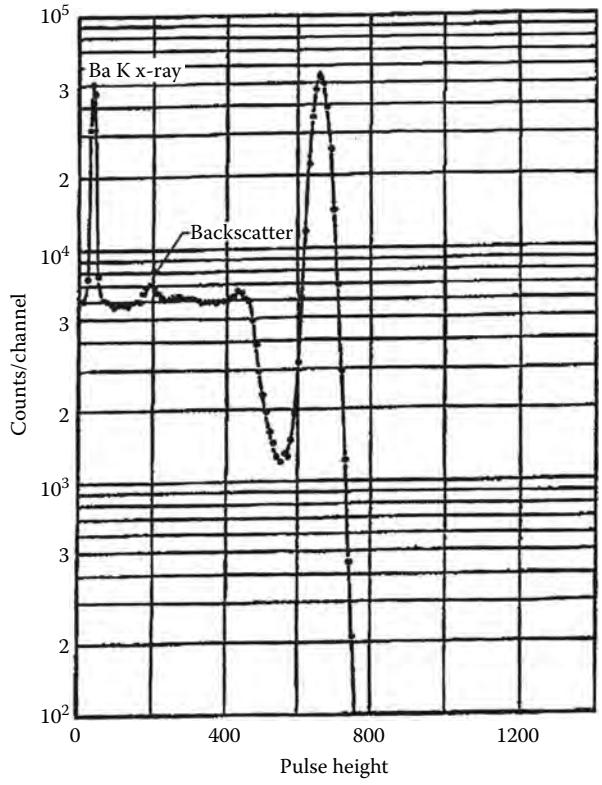


FIGURE 81.6 Typical energy spectrum obtained with a NaI(Tl) scintillator/photomultiplier combination in response to a Cs-137 source (662 keV).

Also useful for spectrographic applications is cesium iodide, which is available with either a thallium or sodium activator. The optical output of CsI(Tl) is at substantially longer wavelengths than either NaI(Tl) or CsI(Na) and is well matched to the spectral response of silicon. This has been successfully exploited to produce small, rugged CsI(Tl)–silicon PIN devices. A spectrum of Cs-137 obtained with a CsI(Tl)/PIN detector combination is shown in Figure 81.7. Exceptional resolution can be obtained when CsI(Tl) is used with a mercuric iodide photocell. With this combination, resolution at 662 keV of better than 5% has been reported [14].

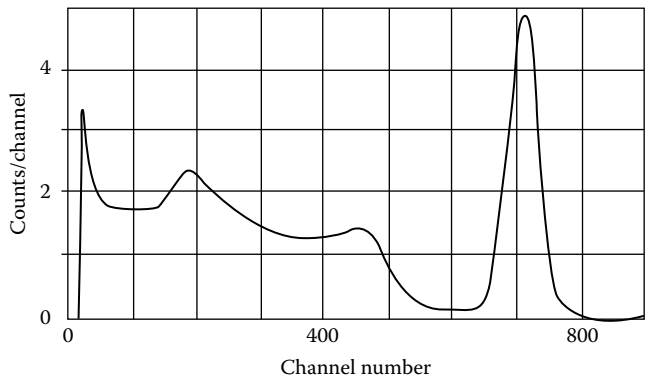


FIGURE 81.7 Energy spectrum of Cs-137 (662 keV) obtained with a CsI(Tl)/PIN combination. Specifications: CsI(Tl) (1 cm × 2 cm × 2 cm) and PIN (1 cm × 2 cm).

TABLE 81.2 Properties of Common Scintillators

Scintillator	Wavelength of Emission Maximum (nm)	Density (g/cm^3)	Principle Decay Constant (ns)
NaI(Tl)	415	3.67	230.0
CsI(Tl)	530	4.51	1000.0
CsI(Na)	430	4.51	630.0
BaF_2	220	4.88	0.6
$Bi_4Ge_3O_{12}$	460	7.13	300.0
BC-505 (liquid)	425	0.88	2.5
BC-400 (plastic)	423	1.032	2.4

Source: Courtesy of Bicron, Newberry, OH.

Two unactivated inorganic crystals, barium fluoride (BaF_2) and bismuth germinate ($Bi_4Ge_3O_{12}$), warrant comment. Barium fluoride has a weak subnanosecond component in the ultraviolet that is useful for the timing and the recording of fast transients. With proper attention to ultraviolet transmission and photocell response requirements, gamma spectra can be obtained. Bismuth germanate has lower light output than either CsI or NaI and inferior energy resolution. However, because of its high density (7.1 g/cm^3) and the high atomic number (83) of bismuth, it is a good choice when high counting efficiency per unit volume rather than energy resolution is required.

81.1.11.2 Organics

Organic scintillators are available in both solid and liquid forms. This group is characterized by very fast rise times (often subnanosecond) and decay times of a few nanoseconds. Their scintillation efficiency is about one-third that of the more efficient inorganics, however. Because of the low atomic number, photoelectric absorption is not significant. As a result, gamma spectra consist mainly of a Compton continuum with, in large samples, a distinct Compton edge. Because of the absence of a photopeak, they are not well suited for spectroscopy applications. (High-*Z* additives such as tin or lead have been shown to enhance the photoelectric response but sometimes at the cost of light output.) They are particularly well suited for counting applications when very large detector volumes are required or for wide bandwidth measurements of fast transients. Because of the large number of available fluorescent organic compounds, organic scintillators are much more easily tailored to meet emission wavelength and decay time requirements than inorganics. The liquids typically consist of one or more fluorescent organic compounds dissolved in an aromatic solvent. A large number of formulations are commercially available. They can be obtained in sealed vessels suitable for mounting on a photomultiplier or as part of integrated scintillator–photomultiplier combination.

Solid organic solutions or plastic scintillator is prepared by dissolving fluorescent compounds in monomers of styrene or vinyl toluene, which then are polymerized. Like liquid organics, several different solutes can be used to achieve the desired wavelength and efficiency. Slabs of plastic scintillator several meters in length can be prepared in this manner. Alternatively, custom shapes can be either cast directly or machined.

The characteristics of several representative scintillators are shown in Table 81.2.

81.2 Germanium Detectors

High-purity germanium (HPGe) detectors are widely used for gamma-ray spectroscopy due to their combination of efficient absorption and high-energy resolution. Figure 81.8 shows the cross sections of germanium and silicon for photoelectric absorption, Compton scattering, and absorption by electron–positron pair production in several materials used for solid-state nuclear radiation detectors. Attenuation is significantly stronger in germanium than in silicon. Over much of the gamma spectrum, the dominant interaction is Compton scattering. However, it is principally the stronger photoelectric absorption in germanium that makes it more

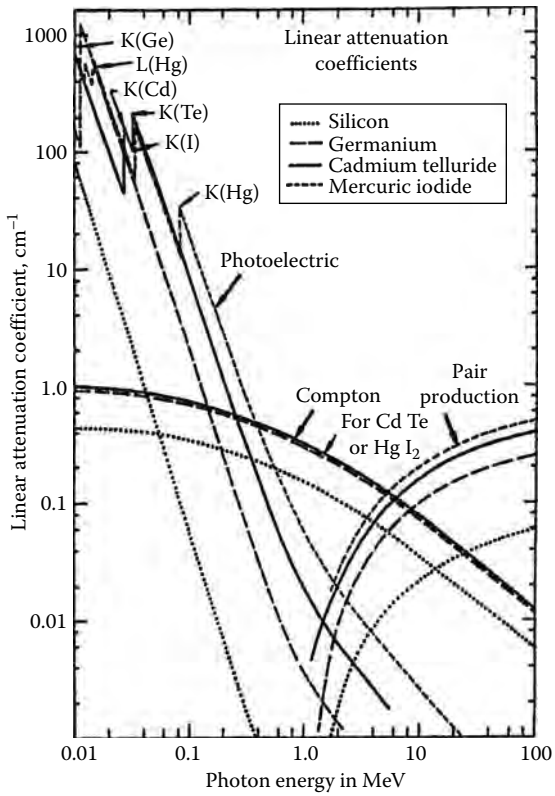


FIGURE 81.8 Attenuation coefficients vs. energy in common semiconductor materials. (From Schlesinger, T. E. and James, R. B. (eds.), *Semiconductors for Room-Temperature Nuclear Detector Applications*, Vol. 43, Semiconductors and Semimetals, Academic Press, San Diego, CA, 1995.)

suitable than silicon for gamma-ray spectroscopy. In the typical-size germanium detector, a gamma ray may be scattered several times before it is photoelectrically absorbed. Thus, the energy of the gamma ray is primarily transmitted directly to a small number of electrons. These energetic electrons in turn interact with electrons in the valence bands to create mobile pairs of electrons and holes. In a detector with sufficiently large volume, the average number of electron–hole pairs N produced by an absorbed gamma ray of energy E becomes independent of the details of the initial reaction path and varies linearly with E as follows:

$$N = \frac{E}{\varepsilon} \tag{81.6}$$

This relationship is more broadly valid and is the foundation of energy spectroscopy of gamma rays using semiconductors, gases, and cryogenic liquids (ε depending on the material). While ε is independent of the gamma-ray energy (and is also virtually the same for energy deposited by charged particles), ε in germanium does increase slightly with decreasing temperature, as does the energy gap. At 77 K, ε is 2.96 eV, and the energy gap is 0.72 eV.

The practical exploitation of Equation 81.6 depends on electronically detecting the motion of the ionized charge in an electric field. The signal-to-noise ratio is improved by reducing current flow in the detector from other mechanisms. In germanium, this is achieved by producing a rectifying and a blocking contact and by cooling to about 100 K. For a planar detector, a slice of HPGe is diffused with the donor lithium on one side, forming a strongly n-type layer. The opposite side is implanted with the

acceptor boron, forming a *p*+ layer. When voltage is properly applied, the electric field direction prevents the majority carriers in the contact regions from being injected across the device. As the voltage is applied, a region depleted of holes will advance into the slice from the *n*+ contact if the slice is *p* type. If the slice is *n* type, a region depleted of electrons will advance from the *p*+ contact. At the depletion voltage V_d, the depletion region reaches the opposite contact. For germanium,

$$V_d = 565\,\text{V} \times \frac{N_e}{10^{10}\,\text{cm}^{-3}} \times \frac{d^2}{\text{cm}^2} \tag{81.7}$$

where

N_e is the net charge density in the depleted or active region of the detector
d is the thickness of this region

This is a key relationship in HPGe technology, as it quantifies the effect of the residual impurity concentration on device size and depletion voltage. Techniques to grow germanium pure enough for gamma detectors were pioneered by Hall [15] and the detector group at Lawrence Berkeley Laboratory [16], based on purification methods of Pfann [17] and crystal-growing techniques developed by Teal and Little [18] to produce crystals for germanium transistors.

81.2.1 Leakage Current

Germanium detectors need to be cooled to reduce leakage current. There are several potential sources of leakage current, including

- Diffusion of minority carriers from either doped contact into the depletion region
- Thermal generation of carriers at either bulk or surface defects in the depletion region
- Electrical breakdown at points where the electric field is concentrated due to irregularities in the contact geometry, large-scale inhomogeneities in the bulk, or surface states

Current will also be generated if the detector is not shielded from room-temperature infrared radiation. Background nuclear radiation from materials near the detector and cosmic radiation also generate leakage current.

Germanium detectors are typically liquid-nitrogen cooled and operated between 85 and 100 K. In this temperature range, leakage current is typically less than 40 pA in "good" detectors and is not a significant contributor to system noise (400–900 eV). Leakage current increases with temperature and eventually becomes the predominant noise component. Pehl et al. [19] reported a leakage current-driven system noise of 2 keV at 150 K and 7 keV at 170 K for an 8 cm^3 planar detector. These authors also reported that, above about 120 K, the leakage current had an activation energy of approximately one-half the bandgap and attributed this to generation at mid-gap surface states. Below 120 K, the temperature dependence was milder.

A typical detector/cryostat configuration is shown in Figure 81.9. The detector resides in an evacuated cryostat and is cooled by means of a copper rod inserted into a liquid-nitrogen dewar. The first stage of amplification is an FET, also cooled, positioned nearby the detector. Mechanical fixturing is designed to stabilize the detector and the mechanisms for contacting it, to provide a cooling path from the detector to liquid nitrogen, and to electrically insulate the high-voltage contact.

A variety of detector geometries are shown in Figure 81.10. These different electrode configurations allow efficiency and energy resolution to be optimized for different gamma-ray energies and applications. For example, the detector in Figure 81.10c minimizes noise by the lower capacitance of its electrode configuration at the expense of the reduced stopping power. Thus, this detector would be more suitable for lower-energy gamma rays.

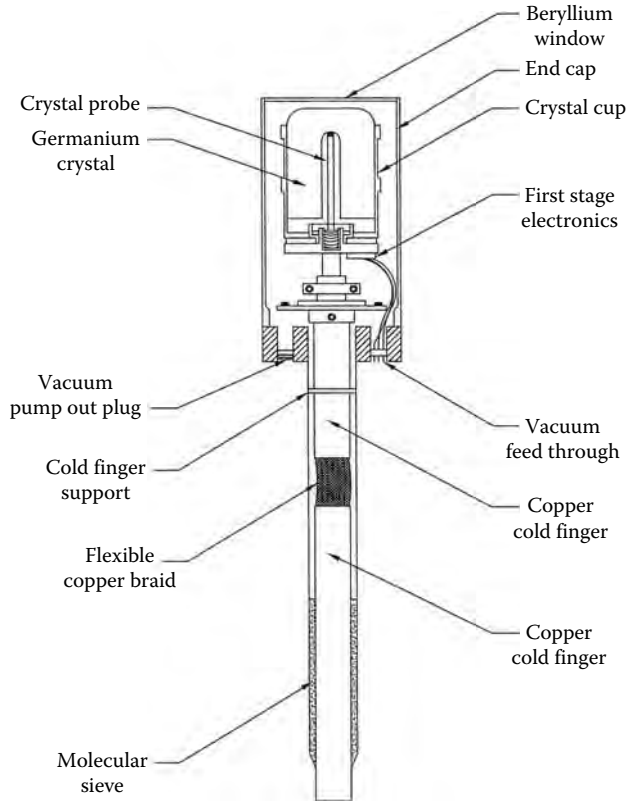


FIGURE 81.9 Schematic cross section of a dipstick cryostat. (From Darken, L. S. and Cox, C. E., in *Semiconductors for Room-Temperature Nuclear Detector Applications*, Academic Press, New York, 1995, reprinted with permission of Oxford Instruments, Inc.)

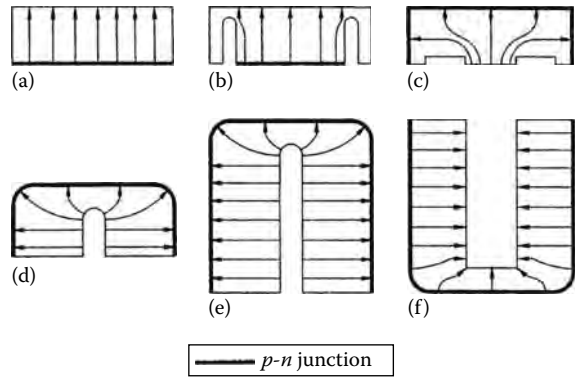


FIGURE 81.10 Schematic cross section and electrostatic field distribution in HPGe detectors. The dark line represents the *p-n* junction: (a) true planar, (b) grooved planar, (c) low-capacity planar, (d) truncated coaxial, (e) closed-end coaxial, and (f) well geometry.

81.2.2 Coaxial Detectors

The detector type shown in Figures 81.9 and 81.10e has a closed-end coaxial geometry. Nearly all of the largest-volume (active volumes of 100–800 cm^3) HPGe detectors are of this type. This electrode geometry reduces both capacitance and depletion voltage with respect to a planar detector of the same volume.

This latter benefit relaxes the constraint on material purity. In addition, charge collection distances are shortened, and the uncontacted surface area, frequently troublesome in processing, is reduced. Also, the HPGe is grown by the Czochralski technique and is therefore nearly cylindrical, even before machining. It is important, however, to note that the reduction in depletion voltage is realized only when the device is contacted so that it depletes from the outer contact to the inner contact. Thus, *p*-type HPGe to be fabricated into a coaxial detector is lithium diffused on the outer diameter, and in the case of *n*-type HPGe, the outer diameter is boron implanted.

The boron-implanted contact (depth approximately 0.2 μm) is thinner than the lithium-diffused contact (depth approximately 750 μm), so the *n*-type coaxial detector can detect lower-energy radiation and is usually built with a beryllium window in the aluminum end cap to take full advantage of this feature. The difference in the range of use is illustrated in Figure 81.11. The geometric asymmetry of the contacting electrodes in the coaxial detector makes charge collection more dependent on the carriers (electrons or holes) traversing to the inner contact. As more gamma rays are absorbed near the outer contact, the carriers traversing to the inner contact must travel on average a longer distance. Also, charge traversal near the inner contact is particularly effective in inducing current in the external circuit [20], Thus, the *p*-type coaxial detector with positive bias on the outer electrode is more sensitive to hole collection, and the *n*-type coaxial detector with negative bias on the outer electrode is more sensitive to electron collection. This is a crucial consideration in applications where the hole collection is going to be degraded during use by exposure to fast neutrons or other damaging radiation. The superior neutron damage resistance of the electrode biasing polarity on *n*-type coaxial detectors was demonstrated by Pehl et al. [21].

A typical gamma-ray spectrum of a Co^{60} source taken with a coaxial HPGe detector is shown in Figure 81.12. The salient features are the full-energy peaks at 1.17 MeV and at 1.33 MeV and the lower-energy plateaus due to incomplete energy absorption of Compton-scattered gamma rays. The peak-to-Compton ratio [22] is generally 40–100, depending on the size and quality of the detector. The 1.33 MeV peak is shown separately in Figure 81.13. The energy resolution measured as the full width at half the peak maximum (FWHM) for typical coaxial germanium detectors is between 1.6 and 2.1 keV for 1.33 MeV gamma rays, again depending on the size and quality of the detector. The variance in the peak L^2 (FWHM = 2.35 × L, L being the standard deviation for a Gaussian distribution) can be divided into three additive components: the electronic noise

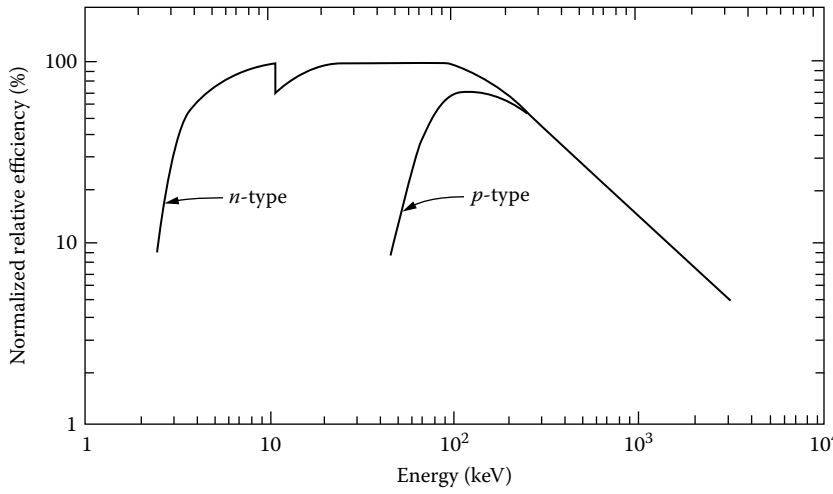


FIGURE 81.11 Relative absorption efficiencies for typical *n*- and *p*-type detectors. (From Darken, L. S. and Cox, C. E., in *Semiconductors for Room-Temperature Nuclear Detector Applications*, Academic Press, New York, 1995, reprinted with permission of Oxford Instruments, Inc.)

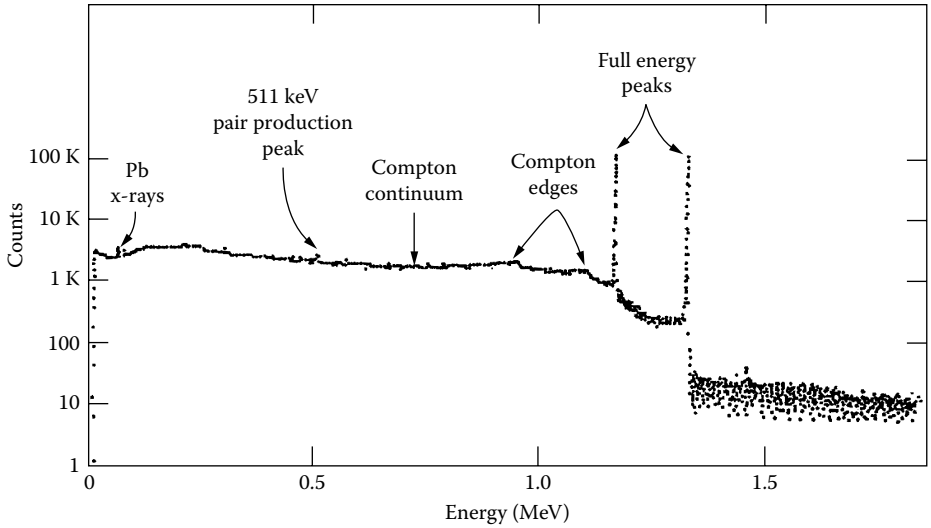


FIGURE 81.12 ^{60}Co spectrum collected with a 15% *p*-type detector showing typical features of germanium detector spectrum.

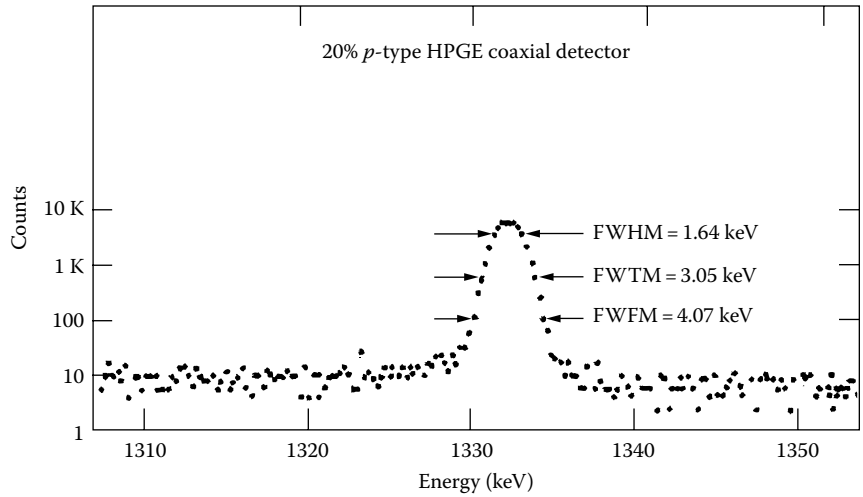


FIGURE 81.13 ^{60}Co spectrum collected with a 22% relative efficiency *p*-type detector. (From Darken, L. S. and Cox, C. E., in *Semiconductors for Room-Temperature Nuclear Detector Applications*, Academic Press, New York, 1995, reprinted with permission of Oxford Instruments, Inc.)

component L_N^2, a component reflecting the variance in the number of electron–hole pairs created L_F^2, and a component due to incomplete charge collection L_T^2:

$$L^2 = L_N^2 + L_F^2 + L_T^2 \tag{81.8}$$

$$L_F^2 = \varepsilon EF \tag{81.9}$$

F is called the Fano factor and has been experimentally determined to be no greater than 0.08 for germanium [23]. $F < 1$ implies that electron–hole pair creation events are not uncorrelated. L_T^2 is usually dominated by the trapping of electrons and holes at defect sites. However, shorter electronic shaping times,

lower electric fields, and larger detectors accentuate ballistic deficit (loss of collected charge in the external electronics due to the finite traversal time of the electrons and holes across the detector). L_N^2 is independent of gamma ray E and is the dominant resolution-limiting factor at low energies. L_F^2 depends linearly on E and, for a coaxial detector, usually dominates L_N^2 for E over a few hundred keV. The energy dependence of L_T^2 is not given simply from first principles for an arbitrary trap distribution, but an E^2 dependence seems to fit under many circumstances. Thus, at high enough E, L_T^2 is expected to be the largest component. For "good" detectors at 1.33 MeV, however, L_T^2 is always smaller than L_F^2. In addition, the magnitude of L_T^2 is variable enough between detectors that it distinguishes between acceptable, very good, and excellent detectors. L_T^2 is usually also the only component of resolution drawn from a non–Gaussian distribution and is thus responsible for any low-energy tailing of the peak.

81.2.3 X-Ray Detection

Both silicon and germanium detectors are used in low-noise systems for the detection of fluorescent x-rays produced by electron beams (usually in an electron microscope) or x-rays (XRF). For both materials, the detector is liquid-nitrogen cooled to reduce leakage current, and small-volume devices (Figure 81.10b, typically 10 mm^2 active area and 3 mm depth) are used to decrease capacitance and therefore to further reduce electronic noise. Lithium-drifted silicon (SiLi) detectors were used first for these applications. Early germanium detectors displayed poor peak shape for x-ray energies just above the L absorption edges (attributed to diffusion against the field to the front contact by some electrons and their resulting loss to the photopeak [24]). However, as was first demonstrated by Cox et al. [25] this is not a fundamental problem but can be solved by the contacting technology. An x-ray spectrum taken with HPGe detector is shown in Figure 81.14. Germanium has the advantages with respect to silicon of a smaller ε (2.96 eV per pair versus 3.96 eV per pair at 77 K) for better energy resolution and a higher Z (32 versus 14) for better photoelectric absorption of higher-energy x-rays.

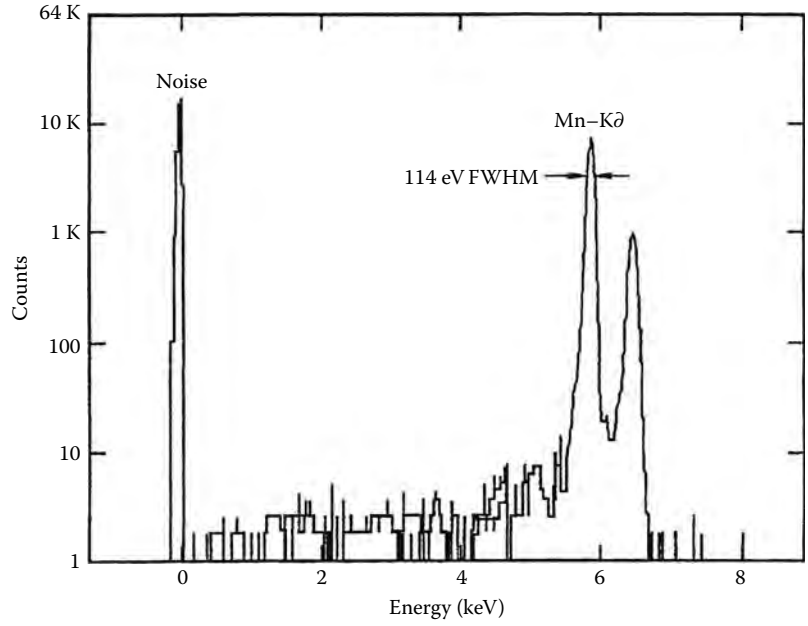


FIGURE 81.14 Manganese x-ray spectrum from ^{55}Fe source collected with an HPGe detector. (From Darken, L. S. and Cox, C. E., in *Semiconductors for Room-Temperature Nuclear Detector Applications*, Academic Press, New York, 1995, reprinted with permission of Oxford Instruments, Inc.)

81.2.4 Current Status of HPGe Detector Technology

HPGe detectors are a mature commercial technology. Process development in crystal growing and diode fabrication has been conducted in private industry where significant advances are proprietary. However, the results of technological advances in these areas are quite evident in the continual improvement in the size, performance, and availability of HPGe detectors. Maximum photopeak efficiency for HPGe gamma-ray detectors is doubling every 6–8 years. Concurrently, energy resolutions are moving toward the theoretical limits of Equation 81.8 as the concentrations of trapping centers are reduced.

The reliability as well as the performance of germanium gamma-ray detectors has also continued to improve, although this is harder to quantify. Cryostats have been redesigned to reduce virtual and direct leaks, reduce microphonics, implement modular design, and improve ruggedness. Detector makers are also making more serious attempts to offer models with reduced backgrounds by judicious design changes and careful selection of materials.

New applications for gamma-ray spectroscopy have emerged. The HPGe detector industry has recently supplied over 100 detectors each to two different experimental facilities (Gammasphere in the United States and Euroball in Europe), where they were arranged spherically in a modular fashion around the target of an ion accelerator to study the decay of nuclei from excited states of high angular momentum.

For users of single-detector systems, developments in the pulse-processing electronics necessary for data acquisition and in the hardware and software for data analysis have resulted in both more compact and more flexible systems. Plug-in cards for a personal computer are available that contain not only the functions of the ADC and multichannel analyzer but the high-voltage power supply and amplifier as well. Software developments also allow for control of many pulse-processing parameters that were previously set manually.

81.3 Silicon Detectors

As with germanium for gamma-ray spectroscopy, the impurity requirements on silicon for nuclear radiation detectors are also stringently low—and difficult to obtain. Such silicon must be grown by the float-zone technique to eliminate contamination from a crucible. Unlike with germanium, little dedicated effort has been expended trying to improve silicon growth techniques to achieve superior detector characteristics. Most progress in material quality has come from technology improvements aimed at other applications. The purest silicon commercially available typically has a net electrically active impurity concentration of a few times 10^{11} cm^{-3} (compared to 10^{10} cm^{-3} for HPGe), which usually limits device thicknesses to less than 1 mm. However, this thickness is sufficient for many applications. Silicon detectors are widely used for heavy-charged-particle (alpha, proton, and ion) spectroscopy. When thicker silicon devices are required (x-ray spectroscopy, beta spectroscopy, or medium-energy protons: $E > 25$ MeV), silicon of higher net purity may be obtained by lithium drifting [26], but such material cannot be subsequently processed above room temperature.

In contrast to germanium detectors, silicon detectors can be operated at room temperature. Compared to gas and scintillation detectors, silicon detectors have good energy resolution and are reasonably compact. They are fabricated from slices of a silicon single crystal and are available in a variety of areas (25–3000 mm^2), and the active thickness is usually a few hundred micrometers. Specialized detectors have been developed for a wide variety of applications.

Energetic heavy-charged particles lose kinetic energy continuously along a linear path in an absorbing material. Energy is transferred primarily to the electrons in the absorbing material but to a lesser extent to the nuclei also, via Rutherford scattering. Although only energy transferred directly to the electronic system generates electron hole pairs, Equation 81.6 (with epsilon = 3.62/pair for silicon at 300 K) is still a good approximation. Energy loss is characterized by two parameters: the specific ionization loss dE/dx, which depends on the incident particle, its energy, and the absorbing material,

and the range R (i.e., the penetration depth of the particle), which determines the detector thickness required for complete energy absorption. The continuous nature of energy loss leads to substantial window effects.

81.3.1 Diffused Junction Detector

Silicon detectors can be generically categorized by the type of rectifying contact employed. The diffused junction detector is fabricated by diffusing phosphorus from the gas phase into *p*-type silicon. This is a high-temperature (900 °C–1200 °C) operation that is prone to introducing faster diffusing metals into the bulk that can act either as generation centers increasing leakage current or as trapping centers degrading charge collection. The thickness of the diffused region, from 0.1 to 2.0 μm, also presents a dead layer to incident particles that are reduced in alternative technologies. Nonetheless, these detectors find use due to their ruggedness and economy.

81.3.2 Surface Barrier Detector

Surface barrier junctions are fabricated by evaporating either gold onto *n*-type silicon or aluminum onto *p*-type silicon. A typical entrance window is equivalent to 80 nm of silicon. The rectification properties depend on the charge density of surface states of the silicon and of the thin oxide layer over the silicon, as well as on the evaporated metal. The wafer is epoxied in an insulating ring before metallization. The finished detector is encapsulated in a can that has a front window for particle entry and a single contact in the back for the combined function of applying bias and for extracting the signal pulse. Devices can be operated either in the partially depleted or totally depleted mode. As fabrication is entirely at room temperature, there is no opportunity for metal contamination by diffusion. Generally, surface barrier detectors have lower leakage current and less system noise than a diffused junction detector of comparable area and depth. However, detectors currently fabricated by ion implantation have still lower leakage current and electronic noise, together with a thinner and more rugged front contact. On the other hand, implanted detectors are not available in the same range of active thicknesses as surface barrier detectors. Below 100 μm and above 500 μm, only surface barrier detectors are currently available. Surface barrier detectors can be made in small quantities with rather simple equipment.

81.3.3 Ion-Implanted Detectors

A simplified representation of ion-implanted detector fabrication is shown in Figure 81.15. The first successful implementation of silicon planar processing to silicon detectors was reported by Kemmer [27]. The procedure starts with the thermal growth of an oxide film on a high-purity, *n*-type silicon wafer. Windows are then opened in the oxide by photolithographic techniques. The front contact area is implanted with boron to form the rectifying contact, and arsenic is implanted into the back side. The wafer is then annealed to activate the implant, and aluminum is evaporated on both sides to reduce sheet resistivity. Typical entrance windows are 50 nm silicon equivalent. Electrical connections are made by wire bonding to the aluminum layers. Finished detectors are canned in a manner similar to surface barrier detectors. More than one detector can be fabricated on the same wafer using the appropriate masks during photolithography. In fact, quite elaborate detector geometries can be achieved via photolithography. The detector in Figure 81.15 is actually a strip type.

This ion implantation planar process technology is well suited for mass production of wafer sizes compatible with the rest of the silicon industry. Minimum wafer diameters are now 4 or 5 in. At this diameter, breakage during fabrication is an issue for thicknesses less than 150 μm. For thicknesses greater than 500 μm, the availability of enough sufficiently pure material to justify the cost of photolithographic masks is an issue. Ion-implanted detectors can be baked at 200 °C to reduce outgassing. This is a

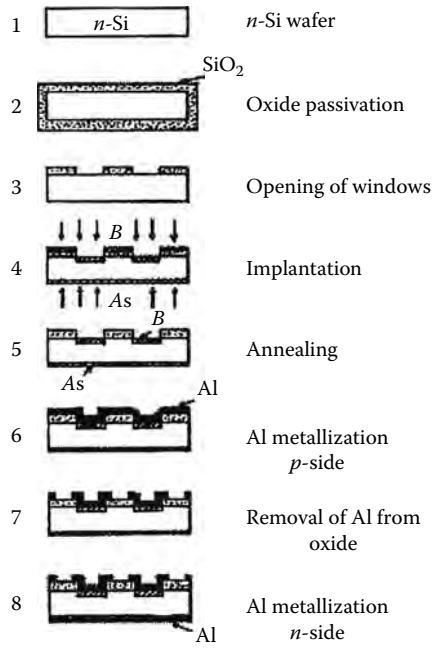


FIGURE 81.15 Steps in the fabrication of passivated-planar silicon diode detectors. (From Kemmer, J. et al., *IEEE Trans. Nucl. Sci.*, NS-29, 733, 1982.)

significant improvement over surface barrier detectors, which irreversibly degrade by device processing above room temperature. This is a useful feature, as most heavy-charged-particle spectroscopy is done in a vacuum.

Leakage currents, at room temperature, are typically 1–10 nA per cm^2 active area and per 100 μm depletion depth. These values represent an order of magnitude reduction in leakage current with respect to surface barrier detectors. Two factors are relevant. Passivation of silicon surfaces by thermal oxidation is extremely effective in reducing leakage current around the rectifying contact. Also, the bulk generation current is reduced by the gettering of metal impurities during the high-temperature oxidation. Float-zone silicon for radiation detectors usually has a minority carrier lifetime longer than 1 ms, and this can be increased to an order of magnitude during detector fabrication [28]. Thus, not only is leakage current reduced, but potential charge collection problems are also eliminated.

81.3.4 Energy Resolution

A typical spectrum of an Am-241 alpha-particle source taken with an ion-implanted detector is shown in Figure 81.16. While the factors considered in Equation 81.8 for germanium gamma-ray spectrometers are still valid, additional considerations also apply. In particular, if the source is moved closer to the detector to improve collection efficiency, larger differences in the angle of incidence will produce peak broadening due to larger variation in effective window thickness. Even when the source is sufficiently distanced from the detector, there will still be spatial variations in window thickness, as well as some variation in energy lost escaping from the source and traversing to the detector.

Another source of peak broadening is the variation in the small amount of particle energy lost during Rutherford scattering. This energy is transmitted directly to the scattering nuclei and does not generate electron–hole pairs, and a small pulse deficit results. These events are relatively few but large and therefore contribute disproportionately to peak variance. The FWHM contribution of this effect on a 6 MeV alpha-particle peak has been estimated to be 3.5 keV [29].

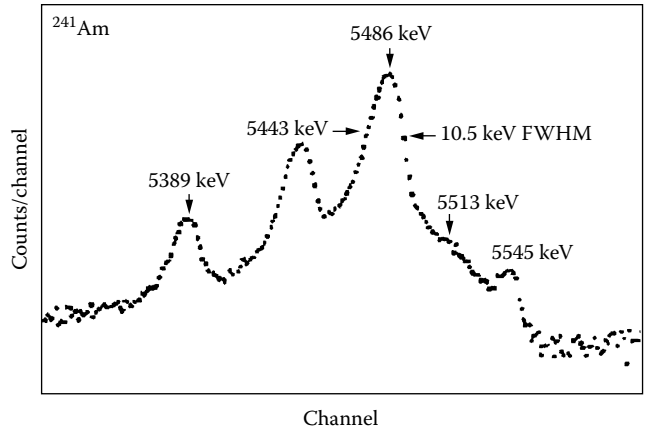


FIGURE 81.16 Spectrum of a ^{241}Am alpha-particle source (log scale) measured with an IP detector (25 mm^2 area, 300 μm thick) at room temperature. Resolution at 5.486 MeV is 10.6 keV (FWHM). (From Kemmer, J. et al., *IEEE Trans. Nucl. Sci.*, NS-29, 733, 1982.)

81.3.5 Spatial Resolution

The uninterrupted progress of the semiconductor silicon industry in achieving both larger wafers and smaller device features has allowed the development of larger and more complex silicon detectors that can provide position information in addition to (or instead of) energy information. Spatial detection can be obtained by fabricating detectors as pixels (2D) or strips (1D) on the same wafer. For penetrating radiation, two strip detectors, one behind the other but with the strip pattern rotated 90°, provide 2D positioning. Frequently, such detectors are individually designed and fabricated for a particular application. Strip detectors, drift detectors, and charge-coupled device (CCD) detectors will be discussed here.

81.3.6 Strip Detectors

Silicon strip detectors are currently fabricated on silicon wafers (typically approximately 300 μm thick) by using photolithographic masking to implant the rectifying contact in strips [30]. The strips usually have a pitch on the order of 100 μm and a width less than half of this size to minimize strip-to-strip capacitance and hence electronic noise [31]. The device is biased past depletion, and the back-blocking contact is continuous. Each strip requires, in principle, its own signal processing electronics; however, charge division readout (capacitive or resistive) can reduce the number of amplifiers by a factor of 10. Detectors are fabricated in rectangular segments from a single wafer and can be ganged together if a larger area is needed.

Strip detectors are well established in high-energy physics experiments for reconstruction on the micron scale of the tracks of ionizing particles. The particles being tracked result from the collision of accelerated particles with a target and are highly energetic (>10^{10} eV). Frequently, experimental interest is focused on short-lived particles created in the collision but which decay before they can be directly detected. The spatial resolution of the decay vertex from the original collision is necessary to detect such a particle and to determine its lifetime.

The requirements of new high-energy experiments and advances in silicon technology have produced much evolution and innovation in the strip detector concept. For example, a double-sided microstrip detector with an oxide–nitride–oxide capacitor dielectric film has been reported [32]. The use of intermediate strips to improve spatial resolution has become common [33], and the biasing network has been integrated onto the detector [34].

81.3.7 Drift Detectors

Silicon drift detectors were first proposed by Gatti and Rehak [35] as an alternative to silicon strip detectors in high-energy physics experiments. The primary motivation was to significantly reduce the number of readout channels. Drift detectors have subsequently been adapted for x-ray spectroscopy. These detectors are usually fabricated on *n*-type silicon wafers with holes collected to either a *p*+ contact on the back side of the detector or a concentric annular *p*+ contacts on the front side. The detector is depleted from both sides. The reverse bias applied to the *p*+ annular rings is varied in such a way that electrons are collected radially in a potential energy trough to an *n*+ anode at the center of the detector on the front side.

A cross section through a circular drift detector is shown in Figure 81.17. The electron collecting anode ring surrounds the integrated FET used for the first stage of signal amplification. Enough negative bias is applied to the back contact (actually the entrance window) to deplete the wafer to the anode, which is near ground potential. At the same time, negative bias, progressively increasing in magnitude, is applied from the ring next to the anode (near ground potential) to the outermost ring, which is maintained at about two times the bias of the back contact. These applied biases deplete the detector in such a way that there is an electrostatic potential minimum for electrons that varies in depth across the detector from right under the front surface at the anode to near the back contact at the last ring. Ionized electrons will drift first to this minimum and then drift radially to the anode as shown in Figure 81.17. A feature of this contacting arrangement is that the anode capacitance, and hence amplifier series noise, is low and nearly independent of the active area of the detector.

Silicon drift detectors have been designed in several different topologies of various sizes for different experimental needs. Spatial resolution for tracking and vertexing of high-energy particles is obtained by segmenting the cathode (for angular position) and analysis of signal rise time (for radial position [36]). Drift detectors with integrated electronics have been demonstrated for high-resolution room-temperature x-ray spectroscopy [37].

81.3.8 CCD Detectors

The design of CCD detectors has similarities to the silicon drift detector [35]. The CCD detector is normally fabricated on an *n*-type silicon wafer depleted both from the back side with a continuous *p*+ contact on the back and from *p*+ CCD registers on the front. Reverse bias voltages are such that the wafer is totally depleted and the electron potential minimum is about 10 μm below the CCD registers.

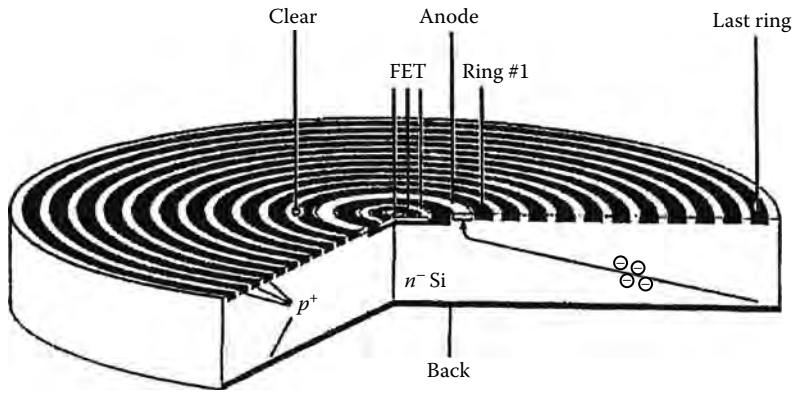


FIGURE 81.17 Cross section of a cylindrical silicon drift detector with integrated *n*-channel JFET. The gate of the transistor is connected to the collecting mode. The radiation entrance window for the ionizing radiation is the nonstructured back side of the device. (From Lechner, P. et al., *Nucl. Instru. Meth.*, A377,346, 1996.)

After an ionizing event, holes are collected to the *p*+ contacts, and electrons are trapped under a nearby register, then transported down a channel of registers by properly clocked voltage pulses to the registers. Each channel has its own readout anode, which can be made small to minimize capacitance, a prerequisite for minimizing noise. The first stage of amplification is frequently integrated onto the same wafer. Spatial resolution is limited to the register (pixel) size. Brauniger et al. [38] described initial results on a 6 × 6 cm CCD array of 150 × 150 μm pixels intended for satellite x-ray imaging. The system also had an energy resolution of 200 eV FWHM for 5.9 keV x-rays at room temperature.

Silicon pixel detectors have also been designed using other highly integrated device structures to optimize particular performance aspects such as timing resolution. Pixel detectors using MOS transistors [39] and using reverse-biased diodes with individual readout circuitry [40] have been described.

81.3.9 Present Status of Silicon Detector Technology

The simple-structured silicon detectors fabricated with parallel contacts on a silicon wafer continue to serve a well-established need for charged-particle spectroscopy. Where economies of scale can be applied, ion-implanted detectors have replaced surface barrier detectors. In x-ray spectroscopy for microanalysis (SEM and XRF), liquid-nitrogen-cooled Si:Li detectors are being challenged by similarly sized HPGe detectors, but Si:Li are still more widely used. In projects of sufficient size to support their development, specialized low-noise silicon drift detectors and CCD-based detectors have been designed and fabricated with promising room-temperature energy resolution: 200 eV FWHM at 5.9 keV. These highly structured detector technologies may find future application in liquid-nitrogen-cooled or room-temperature systems for microanalysis using x-ray spectroscopy.

In high-energy physics, the use of various strip, drift, and pixelated detectors for tracking and vertex determination has flourished. These efforts will intensify as experimental requirements for spatial resolution increase. However, radiation damage to the detector is already an issue in this application, and higher-luminosity beams will only increase the problems. Nevertheless, it appears that the continuing need of the high-energy physics community for a higher number and density of signal paths forecasts continued reliance on the ever-improving integration technology of the semiconductor silicon industry.

81.4 Room-Temperature Semiconductors

Applications arise that require energy resolution beyond the capability of scintillator systems and where cryogenically cooled semiconductors are not suited. Examples include detector probes for monitoring restricted areas, monitoring at remote sites where replenishing the coolant is impractical, spectral imaging, and many portable instrument applications. There is available a class of semiconductor detectors that satisfy many such needs by providing energy resolution substantially better than the best scintillators (although inferior to cooled semiconductors) while operating at ambient temperature. In addition to spectroscopy, these devices are also useful for counting applications where high detection efficiency per unit volume is required. In these applications, the devices are operated in pulse mode wherein the charge associated with single-photon absorption events is recorded. They also can be operated in a current mode in the manner of a solid-state ion chamber. In their current stage of development, room-temperature detectors are limited in size and best suited for the energy region below 1 MeV.

The room-temperature detectors are distinguished from cryogenic semiconductors by the magnitude of the energy gap that separates the normally vacant conduction band from the highest filled band. If this energy gap is small, as is in the case of silicon (1.14 eV) and germanium (0.67 eV), electrons can be thermally stimulated across the bandgap at room temperature. The resultant current competes with the gamma-ray-generated signal precluding room-temperature operation of germanium and high-resolution applications of silicon. Thermally stimulated current is reduced to acceptable levels at

bandgaps energies of about 1.4 eV and above. This phenomenon has been successfully exploited in the development of room-temperature detector materials including cadmium zinc telluride (acronym, CZT), cadmium telluride (CdTe), and mercuric iodide (HgI_2).

81.4.1 Theory of Operation

Operating principles of room-temperature detectors are similar to those governing the more familiar cryogenic semiconductor devices. Gamma radiation is absorbed in the material and generates electron–hole pairs that move under the influence of an applied electric field to contacts and external electronics for processing and production of the familiar pulse-height spectrum. The process is shown schematically in Figure 81.18. Fundamental to the charge transfer process is the carrier mobility (μ) and the carrier life time (τ). The product $\mu\tau E$ defines a drift length (λ) that should be long compared to the inter-contact dimensions. Owing to the substantially higher average atomic number of the room-temperature detector materials in the gamma absorption cross sections, the probability of gamma-ray absorption is much higher than in silicon or germanium (Figure 81.8). As a result, room-temperature detectors provide greater detection efficiency per unit thickness.

The energy required to produce an electron–hole pair (ε) is typically a few times the energy bandgap of the material. In silicon where the bandgap is 1.14 eV, the energy to produce an electron–hole pair (ε) is about 3.5 eV. The absorption of a 1 MeV photon in silicon thus produces about 285,000 pairs. Values of ε for room-temperature materials are in the region 4.2–5.0 eV per e–h pair (see Table 81.3), and consequently, fewer electron–hole pairs are generated per unit of absorbed energy. Complete collection of the

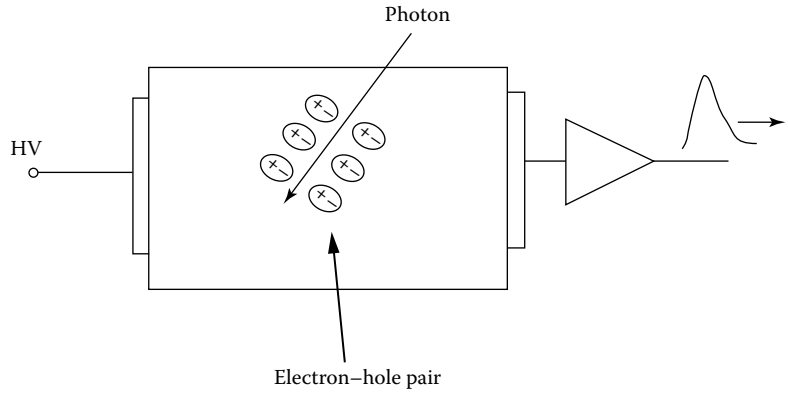


FIGURE 81.18 Schematic illustration of charge generation in a planar detector.

TABLE 81.3 Physical Parameters of Common Room-Temperature Semiconductor Materials

Material	E_g (eV)	1	ε (eV)	ρ (Ω)	$(\mu\tau)_e$ (cm²/V)	$(\mu\tau)_h$ (cm²/V)
CZT	1.65	48	5.0	10^{11}	1×10^{-3}	6×10^{-6}
Cadmium telluride	1.5	50	4.4	10^{9}	3.5×10^{-3}	2.3×10^{-4}
Mercuric iodide	2.13	62	4.2	10^{13}	1×10^{-4}	4×10^{-5}

Source: James, R.B. et al. (eds.), *Semiconductors for Room-Temperature Radiation Detector Applications*, Vol. 32, Materials Research Society, Pittsburgh, PA, 1993.

E_g, bandgap energy; Z, average atomic number; ε, energy to create an electron–hole pair; ρ, resistivity.

[1]Note also that ionizing radiation can cause both permanent and transient resistivity losses in insulating materials and should be considered in high-flux or high-fluence applications.

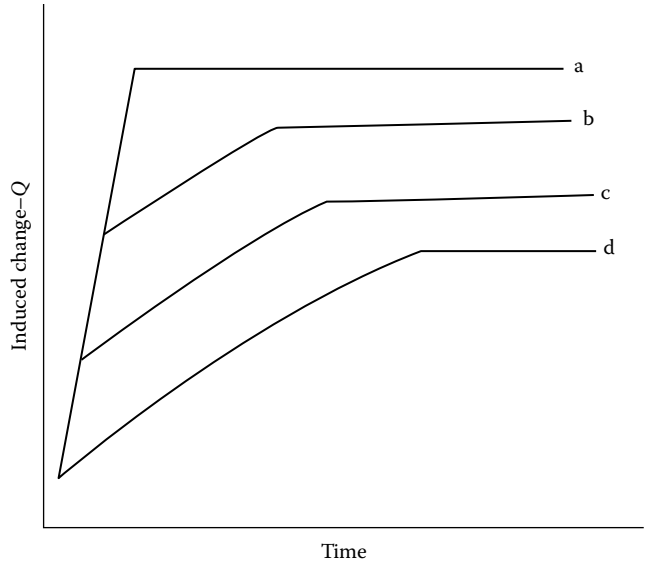


FIGURE 81.19 Charge collections in planar detector for single-photon interaction in a planar detector. Curves a through d depict the charge from photon interactions at increasing depths below the cathode.

charge is desired, although charge trapping, which may not affect the two carrier types equally, prevents this in most cases. The drift length for holes (λ_h) in these materials is often less than the intercontact dimensions and creates a condition where the collection efficiency depends on the photon interaction depth. This phenomenon is illustrated in Figure 81.19, where induced charge from single gamma absorption events originating at various depths in the material is plotted as a function of time. The initial fast-rising segment is due to the more mobile electrons; the slower component is due to holes. In this example, hole trapping is assumed and is manifest in the curvature of the hole segment. The charge collection efficiency (η) can be derived from the Hecht relation [41]. For a photon absorbed at a distance x from the cathode of a planar detector of thickness L operated with a uniform electric, the relationship becomes

$$\eta = \frac{\lambda_e}{L}\left[1 - \exp\frac{(L-x)}{\lambda_e}\right] + \frac{\lambda_h}{L}\left[1 - \exp\frac{-x}{\lambda_h}\right] \tag{81.10}$$

The dependence of the collection efficiency on interaction depth reduces energy resolution and without mitigation would limit high resolution to thin devices. Fortunately, methods have been developed that permit high-energy resolution to be achieved in relatively thick samples. As with cooled semiconductor detectors, the energy resolution of the combined detector-electronics system is normally specified by the full width of a monoenergetic spectral peak at its half amplitude points (ΔE). The FWHM is in turn related to the variance in the peak L^2 (see Equation 81.8). It is useful to note that the energy resolution is related to the reciprocal of the product $\mu\tau$.

81.4.2 Operational Considerations

Important physical parameters for the leading room-temperature detectors are summarized in Table 81.3. Detectors are available with surface area of a few square centimeters and thicknesses up to about 1 cm. The performance of detectors based on the different materials varies considerably, as can the performance for detectors of the same material. The choice of specific detector material is normally dictated by the application. The exceptionally high-resistivity and high-photoelectric cross section in mercuric

iodide permits good resolution and high efficiency in the x-ray region, particularly below 10 keV. For example, ΔE of 4% has been reported [42] with typical values in the region of 10%. For applications in the region of 0.5 MeV, a trade-off between efficient gamma absorption and resolution may be required. If energy resolution is the primary concern, thinner devices that minimize charge trapping are generally required. Considerable progress is being made in achieving both high efficiency and high resolution in the region of 0.5 MeV, particularly with CZT. For example, the resolution of better than 3% has been achieved in 1 cm thick detector at 511 keV and about 5% at 662 keV in a 2.5 cm thick device [4]. Improvements in material quality can be expected that further improve the performance of thick detectors, as the $\mu\tau$ values are further increased.

Procedures have been developed to overcome many of the thickness and surface area limitations of currently available devices. For instance, electronic circuits have been developed that permit the operation of planar arrays that provide spectral resolution approaching that of single units while providing substantially greater area [43]. Similarly, high gamma absorption efficiency with useful spectral resolution has been obtained with stacks of thin spectrometer-quality detectors [44]. Additionally, pulse-processing and single-charge collection procedures have been demonstrated that enhance spectrometer performance. This development can be expected to substantially improve the availability and price of spectrometer grade devices. Further details concerning the performance of these devices, as well as electronic processing and design details, are available in the literature [2,3,45,46].

Detectors based on the materials in Table 81.3 are available commercially. Due to the evolving nature of this technology, it is recommended that buyers' guides be consulted for suppliers and current availability.

81.5 Prices and Availability

The detectors described in this chapter are available commercially. Their prices vary widely, depending on type, size, and performance. Gaseous detectors are normally in the range of a few hundred dollars for standard designs. Scintillator–photomultiplier combinations range from about a thousand to several thousand dollars, depending on size and resolution. Room-temperature semiconductor detectors range from less than one hundred dollars for small, low-resolution devices to over a thousand dollars for large (1 cm^3), high-resolution devices. Pricing of coaxial HPGe detectors is based largely on their gamma-ray efficiency, which is specified relative to a 3 × 3 in. sodium iodide scintillator at 1.33 MeV. Coaxial detectors are available with relative efficiencies up to about 150% with cost in the area of several hundred dollars per percent efficiency. Planar HPGe detectors are normally less expensive than coaxial designs. In either case, the price includes cryostat, dewar, and preamplifier. Cryogenic silicon detectors are available in area up to several tens of square millimeters. Cost ranges to >$10,000, depending on size, performance, and complexity of design.

References

1. G. F. Knoll, *Radiation Detection and Measurement*, New York: John Wiley & Sons, 1979.
2. R. B. Rossi and H. H. Staub, *Ionization Chambers and Counters*, New York: McGraw-Hill, 1949.
3. M. J. Weber, P. Lecoq, R. C. Ruchti, C. Woody, W. M. Yen, and R. Y. Zhu (eds.), *Scintillator and Phosphor Materials*, Vol. 348, Pittsburgh, PA: Materials Research Society, 1994.
4. T. E. Schlesinger and R. B. James (eds.), *Semiconductors for Room-Temperature Nuclear Detector Applications*, Vol. 43, Semiconductors and Semimetals, San Diego, CA: Academic Press, 1995.
5. R. B. James, T. E. Schlesinger, P. Siffert, and L. Franks (eds.), *Semiconductors for Room-Temperature Radiation Detector Applications*, Vol. 302, Pittsburgh, PA: Materials Research Society, 1993.
6. J. Fraden, *AIP Handbook of Modern Sensors*, New York: American Institute of Physics, 1993.
7. M. Cuzin, R. B. James, P. F. Manfredi, and P. Siffert (eds.), *Proceedings of the 9th International Workshop on Room Temperature Semiconductor X- and Gamma-Ray Detectors*, Grenoble, France, September 18–22, 1995, *Nucl. Instru. Meth.*, 380, 1996.

8. T. E. Bortner, G. S. Hurst, and W. G. Stone, Drift velocities of electrons in some commonly used counting gases, *Rev. Sci. Instru*., 28(2), 103, 1957.
9. W. J. Price, *Nuclear Radiation Detection*, New York: McGraw-Hill, 1958.
10. F. H. Attix and W. C. Roesch, *Radiation Dosimetry*, 2nd edn., New York: Academic Press, 1966.
11. W. H. Tait, *Radiation Detection*, Boston, MA: Butterworth, 1980.
12. G. N. Whyte, *Principles of Radiation Dosimetry*, New York: John Wiley & Sons, 1959.
13. J. B. Birks, *Theory and Practice of Scintillation Counting*, New York: MacMillan, 1964.
14. J. Markakis, High resolution scintillation spectroscopy with HgI_2 as the photodetector, *Nucl. Instru. Meth*., A263, 499, 1988.
15. R. N. Hall, Chemical impurities and lattice defects in high-purity germanium, *IEEE Trans. Nucl. Sci*., 260, NS-21, 260–272, 1974.
16. E. E. Haller, W. L. Hansen, and F. S. Goulding, Physics of ultra-pure germanium, *Adv. Phys.,* 30(1), 93–138, 1981.
17. W. G. Pfann, *Zone Melting*, New York: John Wiley & Sons, 1966.
18. G. K. Teal and J. B. Little, Growth of germanium single crystals, *Phys. Rev*., 78, 647, 1950.
19. R. H. Pehl, E. E. Haller, and R. C. Cordi, Operational characteristics of germanium detectors at higher temperatures, *IEEE Trans. Nucl. Sci*., NS-20, 494, 1973.
20. L. S. Darken and C. E. Cox, High-purity germanium detector, pp. 23–80, in *Semiconductors for Room-Temperature Nuclear Detector Applications*, New York: Academic Press, 1995.
21. R. H. Pehl, N. W. Madden, J. H. Elliott, T. W. Raudorf, R. C. Trammell, and L. S. Darken, Jr., Radiation damage resistance of reverse electrode GE coaxial detectors, *IEEE Trans. Nucl. Sci*., NS-26, 321, 1979.
22. ANSI, *IEEE Test Procedures for Germanium Detectors for Ionizing Radiation*, ANSI/IEEE Standard 325–1989.
23. R. H. Pehl and F. S. Goulding, Recent observations on the fano factor in germanium, *Nucl. Instru. Meth*., 81, 329–330, 1970.
24. J. Lacer, E. E. Haller, and R. C. Cordi, Entrance windows in germanium low-energy x-ray detectors, *IEEE Trans. Nucl. Sci*., NS-24, 53, 1977.
25. C. E. Cox, B. G. Lowe, and R. Sareen, Small area high purity germanium detectors for use in the energy range 100 eV to 100 keV, *IEEE Trans. Nucl. Sci*., 35, 28, 1988.
26. E. M. Pell, Ion drift in an *n-p* junction, *J Appl. Phys*., 31, 291, 1960.
27. J. Kemmer, Fabrication of low noise silicon radiation detectors by the planar process, *Nucl. Instru. Meth*., 169, 499–502, 1980.
28. J. Kemmer and G. Lutz, New detector concepts, *Nucl. Instru. Meth*., A235, 365–377, 1987.
29. G. D. Alkhazov, A. P. Komar, and A. Vorob'ev, Ionizing fluctuations and resolution of ionization chambers and semiconductor detectors, *Nucl. Instru. Meth*., 48, 1–12, 1967.
30. J. Kemmer, P. Burger, R. Henck, and E. Heijne, Performance and applications of passivated ion-implanted silicon detectors, *IEEE Trans. Nucl. Sci*., NS-29, 733, 1982.
31. T. Dubbs, S. Kashigin, M. Kratzer, W. Kroeger, T. Pulliam, H. E.-W. Sadrozinski et al., Noise determination in silicon micro strips, *IEEE Trans. Nucl. Sci*., 42, 1119, 1996.
32. Y. Saitoh, T. Akamine, M. Inoue, J. Yamanaka, K. Kadoi, R. Takano, Y. Kojima et al., Fabrication of a double-sided silicon microstrip detector with an ONO capacitor dielectric film, *IEEE Trans. Nucl. Sci*., 43, 1123, 1996.
33. P. Chochula, V. Cindro, R. Jeraj, S. Macek, D. Zontar, M. Krammer, H. Pernegger, M. Pernicka, and C. Mariotti, Readout of a Si strip detector with 200 μm pitch, *Nucl. Instru. Meth*., A377, 409–411, 1996.
34. T. I. Westgaard, B. S. Avset, N. N. Ahmed, and L. Eversen, Radiation hardness of punch-through and FET biased silicon microstrip detectors, *Nucl. Instru. Meth*., A377, 429–434, 1996.
35. E. Gatti and P. Rehak, Semiconductor drift chamber—An application of a novel charge transport scheme, *Nucl. Instru. Meth*., 225, 608–614, 1984.
36. P. Rehak, J. Walton, E. Gatti, A. Longoni, M. Sampietro, J. Kemmer, H. Dietl et al., Progress in semiconductor drift detectors, *Nucl. Instru. Meth*., A248, 367–378, 1986.

37. P. Lechner, S. Eckbauer, R. Hartman, S. Krisch, D. Hauff, R. Richter, H. Soltau et al., Silicon drift detectors for high resolution room temperature x-ray spectroscopy, *Nucl. Instru. Meth*., A377, 346–351, 1996.
38. H. Brauninger, R. Danner, D. Hauff, P. Lechner, G. Lutz, N. Meidinger, E. Pinotti, C. Reppin, L. Struder, and J. Trumper, First results with the *pn*-CCD detector system for the XMM satellite mission, *Nucl. Instru. Meth*., 326(1–2), 129–135, 1993.
39. K. Misiakos and S. Kavadias, A pixel segmented silicon strip detector for ultra fast shaping at low noise and low power consumption, *IEEE Trans. Nucl. Sci*., 43, 1102, 1996.
40. E. Beauville, C. Cork, T. Earnest, W. Mar, J. Millaud, D. Nygen, H. Padmore et al., A 2D smart pixel detector for time resolved protein crystallography, *IEEE Trans. Nucl. Sci*., 43(3), 1243, 1996.
41. H. K. Hecht, Zum mechanisms des lichtelektrischen primarstromes in isolierenden kristallen, *Z Phys*., 77, 235–245, 1932.
42. J. S. Iwanczyk, B. E. Patt, Y. J. Wang, and A. K. Khusainov, Comparison of HgI_2, CdTe, and Si (*p-i-n*) x-ray detectors, *Nucl. Instru. Meth*., A380, 186–192, 1996.
43. V. Gerrish, private communication, 1995.
44. R. Olsen, R. B. James, A. Antolak, and C. Wang, *Proceedings of the 1994 International Nuclear Materials Management Conference*, Bruges, Belgium, Vol. 23, p. 589, 1994.
45. B. E. Patt, J. S. Iwanczyk, G. Vikelis, and Y. J. Wang, New gamma-ray detector structures for electron only charge carrier collection utilizing high-Z compound semiconductors, *Nucl. Instr. Meth*., A380, 276–281, 1996.
46. P. N. Luke, Electrode configuration and energy resolution in gamma-ray detectors, *Nucl. Instr. Meth*., A380, 232–237, 1996.

Further Readings

Guerra A. D., *Ionizing Radiation Detectors for Medical Imaging*, World Scientific Publishing, Co. Pte, Ltd, Singapore, 2004

Lutz G., *Semiconductor Radiation Detectors*, 2nd edn., Springer, Berlin, Germany, 2001.

Maas M. C., *Monolithic Scintillator Detectors for High Resolution Positron Emission Tomography*, IOS Press, Amsterdam, the Netherlands, 2008.

Partial List of Vendors and Suppliers

Gamma products, Inc., 7730 W. 114th Pl., Palos Hills, IL 60465, Phone: 708-974-4100, Fax: 708-974-0071, www.gammaproducts.com

Industrial Test Systems, Inc., 1875 Langston Street, Rock Hill, SC 29730, Phone: 803-329-9712, 800-861-9712 (toll free), Fax: 803-329-9743, www.sensafe.com

LPR Global, 720 Bathurst St., 408, Toronto, Ontario M5S2R4, Canada, Phone: 416-423-5590, Fax: 416-406-6141, www.uskoreahotlink.com/

National X-Ray Corp., 595 Old Norcross Rd., Suite G, Lawrenceville, GA 30045, Phone: 866-484-9269 (toll free), Fax: 678-960-0085, www.nationalx-ray.com

Princeton Gamma-Tech, Inc., 303C College Rd. E., Princeton, NJ 08540, Phone: 609-924-7310, Fax: 609-924-1729, www.pgt.com

Radiation Monitoring Devices, 44 Hunt St., Suite 2, Watertown, MA 02472, Phone: 617-668-6800, 800-532-3763 (toll free), 800-LEAD-RMD (toll free), Fax: 617-926-9980, www.rmdinc.com

Rapiscan Systems, 1901 S. Bell St., Suite 325, Arlington, VA 22202, Phone: 703-812-0322, Fax: 703-812-0335, www.rapiscansystems.com/



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

82

Charged-Particle Measurement

82.1 Introduction **82**-1
Interaction of Charged Particles with Matter • Energy Loss of Protons and Ions • Energy Loss of Electrons • Technologies
82.2 Nuclear Emulsions **82**-5
82.3 Gas-Filled Charged-Particle Counters **82**-6
Gas-Filled Proportional Counters • Fill Gases for Proportional Counters • Proportional Counter Operation and Readout • Proportional Counters: Advanced Techniques • Geiger–Müller Counters • Proportional Counter and Geiger–Müller Counter Applications
82.4 Scintillation Counters **82**-11
Introduction • Scintillation Counters: A Detailed Description • Scintillation Medium • Light Coupling to Photon Detectors • Light Detection and Readout • Scintillator Detector Construction • Operating Information • Advanced Techniques • List of Manufacturers
82.5 Solid-State Detectors **82**-15
Signal Generation • Categories of Detector • Detector Systems • Plasma Effects and Pulse Height Defect • New Structures • Manufacturers and Prices
References **82**-20

John C. Armitage
Carleton University

Madhu S. Dixit
Carleton University

Jacques Dubeau
Carleton University

Hans Mes
Carleton University

F. Gerald Oakham
Carleton University

82.1 Introduction

82.1.1 Interaction of Charged Particles with Matter

There are a number of subatomic charged particles that can be detected by their interaction with matter. These include protons (hydrogen nuclei), β-particles (fast electrons), α-particles (helium nuclei), light ions, heavy ions, and fission fragments. Some of these particles are emitted by natural radioactivity on Earth or originate from space and reach the Earth in the form of cosmic rays; others are due to human activities (nuclear industry, accelerators). The kinetic energy of charged particles is given in units of electron volts (eV), where $1\ \text{eV} = 1.6 \times 10^{-19}$ J. Naturally occurring α- and β-particles from radioactive decay have energies up to 10 MeV. Ions, such as Cl^+, are accelerated to 35 MeV and higher in beam analysis procedures. Energetic particles that originate from space can have energies up to 10^{12} GeV. Charged particles traversing matter lose kinetic energy until they eventually come to a halt. In the case of electrons, protons, α-particles, or light ions, most of this energy is lost through interactions with the electrons of the target. Heavier ions also lose a significant amount of kinetic energy through direct collisions with the nuclei of the constituent atoms.

However, for all types of charged particles of concern here, ionization of the target atoms is the dominant mode of excitation of the detection medium. This process is also the main source of radiation damage in biological tissue.

82.1.2 Energy Loss of Protons and Ions

The quantity of energy deposited by an incident energetic charged particle in a detection medium is a function of the atomic mass, A_{med}; the atomic number, Z_{med}; and the density of target atoms that it meets along its path. Hence, the unit of dimension used in this field is the mass thickness, and it is measured most often in units of g cm^{-2}. This is equal to the geometric thickness or path length x, multiplied by the density of the detector volume, ρ. It follows that the rate of energy loss of a particle in the detection volume is $dE/(\rho\, dx)$ and it is given in units of keV (g cm^{-2})$^{-1}$ or MeV (g cm^{-2})$^{-1}$. The energy loss due to ionization is characterized by the Bethe–Bloch equation:

$$-\frac{1}{\rho}\frac{dE}{dx_{ion}} = D\frac{z^2 Z_{med}}{A_{med}}\frac{c^2}{v^2}\left[\ln\frac{2m_0v^2}{I_{med}} - \ln\left(1-\frac{v^2}{c^2}\right) - \frac{v^2}{c^2}\right] \quad \text{MeV (g cm}^{-2})^{-1} \tag{82.1}$$

with $D = 0.307$ MeV (g cm^{-2})$^{-1}$ and c is the velocity of light (3.0×10^8 m s^{-1}). The incident ion is described by z and v, its charge and velocity, respectively. The response of the atoms of the target are further defined by $I_{med} \sim 16\, Z_{med}^{0.9}$ eV, their mean ionization energy. The energy loss is characterized by a peak at low energy and a minimum at higher energies where a particle of $z = 1$ is referred to as a minimum ionizing particle (MIP). The energy loss curves of protons in Si, Ar, and Au are shown in Figure 82.1. The energy loss expressed in MeV (g cm^{-2})$^{-1}$ is lower for higher Z_{med} materials, where the ratio Z_{med}/A_{med} becomes less than 0.5 and I_{med} is large. However, when multiplied by the material's density, the energy loss in MeV cm^{-1} shows a reverse trend; it is greater for high Z_{med} materials because they have a high density. The rate of energy loss varies to the second power of the charge z of the incident ion and as the inverse square of its kinetic energy ($1/v^2$ term). Equation 82.1 holds for protons and α-particles of energy above 500 keV and 1 MeV, respectively, in Si where the target is described by an

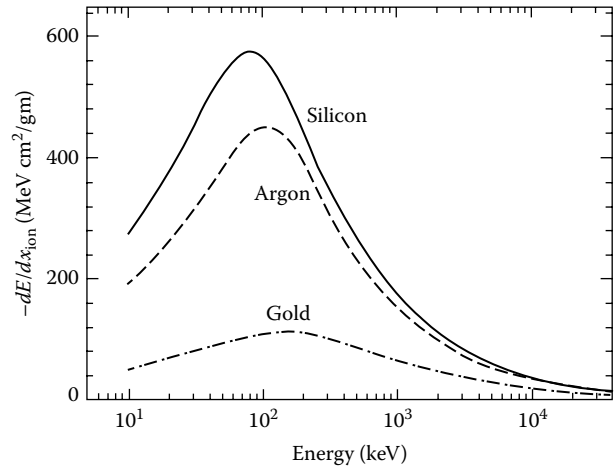


FIGURE 82.1 Electronic energy loss curves for protons in three different absorber elements are shown as a function of kinetic energy. The target element with the lower Z_{med} displays the highest stopping power due to its low average ionizing potential. These curves were obtained using the SRIM computer program. (From Ziegler, J.F., *SRIM: The Stopping and Range of Ions in Matter*, IBM-Research, Yorktown, NY, 1996.)

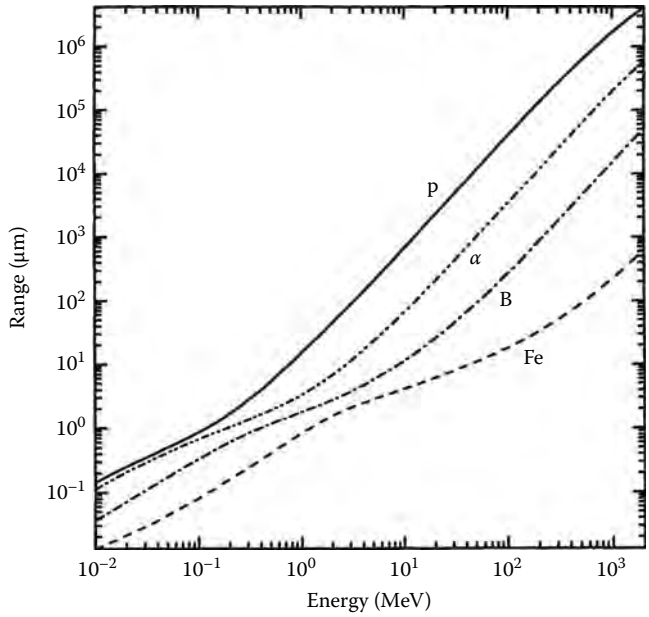


FIGURE 82.2 Ions of widely different masses have widely different ranges in a target element such as silicon. These curves for protons (p), α-particles, boron ions (B), and iron (Fe) were obtained using the SRIM computer program. (From Ziegler, J.F., *SRIM: The Stopping and Range of Ions in Matter*, IBM-Research, Yorktown, NY, 1996.)

average ionization energy. It is not applicable for ions for which $z > 2$, where stripping must be considered. Ziegler et al. [1] give a very useful description of scaling rules that allow the proton energy loss to be scaled to any incident ion and target atoms. They have included these rules in a software package called SRIM [2]. Their empirical relation allows for ion stripping and gives a more exact representation of the energy loss than Equation 82.1 in the case of heavy ions and for low-velocity light ions. Alternatively, one can use published $\mathrm{d}E/(\rho\,\mathrm{d}x)_{\mathrm{ion}}$ plots [3]. The interaction of fast charged particles with the nuclei of the constituent atoms of a target produces an additional contribution to the energy loss, $\mathrm{d}E/(\rho\,\mathrm{d}x)_{\mathrm{nucl}}$, the nuclear stopping power. For protons and α-particles, the electronic stopping power dominates by a factor of 1000 over the nuclear stopping power for most energies of interest. For ions, the nuclear stopping power may dominate at low energies (<1 MeV), with its contribution decreasing at higher energies and becoming negligible. The range of a particle, or its path length, is also a very important parameter in the selection of a detector. It is given by

$$R(E_0) = \int_0^{E_0} \frac{\mathrm{d}E}{\mathrm{d}E/\mathrm{d}x} \tag{82.2}$$

where $\mathrm{d}E/\mathrm{d}x$ in this case is the total stopping power $\mathrm{d}E/\mathrm{d}x_{\mathrm{ion}} + \mathrm{d}E/\mathrm{d}x_{\mathrm{nucl}}$. Range–energy curves for various ions in silicon are shown in Figure 82.2.

82.1.3 Energy Loss of Electrons

The case of electron propagation through matter is more complicated. The electron collides with bodies having a mass equal to it (other electrons) or much heavier (nuclei). The collisions cause it to undergo many changes in direction. This "random walk" type of behavior is superimposed on the forward motion of the electron. Along its random path, an electron loses energy through ionization in a way similar to ions. At increasing kinetic energy, electrons also lose energy through the emission of x-rays,

called bremsstrahlung radiation. In a target constituted of atoms of charge Z_{med} and below a critical energy E_c, ~817/Z_{med} MeV, electrons lose a greater fraction of their energy through ionization. Above E_c, the energy loss due to bremsstrahlung x-ray emission dominates and is given by

$$-\frac{1}{\rho}\frac{dE}{dx_{rad}} = B\frac{Z_{med}^2 E}{A}\left(\ln\left(\frac{183}{Z_{med}^{1/3}}\right)+\frac{1}{18}\right)\text{MeV (g cm}^{-2})^{-1} \tag{82.3}$$

where $B = 1.40 \times 10^{-3}$ for E in MeV and for the molar weight of the target A in grams. Equation 82.3 applies for specified energy ranges, and the interested reader should consult the available literature [4].

The electron range is defined as the distance separating the point of entry in the detection medium to the point where the electron stops. It is much shorter than its actual path length and can be approximated by $R(E)$ (Equation 82.4) for E ranging from 0.01 to 3 MeV [5]. Ultimately, the problem of electron propagation through materials has to be addressed using Monte Carlo programs (see Refs. [6,7]), which follow the individual electron histories to a specified energy or geometric cutoff:

$$R(E) = 0.2115 Z_{med}^{0.26} E^n (\text{g cm}^{-2}) \tag{82.4}$$

$$n = 1.265 - 0.0954 \ln(E)$$

82.1.4 Technologies

The most appropriate detection technique selected depends on the particle under study, its energy, and the type of information sought, such as source activity, position or energy measurement, or particle identification. For all cases reviewed here, except nuclear emulsions, the detection process is active and accompanied by electronic signal processing. Nuclear emulsions permanently record the ionization tracks left by traversing particles and can only be examined after their development. The active technologies include gas-filled detectors (ionization chambers, proportional counters, Geiger–Müller counters), scintillators, and solid-state detectors. The active volume consists of a radiation-sensitive material that is raised to an excited state through ionization by the passage of charged particles. For active detectors, this ionization is sensed by the external electronics as an induced charge in the case of gas-based and solid-state detectors or as visible or near-visible light in the case of scintillators. The sensors described here are the front ends of complex detection systems.

Nuclear emulsions are a special preparation of photographic emulsion designed to record "tracks" left by ionizing particles as they traverse the emulsion. The emulsion is sensitive up to the time it is processed, and the image is latent for many months or years if the emulsion is cooled. Once processed, the emulsion is a permanent record of the ionizing radiation that traversed it. In a situation where a prompt signal is not required, the nuclear emulsion may be a suitable detector. Its primary advantage is that it is compact, is self-contained, and has high spatial resolution. Since no associated equipment is needed during exposure, it is well suited for remote applications, such as studying cosmic rays and for recording the cumulative exposure to ionizing radiation in space.

Gas-filled detectors were among the first instruments designed for radiation detection. A charged particle passing through a gas-filled container leaves a trail of ionization electrons and ions. With a suitably configured external electric field, the ionization can be used to produce an electric signal for charged-particle detection. Depending on the magnitude of the applied electric field, a variety of ionization phenomena can arise in gases that have been used over the years to develop different types of gaseous charged-particle detectors, for example, ionization chambers, proportional counters, Geiger–Müller counters, and spark chambers. Spark chambers are rarely used these days because of electric

TABLE 82.1 Detection Applications and Suggested Instruments

Detector Type	Typical Applications	Advantages	Shortcomings
Nuclear emulsion	Cosmic ray studies, dosimetry, low statistic (single track) studies	Passive detector, high spatial resolution, unattended operation, particle identification, and energy measurement	No time information, manual scanning, chemical development
Gas-filled counter	Radiation monitoring, dosimetry	Low density, large area at low cost, nuclear and particle physics experiments, real-time particle identification	Needs ultrapure flammable gases, high voltages required, signals need amplification
Scintillation counter	Energy measurement, particle counting, triggering	Large volumes, flexible shapes, fastest timing, high efficiency	Poor spatial resolution, poor energy resolution
Solid-state detector	Energy measurement, particle identification, position measurement, radiation monitoring, beam analysis techniques	High-energy resolution, compact size	High cost per area, signal needs amplification

noise and limited counting rate ability. The other three devices are used primarily as instruments for radiation protection and monitoring. Their advantage is that they are inexpensive, easy to operate, and available commercially. In addition, many variants of gas-filled proportional counters have been developed for use in experimental nuclear physics and elementary particle physics. These, however, need to be specifically designed and fabricated for particular applications.

Scintillation counters consist of a scintillation material, a photodetector, and a light guide to link the two. Charged particles crossing the scintillator material deposit energy in the material, which responds by emitting light isotropically. A large fraction of the light can be trapped in the material by internal reflection and directed toward a photodetector, such as a photomultiplier (PM) tube. This kind of detector can be used at very high counting rates and can achieve efficiencies approaching 100%. Components for scintillation counters are readily available and it is possible to purchase assembled counters.

Solid-state detectors are large semiconductor diodes made of very high-resistivity material and are operated under reverse bias. The small bandgap in these materials ensures that the kinetic energy of the incident charged particle is efficiently converted into an electric signal through the creation of electron–hole pairs. A signal is induced on the metal contacts as the charge carriers drift under the applied electric field present in the depletion region. Their advantages include a fast response, good energy resolution, and compact size.

Table 82.1 presents some radiation-sensing applications, the particles under study, and the suggested sensor–system combination. The following subsections present each sensing technology in more detail.

82.2 Nuclear Emulsions

Nuclear emulsions are a suspension of silver halide grains, primarily AgBr, in a gelatin medium. The silver halide grains are small, on the order of 0.2–0.5 μm, and constitute about half the volume and 80% of the total weight. The "pellicles" of nuclear emulsion range from 25 to 600 μm in thickness and can be as large as 70 cm by 30 cm. The pellicles may be supported on a glass or plastic plate during exposure or stacked to increase the thickness of the emulsion detector, in which case the pellicles will be placed on a supporting plate as part of processing. If pellicles are stacked for an exposure, then a reference grid is usually printed on the bottom surface of each pellicle to assist in following tracks from one pellicle to the next.

The ionizing radiation crossing a grain of silver halide leaves the grain in an excited state. During development, the excited grains are reduced to elemental silver. The undeveloped grains are washed out

during the fixing process. Due to their thickness and to ensure uniform response, a special development process, involving presoaking the emulsions at a reduced temperature, must be used. After development, the nuclear emulsion will have shrunk to about 40% of its original thickness. The thickness is also sensitive to the ambient humidity, and the plates are normally kept in humidity-controlled storage (~50% relative humidity). The shrinkage needs to be taken into account for most measurements.

The track image in the emulsion is viewed with a microscope with magnification in the range of 10 × 10 to 10 × 100. With the aid of a special microscope, having a calibrated micrometer vertical focus adjustment and a special high-precision stage with calibrated micrometer x and y movements, the characteristics of the tracks can be determined and used to identify most particles and to measure their momentum. The dE/dx is measured by determining the grain density along the track. The energy or momentum is determined from the range of the particle or from multiple scattering. Additional information can also be obtained from the density of associated delta rays. These measurements can usually be combined to uniquely determine the charge (but not the sign) and momentum of the particle and the mass in the case of a singly charged particle. The interested reader may find additional information in Refs. [8–10].

Nuclear emulsions are currently produced by Ilford Photographic (Ilford Research Laboratory, Ilford Ltd., London, England) and Fuji Photo Film Co. Ltd. (Tokyo, Japan). Several types of emulsion are available with different sensitivities to ionizing radiation.

82.3 Gas-Filled Charged-Particle Counters

A gas-filled counter detects the ionization electrons and positive ions produced by a charged particle passing through a gas volume. An electrostatic field established between a pair of electrodes in a gas-filled container causes the electrons to move toward the anode and the ions toward the cathode. The charge movement produces an electric pulse that can be detected using suitable electronics. A gas-filled counter can operate as an ionization counter, a proportional counter, or a Geiger–Müller counter. Figure 82.3 shows the three operating regions with characteristics determined by the electric field region in which they operate.

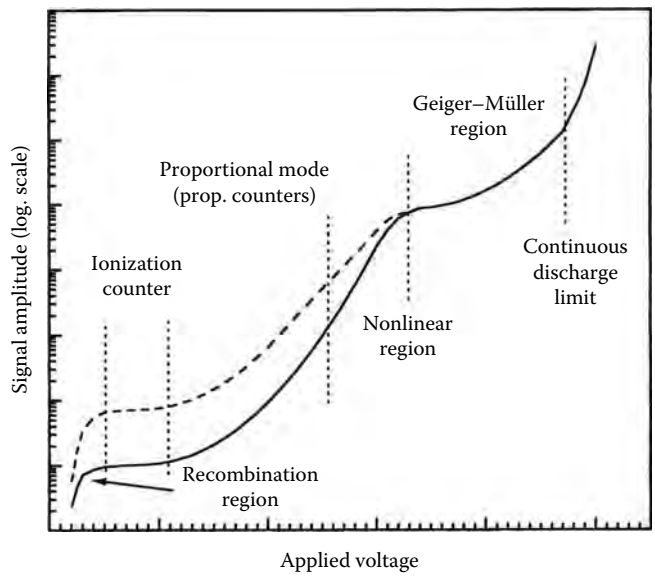


FIGURE 82.3 Typical behavior at atmospheric pressure for a gas-filled counter. Three principal operating regions are identified: ionization region, proportional region, and Geiger–Müller region. The lower curve is for minimum ionizing charged particles and the upper one for heavily ionizing ones. At low operating voltages, the ions and electrons recombine. Continuous discharge occurs at high voltages above the Geiger–Müller region of operation.

Ionization counter/chamber: At low fields, the recombination of electron–ion pairs reduces the detector signal, making it electric field dependent. With increasing field, nearly all the charges can be collected before they recombine. In an ionization counter, the signal-to-noise ratio is quite unfavorable, except when used for the detection of heavily ionizing particles (e.g., alpha particles). For radiation monitoring, the device is generally used in integrating mode as an ionization chamber. The current signal in an ionization chamber is proportional to the rate of ionization produced by charged particles interacting in the detector gas volume.

Proportional counter: At still higher fields, the electrons acquire enough energy between collisions with gas molecules to produce secondary ionization in the gas. The secondary electrons in turn can produce further ionization along their path, leading to an electron avalanche that stops only when all the electrons are collected at the anode. This large increase in the number of primary electrons leads to a detector signal substantially larger than that produced by a pulse ionization counter. The signal amplitude in a proportional counter is directly proportional to the ionization energy deposited in the gas by the ionizing particle. The proportional counter can detect charged particles without loss of gain at incident particle rates approaching 10^4 s^{-1} for each millimeter length of anode wires.

Geiger–Müller counter: Increasing the electric field still further results in increasing loss of proportionality until the counter enters the Geiger–Müller region of operation. In the Geiger–Müller region of operation, a single ionizing event induces an electric discharge along the entire length of the anode wire. The output pulse amplitude is large and independent of the energy deposited by the ionizing particle in the gas. The Geiger–Müller counter has a relatively long dead time of a few hundred microseconds, and the detector saturates at counting rates above a few thousand particles per second.

Proportional counters have many advantages over ionization chambers and Geiger–Müller counters, both of which are now used mainly as radiation monitors in the laboratory. A gas-flow proportional counter is sensitive to the ionization produced in a gas volume by a single charged particle. It can be used for counting charged particles, to measure their energy loss in the gas and to measure their track coordinates. The proportional counter is more sensitive than an ionization counter because of a built-in gas multiplication process that amplifies the initial ionization pulse. Because of its relatively short pulse duration, a proportional counter recovers quickly and can count particles at higher rates than a Geiger–Müller counter with its long dead time.

82.3.1 Gas-Filled Proportional Counters

82.3.1.1 Basic Design and Operating Principle

Figure 82.4 shows the schematic diagram of gas-filled proportional counter and associated readout electronics. An electrically conducting gas-filled tube serves as the proportional counter cathode. A fine wire 20–25 μm in diameter (usually made of gold-plated tungsten) at the center of the tube, supported by insulators on both ends, forms the anode. For reasons of mechanical and electric stability, the anode wire is strung under a tension of several tens of grams, depending on the wire length and operating voltage of the detector. Mylar windows thin enough to allow passage of α- and β-particles close the gas volume. The cathode is usually kept at ground potential and a positive voltage V_0 is applied to the anode. The electrostatic field is radial with magnitude given by

$$E = \frac{V_0}{(r \ln(b/a))} \tag{82.5}$$

where

r is the distance from the central axis
a is the radius of the anode wire
b is the radius of the tube

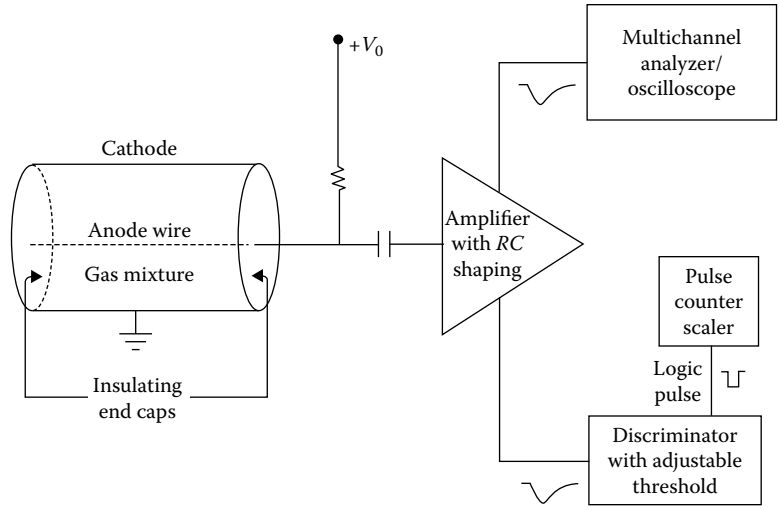


FIGURE 82.4 Schematic diagram of a gas-filled proportional counter and measurement system.

The $1/r$ dependence leads to the existence of an intense electric field near the anode wire; for example, at $r = 20$ μm, the electric field is 25 MV m^{-1} for a 20 μm diameter anode wire held at 1.5 kV within a cathode tube of radius 1 cm. A charged particle passing through the detector ionizes the gas molecules along its path. Under the influence of electric field, the primary ionization electrons start drifting toward the anode. In the strong anode field region a few wire radii from the anode surface, the electrons produce additional ionization along their paths, which leads to the production of an electron avalanche. The large increase in number of primary electrons is called gas multiplication or gas gain, which becomes significant in the high-field region near the anode. The gas gain is related to the mean free path of the electrons in the gas for secondary ionizing collisions. The electron mean free path is a function of the electric field and the gas pressure. Gas gains as high as 10^8 can be achieved by proper choice of gas mixture and mechanical construction before reaching the limit imposed by electric breakdown in the gas. Most commonly used proportional counter systems have gas gains in the range of 10^4–10^6 [4,11,12].

The voltage pulse observed on the anode is the result of a change in the electrostatic potential energy of the system. Most of the potential energy change is due to the movement of positive ions away from the high anode field region. The contribution of electrons to the total signal is quite small since most of the avalanche electrons traverse only a short distance before being collected by the anode. The fast movement of the ions away from the intense electric field region near the anode produces an initial rapid rise of the signal. About 50% of the pulse height is reached in roughly 0.1% of the total time, followed by a slow logarithmic rise lasting a few hundred microseconds. The signal is proportional to the total amount of ionization produced in the gas, that is, the product of the gas gain and the number of primary electron–ion pairs. The proportional counter energy resolution is determined by statistical fluctuations in the number of ions and by the statistical nature of the gas gain process itself. Deviations from proportionality occur when the space charge density of the positive ion sheath around the anode wire becomes large enough to locally reduce the electric field. The space charge effects become appreciable, depending on the gas gain at counting rates above 10^4 s^{-1} mm^{-1} of anode wire length.

82.3.2 Fill Gases for Proportional Counters

The choice of proportional counter fill gases is dictated by practical considerations. If low gas gain is acceptable, such as for the detection of α-particles or heavy ions, both of which produce extremely high specific ionization (initial number of electron–ion pairs produced per cm), almost any gas (even air)

TABLE 82.2 Characteristics of Gases Commonly Used in Proportional Counter Gas Mixtures

Gas	W (eV)	dE/dx_{ion} (keV cm^{-1})	Specific Ionization (Ion Pairs per cm)
Ar	26	2.44	94
Xe	22	6.76	307
CO_2	33	3.01	91
CH_4	28	1.48	53
C_4H_{10}	23	4.50	195

Note: dE/dx_{ion} and specific ionization values are for minimum ionizing charged particles in gas at atmospheric pressure.

can be used. For other particles, one requires a gas that has high specific ionization, high gas gain, low operating voltage, good proportionality, high single-particle counting rate ability, and long operating lifetime. Noble gases such as argon meet the criteria of reasonable gas gain at low operating voltage and high specific ionization. However, the ultraviolet photons produced by excited argon atoms can cause secondary electron emission from cathodes, leading to electric breakdown at higher operating voltages. The problem can be cured by adding a quenching polyatomic gas with rotational and vibrational levels that can readily absorb the ultraviolet photons. The addition of a quenching gas permits stable high-gain operation of proportional chambers. A mixture of 90% argon and 10% methane (CH_4) called P10 is a commonly used proportional counter gas. Several different gas mixtures exist that are capable of achieving proportional gas gains in the neighborhood of 10^6.

The specific ionization for a gas can be calculated from the knowledge of dE/dx_{ion}, the rate of charged-particle energy loss in the gas, and W, the mean energy needed to create a single electron–ion pair. The properties of some common proportional counter gases are shown in Table 82.2. The values of dE/dx_{ion} and the specific ionization shown in the table are for minimum ionizing charged-particle tracks. The specific ionization for a gas mixture is the average specific ionization calculated with partial pressures of the component gases as weights.

The use of organic quenching gases as quenchers in proportional chamber gas mixtures can lead to the formation and deposition of polymers on the electrodes that reduce the operating life of a proportional chamber. To improve chamber lifetime and reduce aging, nonpolymerizing agents such as methylal are frequently added in small quantities to the gas mixture. For proportional counters with sealed gas volumes, the detector performance will degrade due to the contamination of gas in a sealed volume caused by component outgassing. Cleanliness and careful choice of components during fabrication can minimize these effects. The performance of a sealed proportional counter will also degrade with the degradation of the quenching gas component, which can happen when the device is exposed to large fluxes of radiation or after prolonged use. To minimize these problems, proportional counters are often used with external gas supply systems capable of refreshing the gas continuously.

82.3.3 Proportional Counter Operation and Readout

It takes a few hundred microseconds to develop the full voltage pulse in a proportional counter after a leading edge rise time of a few hundred nanoseconds. Even with a gas gain of 10^4–10^5, the pulse amplitude is only a few millivolts. Further amplification and shaping is needed to obtain signals suitable for measurements. Often, low-noise amplifiers with built-in *RC* differentiation to shorten the output pulses to 1 μs or less are used (see Figure 82.4). These improve the ability of proportional counters to measure particles at high counting rates without pulse pileup distortions caused by closely spaced pulses overlapping in time. The energy loss of a charged particle in a gas can be measured by measuring the pulse height of the anode signal with an oscilloscope or a multichannel analyzer (MCA). For charged-particle

counting, a discriminator can be used to produce logic pulses that can be counted with a scaler. For coincidence applications, timing resolution of a few nanoseconds is typical for a proportional counter. Good RF shielding and careful electric grounding are essential to maximize the sensitivity of most proportional counters.

The performance of a proportional counter is determined by operating voltage. At lower operating voltages, the best linearity is achieved at the expense of lower gas gain and, consequently, a low signal-to-noise ratio and low efficiency for particle detection. Increasing the voltage leads to larger gain, better signal-to-noise ratio, reduced linearity, faster aging, and susceptibility to electric sparking and damage. In general, the lowest possible anode voltage consistent with achieving good particle detection efficiency is preferred for the operation of proportional counters. This can be accomplished by choosing the operating voltage roughly in the middle of the high-voltage (HV) plateau for the device. On the HV plateau, the counting efficiency of the device remains constant for any variations of anode HV [12]. Typical proportional counter operating voltage plateaus are reached in the voltage range of 1.5–2.5 kV.

82.3.4 Proportional Counters: Advanced Techniques

Many variants of the simple, single wire design have been developed for nuclear physics and particle physics experiments [4,11,12]. Multiwire proportional counters were invented by George Charpak in 1968. In a multiwire proportional chamber, a parallel equidistant row of anode wires is enclosed between two cathode planes. Each anode acts as an independent proportional counter that is read out using suitable electronics. Segments of charged-particle tracks can be localized in a multiwire proportional counter in the anode wire plane in a direction perpendicular to the anodes with accuracies on the order of 0.5 mm. A second coordinate along the length of the anode can also be obtained by a measurement of induced signals on a cathode that is segmented in strips in a direction normal to the anode wires. Much more precise particle track coordinate measurements with accuracies of tens of micrometers are possible in another variant called the drift chamber. This is achieved from knowledge of drift velocities of electrons in the chamber gas as a function of electric field and by measuring the drift time of the ionization electrons to the anode. Complete 3D determination of particle tracks is possible in a device known as the time projection chamber, which is a combination of a multiwire proportional chamber and a drift chamber. Gas microstrip detectors [13] improve on the high count rate abilities and spatial resolution of a multiwire proportional chamber using closely spaced anode strips that are printed photolithographically on a rigid substrate. Gas microstrip detectors with anode spacings of 200 μm are able to achieve spatial resolution of under 40 μm and count rate abilities in excess of 10^6/(s mm^2).

82.3.5 Geiger–Müller Counters

82.3.5.1 Basic Design and Operation

A Geiger–Müller counter [14] is similar in construction to a proportional counter. Operating at electric fields above the proportional region, it is a controlled discharge counter that produces a large output pulse of a few volts for every ionizing event in the gas independent of the amount of energy deposited in the gas. The ultraviolet photons from the electron avalanche in a Geiger–Müller counter induce an electric discharge along the entire length of the anode wire. Gas mixtures used in Geiger–Müller counters generally contain a noble gas mixed with ethyl alcohol or halogen containing a vapor such as ethyl bromide, which acts as a quencher and eventually stops the discharge. After the discharge, the Geiger–Müller counter has a large dead time on the order of a few hundred microseconds, while the excess ionization is swept out of the chamber, during which it cannot register another pulse. If ethyl alcohol is the quenching agent, the gradual decomposition of molecules causes the counter characteristics to degrade with usage, and the counter eventually ceases to function. Counters with a halogen compound used as a quenching agent do not suffer from this defect; however, because halogens are

TABLE 82.3 List of Manufacturers for Gas-Filled Proportional Counters and Geiger–Müller Tubes

Aptec Engineering Ltd., East 50 B, Caldari Road, Concord, Ontario L4K 4N8, Canada, Phone: +1 (905) 660-5373, Fax: +1 (905) 660-9693
Canberra Industries, Inc., 800 Research Parkway, Meriden, CT 06540, Phone: +1 (203) 238-2351, Fax: +1 (203) 235-1347, http://www.canberra.com/
EG & G Berthold, P.O. Box 100163, D-75312, Bad Wildbad, Germany, Phone:+49 708 1177-140, Fax: +49 708 1177-100, http://www.berthold-us.com/
EG&G Ortec, 100 Midland Road, Oak Ridge, TN 37831, Phone: +1 (615) 482-4411, Fax: +1 (615) 483-0396, http://www.ortec-online.com/
Ludlum Measurements Inc., P.O. Box 810, 501 Oak Street, Sweetwater, TX 79556, Phone: +1 (915) 235-5494, Fax: +1 (915) 235-4672, http://www.ludlums.com/
Oxford Instruments Inc., 601 Oak Ridge Turnpike, P.O. Box 2560, Oak Ridge TN 37831-2560, Phone: +1 (615) 483-8405, Fax: +1 (615) 483-5891, http://www.oxford-instruments.com/Pages/home.aspx
Panasonic, Two Panasonic Way, Secaucus, NJ 07094, Phone: +1 (201) 392-6044, Fax: +1 (201) 392-4315, http://www.panasonic.com/
TGM Detectors Inc., 160 Bear Hill Road, Waltham, MA 02154-1075, Phone: +1 (617) 890-2090, Fax: +1 (617) 890-4711
Cardinal Health (Victoreen Instruments), 6000 Cochran Road, Cleveland, OH 44139, Phone: 800-850-4608, Fax: 440-248-9300, www.cardinal.com/rms

electronegative and capture electrons, such counters suffer from the problem of variable output signal delays. Since the Geiger–Müller counter pulses are several volts high, they require no further amplification for measurement purposes.

82.3.6 Proportional Counter and Geiger–Müller Counter Applications

Proportional counters and Geiger–Müller counters are often used in instruments for radiation protection and monitoring. Depending on the construction, they can be used for the detection of both α- and β-particles. The simple Geiger–Müller counter is perhaps more commonly used as a radiation monitor because it is easier to use and it has high efficiency, close to 100%. However, because of its relatively long dead time, a Geiger–Müller counter becomes less useful as a monitor in high radiation environments where particle rates exceed a few thousand counts per second. Also, a Geiger–Müller counter provides no energy information. For high counting rate applications or when energy discrimination is needed, it is better to use a proportional counter. Unfortunately, they can be cumbersome to use if an external gas supply is required. Geiger–Müller counters and proportional counter systems can be purchased from many different vendors (see Table 82.3). Prices vary from $100 to $500 for Geiger–Müller and proportional tubes and up to $2000 for complete systems and accessories (e.g., survey meters, scalers, rate meters, analyzers).

82.4 Scintillation Counters

82.4.1 Introduction

The use of scintillation light is one of the original methods used to detect charged particles. In 1911, Rutherford used an activated zinc sulfide screen to detect α-particles. The flashes of light produced in the screen were counted by a dark-adapted observer. Today, scintillation counters are usually made of plastic, liquid, or crystal materials, and the resulting signal is detected electronically. The principle of operation of these devices is that the energy deposited by charged particles moving through the scintillator material causes some of the molecules to be raised to an excited state. On returning to the ground state, the molecules emit light. The decay time of the light emitted depends on the fluorescence time of the scintillator. This can be very short, producing signal pulses as short as a few nanoseconds and thus allowing this type of detector to be used at very high counting rates. The light produced by the

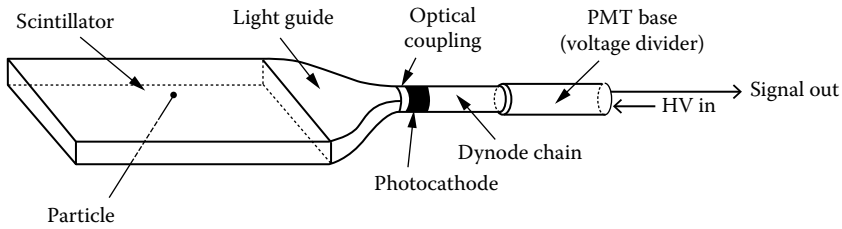


FIGURE 82.5 Typical scintillator system consisting of a scintillation medium, light detector, and light guide.

scintillator is usually detected using a PM. An appropriate choice of scintillation medium and detection method can produce a highly effective charged-particle counter with an efficiency of essentially 100%.

82.4.2 Scintillation Counters: A Detailed Description

A scintillation counter consists of the scintillation medium, a light detector, and a light guide to transfer the light from one to the other (see Figure 82.5). Information on each of the components of a scintillation detector follows [4].

82.4.3 Scintillation Medium

Plastic or liquid scintillators are usually used for the detection of charged particles. Crystal scintillation media such as sodium iodide or bismuth germanate (BGO) are normally used for photon detection or applications in which energy resolution is important. Plastic detectors are inexpensive and have proved to be a reliable way of detecting charged particles. Large areas can be covered using adjacent sheets of plastic scintillator. Such an array of detectors is known as a hodoscope. Primary light output from a plastic scintillator is at short wavelengths, below 400 nm. Light at these wavelengths has a short attenuation length in the plastic, so wave-shifting dyes are employed to convert the light to longer wavelengths. Thus, a plastic scintillator contains two or more additives such as polystyrene doped with *p*-terphenyl and tetraphenyl-butadiene. Proprietary scintillators are available in a large variety of shapes and sizes from companies such as Bicron Corporation. Plastic scintillators will darken upon exposure to radiation in excess of about 10^3 grays, which will lead to inefficiencies. Liquid scintillation materials, such as toluene, are useful for constructing large-volume detectors. Such devices are used in high radiation areas as liquid scintillators are inherently more radiation resistant. It is also possible to replace the liquid if radiation damage occurs. The main problem with this kind of device has been maintaining a leak-proof container and the resulting hazard associated with the toxicity of the liquid.

82.4.4 Light Coupling to Photon Detectors

The light produced in the scintillator is emitted isotropically. A portion of this light is trapped in the scintillator by total internal reflection. For a small detector, it is possible to couple the end of the scintillator directly to a phototube. For larger detectors (exceeding the width of the photodetector), a light guide made of a UV-transparent acrylic plastic is recommended. Acrylic plastic is widely available; Plexiglas by Rohm is one example. Light guides can be made of one piece of shaped plastic, but maximum efficiency can be realized using adiabatic (constant area) light guides that minimize light losses. These light guides are quite artistic (see Figure 82.6) and are made from acrylic strips heated and then bent to the appropriate shape. The cut edges are flame polished to produce a smooth surface for good internal reflection. To match large scintillators to a small area detecting device, a light guide known as a "Winston cone" [15] is used on the light detector. The shape of the Winston cone minimizes light losses.

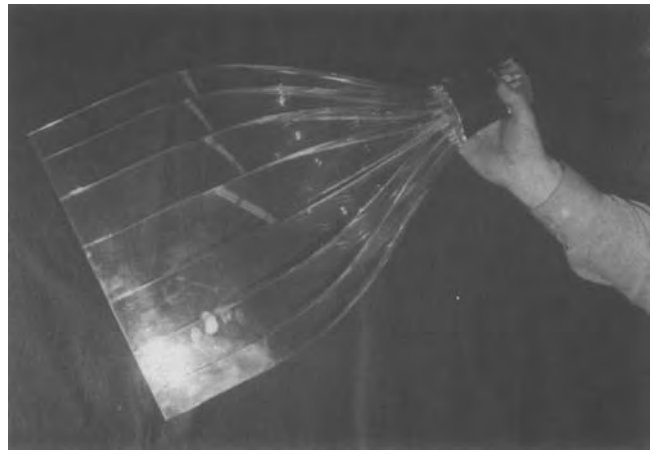


FIGURE 82.6 Photograph of adiabatic Plexiglas light guide. (Courtesy of Science and Technology Centre, Carleton University, Ottawa, Ontario, Canada.)

82.4.5 Light Detection and Readout

The standard method of detecting light from a scintillator is the use of PM tubes. These consist of a light-sensitive photocathode and an amplification stage. The photocathode is a thin layer deposited on the inside of a vacuum tube that has a good probability (>10%) of emitting an electron when struck by a photon. These electrons are electrostatically focused on a secondary electrode (dynode), which is the first element of the amplification structure. The voltage between the photocathode and this first dynode is sufficient to give the electrons enough energy to eject two to four electrons from the dynode. This process is continued through a series of dynodes resulting in a large gain, up to 10^8 in a tube with 14 dynodes. The output pulse is read out from a connection to the anode or the last dynode. The PM has connection pins corresponding to each dynode. Voltages are supplied to these via a resistive divider network contained in a "base" or "divider." The base requires one HV input and provides one or two signal outputs. Details of divider networks are available from PM tube manufacturers. Further information on the operation of phototubes can be found in Ref. [16].

In operation, PM tubes are sensitive to magnetic fields and it is usually necessary to provide magnetic shielding. A tube of a single layer of mu-metal is often sufficient to shield the PM tube from the Earth's magnetic field. For applications where there are higher ambient fields, a shield of layered tubes can be used consisting of two mu-metal shields and up to two soft iron shields. They should extend approximately one tube diameter beyond the PM tube. In applications where an axial field is unavoidable, a bucking coil can be inserted between the two iron layers. A suitable choice of coil current can counteract the axial field.

82.4.6 Scintillator Detector Construction

Assembly of scintillators, light guides, and PM tubes is usually done with optical cement. It is important in the preparation of the adhesive to remove air bubbles that would scatter the light. It is also possible to make demountable joint using optical grease or vacuum grease and a mechanical fixture. Such a joint can be used to simplify the replacement of phototubes. Light levels produced in the scintillator are extremely low; therefore, detector assemblies must be enclosed in light-tight envelopes. For plastic scintillators, this is most easily arranged by wrapping with aluminum foil and protecting the foil with black plastic sheets held in place with black adhesive tape. In large systems of counters, it is useful to install LEDs in each scintillator to monitor the operation of the counter. A system for powering the LED with a short duration pulse allows a test of the counter and readout system.

Components for scintillator systems can be purchased from many manufacturers, some of which are listed at the end of this section. Prices vary widely, depending on both the details of the detector and the quantity of components ordered. Typical prices would be $600 for a piece of scintillator 0.5 cm × 45 cm × 60 cm, $200 for a voltage divider, and $500 to $1200 for a 12-stage, 5 cm diameter phototube. Liquid scintillators are available for $145 per liter. Prices are in U.S. dollars. The cost of custom light guides is dominated by the available labor rate. There are some suppliers (such as Bicron) that will provide complete custom detectors consisting of scintillator, light guide, PM tube, and divider.

82.4.7 Operating Information

A MIP crossing a 1 cm thick scintillator will liberate about 10^4 photons. With typical scintillators and light guides, about 10% of these will reach the photocathode. The efficiency of photocathodes to produce photoelectrons is usually greater than 10%. Given PM tube gains of 10^8, this leads to 10^{10} electrons at the anode, allowing analog pulses of about 1 V produced. The pulses typically have a rise time of about 1 ns and a fall time of 5 ns. The trailing edge of the pulses can be shortened to approximately 2 ns by attaching a short grounded line to the output (clip line). The anode outputs are normally connected to a discriminator that produces a standard digital pulse for all analog pulses exceeding a fixed threshold voltage. More sophisticated circuits, such as constant fraction discriminators, can be used for best timing resolution. Using these devices along with a scintillator geometry that provides good light output, a timing resolution of the order of 300 ps (1σ) can be achieved. The excellent signal characteristics of scintillation counters allow them to count efficiently at rates exceeding 10 MHz.

Background counts in a scintillator are caused by cosmic rays and noise in the tube. An average rate from cosmic rays is 1 min^{-1} cm^{-2} sr^{-1}. Tube noise tends to have a lower amplitude than pulses resulting from the scintillator and will have a rate dependent on the choice of threshold. A counter should be checked to see if its count rate is consistent with cosmic ray backgrounds. High count rates are indicative of a light leak, which can be detected by shining a flashlight on the counter in a darkened room. Diligent use of aluminum foil and black tape will fix the leaks. In applications measuring particles that are energetic enough to pass through a counter, system noise can be reduced by using two layers of scintillator and putting their signals in coincidence. Noise can also be minimized by optimal setting of the tube HV and discriminator threshold. If the discriminator threshold is too low, then a large number of tube noise pulses are picked up.

82.4.8 Advanced Techniques

A plastic scintillation counter hodoscope can also be used to obtain position information. The spatial resolution of such a device is limited by the size of the individual detectors. For example, thin finger-shaped counters can be used for measuring beam profiles. Larger "paddles" 10 or 20 cm wide can be used to cover large areas. It is also possible to measure impact position along counter using phototubes at both ends of a long scintillator and measuring the difference in arrival time of the pulses at the two tubes. Higher-precision spatial resolution is possible using scintillating fibers with diameters of around 1 mm. Such a device has a large number of channels and requires segmented cathode phototubes or solid-state readout devices. One such device is the visible-light photon counter (VLPC), which is a produced by Rockwell International and is a development from a solid-state PM. To keep noise low, VLPCs are used at cryogenic temperatures.

82.4.9 List of Manufacturers

Table 82.4 shows a selection of potential suppliers of components for scintillator systems.

TABLE 82.4 List of Manufacturers for Scintillator System Components

	Scintillator	Bases/Dividers	Phototubes	Systems
Bicron	x	x		x
DEP Scientific	x			
Hamamatsu		x	x	
Philips		x	x	
POL. HI. TECH.	x			
Burle			x	
Thorn EMI		x	x	x
Bicron, 12345 Kinsman Road, Newbury, OH, 44065-9677, Phone:+1 (216) 564-2251, Fax: +1 (216) 564-8047	Hamamatsu Photonics, 325-6, Sunayama-cho, Hamamatsu City, 430, Japan, Phone: +81 53-452-2141, Fax: +81 53-456-7889, http://www.hamamatsu.com/	Thorn EMI, Bury Street, Ruislip, Middlesex, HA4 7TA, U.K., Phone: +44 1895 630771, Fax: +44 1895 635953		
Burle Industries, 100 New Holland Avenue, Lancaster, PA 17601-5688, Phone: +1 (717) 295-6000, Fax: +1 (717) 295-6096	Philips Photonics, 100 Providence Pike, Slatersville, RI 02876, Phone: +1 (401) 762-3800, Fax: +1 (401) 767-4493			
DEP Scientific, P.O. Box 60, 9300 AB Roden, the Netherlands, Phone: 3150-5018808, Fax: 3150-5013510	POL. HI. TECH. SRL KM44,400 S.P. Turanense 67061 Carsoli (AQ)—Italy, Phone: (986) 3997798, Fax: (986) 3995868			

82.5 Solid-State Detectors

Solid-state detectors have proved to be very versatile in the detection of charged particles. They have been particularly useful for energy measurements of α- and β-particles, light and heavy ions, and fission fragments in applications that include nuclear spectroscopy, radionuclide identification in both the environment and the laboratory, dosimetry, radiation level monitoring, and nondestructive testing using particle beam techniques. In high-energy physics, solid-state detectors are used mainly as high-precision tracking detectors in a particle counting mode. Most of the devices used in all these applications are based on silicon technology except for energy measurements of β-particles (>250 keV), where germanium-based detectors are also used.

82.5.1 Signal Generation

A charged particle moving through a semiconductor crystal loses energy according to the stopping processes described previously (Equation 82.1). The energy lost is spent promoting electrons from the valence band to the conduction band of the crystal. This leaves free electrons in the conduction band and free holes in the valence band. Under the action of an electric field, the free electrons and holes drift in opposite directions and give rise to an electric current that can be sensed by external electronics. The amount of charge produced is very small and the semiconductor must be depleted of majority carriers in order for the signal to be measurable. This is accomplished using a diode structure operated with a reverse-bias voltage. A profile of a "typical" diode detector is given in Figure 82.7. The size of the active region of the detector, the depletion region, is determined by the reverse bias applied and the residual fixed charges. In the case where the material has a very high resistivity (~500 Ω cm or more), the built-in field of the junction can be sufficient to generate a useful depletion layer without any reverse bias.

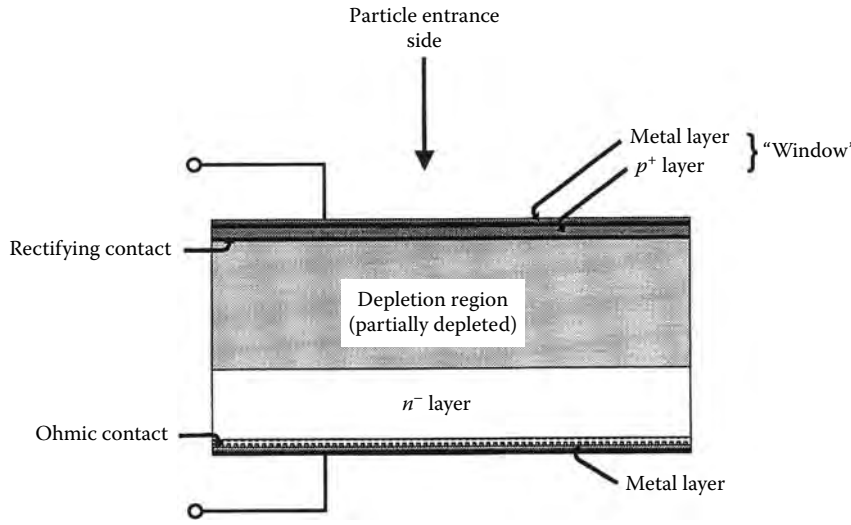


FIGURE 82.7 Solid-state diode detector consists of a thin current blocking layer, a sensitive volume of high-resistivity material, and a back ohmic contact. A near-uniform field may be established in the sensitive volume.

For a quantity of energy E_0 deposited through electronic interactions, the number of free electron–hole pairs created is $N_p = (E_0)/\varepsilon$, where ε is the average energy to produce one pair, a characteristic of the semiconductor, similar to the W value for gas detectors. Typical values of ε are small, 3.62 eV for Si at 300 K and 2.96 eV for Ge at 77 K. The variance of the number of pairs created is FN_p, where F is the Fano factor [17], which is of the order of 0.1 for Ge and Si. Hence, semiconductor materials are very efficient in converting radiation energy to an electric signal, and the pulse-to-pulse variation in the signal size is smaller than predicted by Poisson statistics, where $F = 1$. When all the charge carriers are collected (i.e., the electrons reach the anode side, while the holes reach the cathode side), a total of N_p electrons are available to the front-end electronics for amplification and shaping.

82.5.2 Categories of Detector

82.5.2.1 Device Details

Commercial solid-state detectors (i.e., silicon or germanium devices) are available from a number of manufacturers (see Table 82.5). The following silicon devices are available: silicon surface barrier detectors (SSBD), ion-implanted silicon detectors, diffused junction silicon detectors, *pin* photodiodes, and lithium-drifted silicon detectors (Si(Li)). Some of the manufacturers rename the different types using their own brand names. The first three types actually lend themselves to nearly the same applications: ion detection. Si(Li) detectors are mostly used for soft x-ray detection but can also be used for full-energy measurement of β-particles from 250 keV to 3 MeV. Germanium devices are available as lithium-drifted germanium detectors (Ge(Li)) and high-purity germanium (HPGe) detectors. As with the Si(Li) detector, the primary function of germanium detectors is for photon detection, but they can also be used to detect β-particles. Solid-state detectors are either *pn* or *pin* diodes. Silicon devices fabricated from a starting material of about 500 Ω cm (i.e., ion-implanted Si and diffused junction Si) have the *pn* structure. The SSBD, also made from Si of a few 100 Ω cm, is in fact a hybrid of Schottky junction and *pn* junction due to its gold-on-*n*-type Si interface as well as "*p*-like" surface defects introduced during manufacture. Devices that have a bulk resistivity in excess of 1 kΩ cm, whether it is due to a high-resistivity starting material (such as is used for pin Si and HPGe) or compensated material (Si(Li) and Ge(Li)), are *pin* junctions.

TABLE 82.5 Device Details

Type	Area (mm²)	Sensitive Thickness (μm)	Dead Layers[a] (nm)	Particles Detected	Prices 1996 US$	Manufacturers
SSBD	10–1000	10–5,000	80 225 (back)	p, α, β light and heavy ions	400–6,000	a, b, c, d
Si, ion implanted	Up to 3000	100–1,000	50	p, α, β light and heavy ions	300–3,000	a, b, c, d
Si pin photodiode	1–900	150–400	100	p, α, β light and heavy ions	10–300	e, f, g, h
Si(Li)	25–200	to 5,000	1000 Be +200 Si	β, 0.3–3 MeV	~7,500	a, b, c, d
HPGe	75–800	to 10,000	2500 Be +300 Si	β, 0.3–3 MeV	~9,000	a, b, c, d
Silicon strip	See manufacturer for details					c, d, g, i

Note: (a) EG&G Ortec (Oak Ridge, TN), (b) Oxford Instruments Inc. (United States), (c) Eurisys Mesures (France), (d) Canberra (Australia), (e) EG&G Optoelectronics (Canada), (f) Silonex Inc. (Canada), (g) Hamamatsu Photonics K.K. (Japan), (h) Centronic Inc. (United States), (i) VTT Electronics (Finland).

[a] Values in nm Si unless stated otherwise.

Many aspects dictate the choice of a detector for a given application. To measure the full energy of highly ionizing particles, it is important to keep the surface dead layer of the detector as thin as possible to limit the loss of undetected energy. The dead layer is the combination of the entrance metal electrode and the underlying, radiation insensitive, heavily doped top Si layer (usually *p* type). The required depletion region thickness is determined by the range in Si of the particles of interest. A sufficiently wide depletion region is required for full-energy measurements of ions or to obtain a large signal in the case of energetic particles, such as MIPs, which traverses the whole detector without stopping. Energy measurements of β-particles often require an active volume that is larger than can be achieved with a standard p–n junction, and so Ge(Li), Si(Li), or HPGe detectors are used. Varnish-coated detectors can be obtained to withstand exposures to rugged environments. Bakeable devices, which allow partial recovery from damage, or inexpensive and replaceable photodiodes can be used in situations of heavy radiation damage. For the detection of heavy ions, devices that can be operated at high electric fields (>15 kV cm^{-1}) can be used to overcome recombination effects. One of the most widely used Si detectors is the SSBD. It offers a thin entrance window of less than 100 nm. The diffused junction is the original type of Si detector and it has a rather thick dead layer of 250 nm Si. It is still useful in applications that require a light-insensitive device or for particle physics applications. The implanted junction Si detector results from the implantation of boron ions in an *n*-type Si wafer. It is the most recent addition to commercial Si detectors and it offers a thinner entrance window (less than 50 nm) and a larger surface area than SSBD (see Table 82.5). All these devices can be operated in either partially or fully depleted mode. The thickness of the back dead layer is only relevant to fully depleted, transmission-type detectors for energy loss (ΔE) measurements. Recently, high-quality, inexpensive Si *pin* diodes for visible and UV photon detection have been found to be suitable replacements for SSBD in many applications. The thickness of the dead layer is 100 nm, twice that of ion-implanted Si detectors, and the resistivity is as high as 2–8 kΩ cm for the more expensive versions.

82.5.3 Detector Systems

Three practical systems of increasing sophistication are considered next. The first setup, depicted in Figure 82.8, is used to measure the energy spectrum and the rate of incoming charged particles. One possible application is the measurement of the α-activity from a natural source or from a wipe test sample. This simple count rate measurement requires a *pn* silicon detector or photodiode, a discriminator, and a pulse counter. By adding a pulse analyzer, called an MCA (Figure 82.8), the same system can

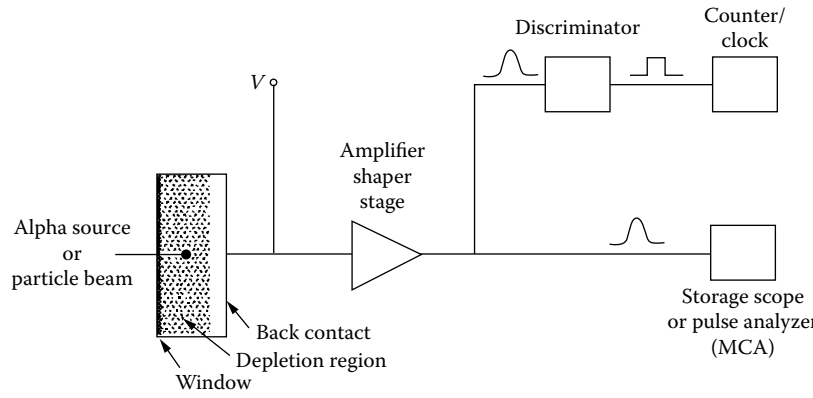


FIGURE 82.8 Using a pulse counter, this simple setup is used to measure the activity of a radioactive sample or the intensity of an accelerator beam. With the addition of a storage oscilloscope or an MCA, the energy deposited in the detector by each incoming charged particle may be measured.

be used to record the energy of the α-particles or the full energy of light ions, such as those encountered at an accelerator site where beam analysis of materials is performed. In this case, the suggested detector is a good quality SSBD or an ion-implanted detector with a thin window and operated at a voltage sufficient to generate a depletion region at least as wide as the range of the ions of interest. The second setup (Figure 82.9) shows a thin transmission-type (ΔE) detector (open front and back), in front of a thick, fully stopping, (E) detector. This configuration is called a ΔE–E telescope and it supplies a simultaneous measure of the energy loss and the full energy of particles going through both devices. This information then uniquely identifies the particle. Finally, Figure 82.10 presents a position-sensitive silicon strip detector for which the localization of a particle hit to better than 5 μm in one dimension is possible [18]. Such a detector is used in particle physics experiments. In many applications, combinations with other detector types, such as scintillators, are also possible.

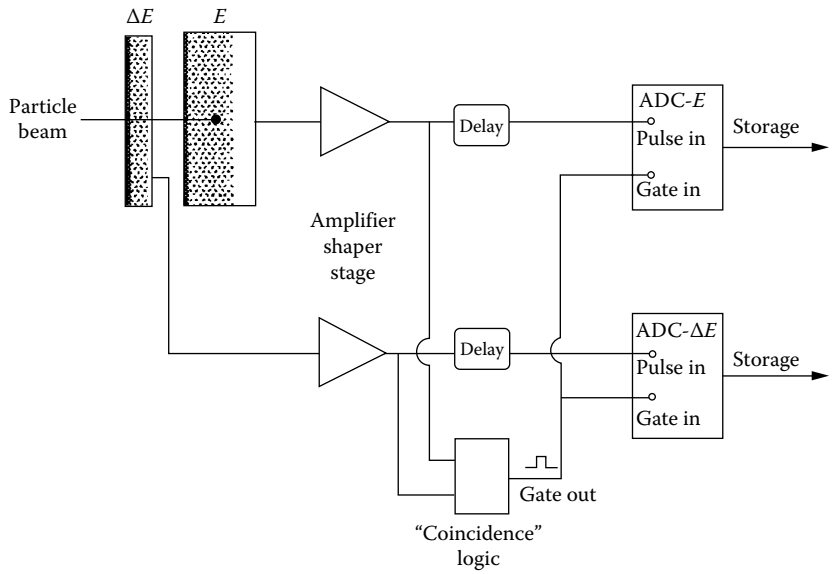


FIGURE 82.9 Knowledge of the amount of energy deposited in a transmission-type detector (ΔE) positioned in line with a full-energy measuring detector (E) allows the identification of a particle due to the unique nature of dE/dx and E combination.

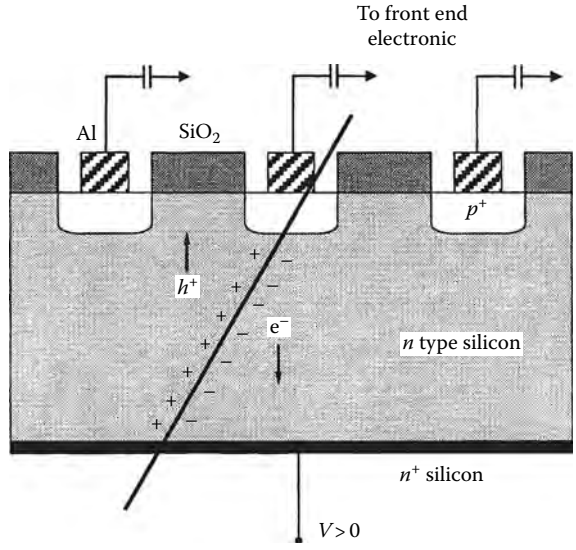


FIGURE 82.10 A silicon strip detector, such as the one depicted here, can be used to localize the hit position of a MIP.

Detectors with sensitive thicknesses of 100–300 μm are used in many nuclear spectroscopy and high-energy physics applications. For typical values of applied voltages, full electron collection occurs in less than 10 ns, while hole collection takes less than 20 ns. The amount of charge in a pulse is very small, of the order of 1.5 (10^6) e for the full energy of an α-particle and 100 e per μm of path in silicon of a MIP. Depending on the application, the signal-processing electronics can be optimized for low noise or fast response. In spectroscopy applications, where precise measurement of the total charge is important, one uses low-noise, charge-integrating preamplifiers, which offer output gains of the order of 1 mV fC^{-1}. The preamplifiers are optimized for specific ranges of detector capacitance. For detectors of 5 pF and smaller, noise values of less than 1 keV (equivalent energy loss signal in Si) can be achieved. In the case of preamplifiers optimized for larger detector capacitances (200–1000 pF), the noise increases with a slope of ~0.02 keV pF^{-1}. Preamplifiers produce a "step" pulse with amplitude proportional to the input signal charge. The rise time of the pulse varies with the detector capacitance, from 5 to 10 ns for <10 pF to 100 ns for 1000 pF. The pulse decay time is determined by internal circuit components and usually falls in the range of 50–300 μs. The signal is further amplified and shaped using an integrating–differentiating amplifier to provide a pulse of Gaussian shape whose width is of the order of 1 μs. This pulse duration limits the counting rate to much less than 10^5 cps before an unreasonably high counting dead time is encountered. In high-energy physics experiments, the signal-processing electronics coupled to position-sensitive detectors are optimized for a high counting rate capability and good timing resolution. Silicon-integrated circuits containing a large number of amplifiers are specially designed for such applications. The individual amplifiers provide large output gain (15 mV fC^{-1}) and output pulses whose width can be as small as 20 ns (baseline to baseline), at the cost of an increase in noise (6 keV Si equivalent). Count rates near 10^6 cps for every detection channel can be achieved with such systems.

82.5.4 Plasma Effects and Pulse Height Defect

Under some conditions, one may observe a loss in proportionality between the energy lost by the particle crossing the detector and the signal. This is called the pulse height deficit (PHD) and it is the sum of three components: E_W, E_N, and E_R, where E_W is the energy lost in the window or dead layer of the detector and E_N is the energy lost to nuclear scattering (nonionizing). The third term, E_R, is harder to define and is

due to a deficit in the charge collection. Ions heavier than the proton create plasma with a charge density as high as 10^{19} e^- cm^{-3} that shields the charge from the collecting field and causes charge loss through electron–hole recombination in the plasma to occur. The PHD increases roughly with the particle mass and energy and ranges from 1% to 3% for ions such as oxygen at 20 MeV.

82.5.5 New Structures

Modern silicon processing technologies allow the design of intricate structures that find applications in charged-particle detection, particularly in high-energy physics experiments. These include position-sensitive detectors, with strip electrodes or pixel segmentation, and large-volume silicon drift chambers. Front-end electronics can also be fabricated on the same wafer as the detecting structure itself. Close coupling provides noise reduction. The reader should be aware that new materials, such as GaAs, CVD diamond, natural diamond, cadmium telluride, HgI_2, and hydrogenated amorphous silicon, are being used in prototype detectors. These are beyond the scope of this chapter.

82.5.6 Manufacturers and Prices

A list of typical devices and pricing for each category of solid-state detector is given in Table 82.5. The names of some manufacturers are given, but the interested reader is also encouraged to consider other sources.

References

1. J.F. Ziegler, J.P. Biersack, and U. Littmark, *The Stopping and Range of Ions in Solids*, New York: Pergamon Press, 1985.
2. J.F. Ziegler, *SRIM: The Stopping and Range of Ions in Matter*, Yorktown, NY: IBM-Research, 1996 (computer software).
3. J.F. Ziegler, M.D. Ziegler, and J.P. Biersack, SRIM: The stopping and range of ions in matter, *Nuclear Instruments & Methods in Physics Research, Section B (Beam Interactions with Materials and Atoms)*, 268(11–12), 1818, 2010.
4. R. Fernow, *Introduction to Experimental Particle Physics*, Cambridge, U.K.: Cambridge University Press, 1986.
5. J.E. Bateman, M.W. Waters, and R.E. Jones, Spatial resolution in xenon filled MWPC x-ray imaging detector—A computing physics approach, *Nuclear Instruments and Methods*, A135, 235–239, 1976.
6. W.R. Nelson, H. Hirayama, and D.W.O. Rodgers, *The EGS4 Code System*, SLAC-Report-265, 1985.
7. J.A. Halbleib, R.P. Kensek, T.A. Melhorn, and G.D. Valdez, *ITS Version* 3.0—*The Integrated Tiger Series of Coupled Electron-Photon Transport Codes*, Albuquerque, NM: Sandia National Laboratory, 1992.
8. C.F. Powell, P.H. Fowler, and D.H. Perkins, *The Study of Elementary Particles by the Photographic Method: An Account of the Principal Techniques and Discoveries*, *Illustrated by an Atlas of Photomicrographs*, New York: Pergamon Press, 1959.
9. W.H. Barkas, *Nuclear Research Emulsions*, New York: Academic Press, 1963.
10. C. Plante and J. Hebert, The p/mc dependence of the grain density of tracks in the Fuji ET-7B nuclear emulsion, *Nuclear Instruments and Methods*, B108, 99–113, 1996.
11. W.R. Leo, *Techniques for Nuclear and Particle Physics Experiments*, 2nd edn., New York: Springer-Verlag, 1994.
12. N. Tsoulfanidis, *Measurement and Detection of Radiation*, New York: McGraw Hill, 1981.
13. A. Oed, Position-sensitive detector with microstrip anode for electron multiplication with gases, *Nuclear Instruments and Methods*, A263, 351, 1988.
14. R.W. Williams, Chapter 1.3—Gas filled counters, *Methods of Experimental Physics*, *Nuclear Physics A*, 5, 1961.

15. H. Hinterberger and R. Winston, Efficient light coupler for threshold cerenkov counters, *Review of Scientific Instruments*, 37, 1094, 1966.
16. *Photomultipier Tubes*, *Principle and Applications*, available from Philips Photonics, International Marketing, BP 520, F-19106, France.
17. G.F. Knoll, *Radiation Detection and Measurement*, New York: John Wiley & Sons, 1979.
18. A. Peisert, Silicon microstrip detectors. In: F. Sauli (Ed.), *Instrumentation on High Energy Physics*, Vol. 9, Singapore: World Scientific, 1992.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

83

Neutron Measurement

83.1 Introduction **83**-1
83.2 Detector Types **83**-2
Detectors Based on Elastic Scattering • Detectors Based on Capture Reactions • Detectors Based on Fission • Counters Based on Inelastic Scattering • Detectors Based on (n,p) or (n,α) Reactions • Standards
83.3 Efficiency Calculations and Measurements **83**-7
83.4 Summary **83**-10
Defining Terms **83**-11
References **83**-11

Steven M. Grimes
Ohio University

83.1 Introduction

Neutrons have characteristics that make them of particular importance in technology and research. Since they are uncharged, they are able to enter the nucleus at very low energy. Furthermore, the lack of energy losses through ionization permits deep penetration into materials.

This latter characteristic makes detection of neutrons more complicated than detecting protons or alpha particles. The energy of these charged particles can easily be determined by detecting them in ionization chambers, proportional counters, solid-state detectors, or scintillators, and ultraprecise measurements can also utilize the steering effect of magnetic fields. Detection efficiencies approach 100%. Neutrons, on the other hand, are not directly detected by these detectors, and some means of converting the neutron through a nuclear reaction to a charged particle or gamma ray must be used.

Nuclear reactions occur when neutrons interact with nuclei (including protons). When a neutron interacts with a nucleus, the following processes are possible:

1. Elastic scattering occurs if the neutron simply changes its direction without giving the nucleus any intrinsic excitation energy. The transfer of energy is analogous to that in a collision of two billiard balls and is governed by the laws of mechanics. The cross section for elastic scattering is usually large below 3 MeV and decreases at higher energies, although it is still significant at energies above 100 MeV.
2. Capture reactions are those in which the neutron is absorbed by the nucleus to form a heavier nucleus. In this case, the energy released comes in the form of one or more gamma rays. The capture cross section typically varies with energy E as $1/\sqrt{E}$ at energies below 1 MeV. At energies above 5 MeV, it is usually quite small.
3. Fission occurs in many cases if a heavy nucleus (mass > $A = 230$) is struck by a neutron. The nucleus divides into two smaller nuclei with a huge energy release ($E > 100$ MeV, where 1 MeV = 10^6 eV = 1.6×10^{-13} J).

4. Inelastic scattering occurs if a neutron transfers energy to the target nucleus and leaves it in an excited state. This energy is usually emitted in the form of a gamma ray, although in some cases positron–electron emission occurs and, in many cases, a number of low-energy gamma rays are emitted rather than one higher-energy gamma ray.
5. (n,z) reactions can also occur, where the neutron is absorbed by the nucleus and a charged particle (usually a proton or alpha particle) is emitted.
6. At high enough energies, multiple particles may be emitted. The (n,2n) reaction is particularly strong in the energy range $9 \leq E \leq 16$ MeV.

Each of these processes could, in principle, be used to detect neutrons. Unfortunately, in many cases, one cannot ensure that only one type of reaction occurs, and even when this condition is met, the resulting pulse spectrum does not always allow determination of the energy of the neutron.

83.2 Detector Types

83.2.1 Detectors Based on Elastic Scattering

Detectors based on elastic scattering are attractive because the elastic scattering cross section is large (usually about half of the total interaction cross section). A fundamental problem with elastic scattering is that the recoil energy as a fraction of the incident energy ranges from 0 (at 0° scattering angle) to $4A/(A+1)^2$ (for 180° scattering angle), where A is the mass number of the target. For $A = 1$ (hydrogen), this recoil energy ranges from 0% to 100%, but for even as light a target as carbon ($A = 12$), the maximum energy of recoil is less than 30%. The fact that a continuous range of energies is produced even from a monoenergetic neutron beam makes the efficiency difficult to calculate, since it depends on accurate knowledge of the cutoff threshold of the electronics. The same feature complicates determining the energy distribution of neutrons when a range of energies is present.

Hydrogen-containing counters are nonetheless frequently used in neutron physics. Below about 500 keV, the pulse height is usually too small to be useful, but above this energy, this type of counter is effective. Proportional counters or ion chambers filled with methane can be used in situations where fast timing is not required. Hydrogen-containing scintillators have poorer energy resolution but better timing information and are available in large volumes. Scintillators are subject to a threshold but can be made highly efficient ($\varepsilon > 50\%$) if the volume is sufficiently large. For either proportional counters or scintillators, gamma rays can produce background pulses through the Compton effect, but some liquid scintillators have the capability of identifying the two types of pulses produced by neutrons and gamma rays, respectively, allowing discrimination against gamma rays. The excellent timing characteristics of scintillators have resulted in extensive use of these detectors for time-of-flight spectrometers. In these spectrometers, the energy of the neutron is deduced by timing the neutron over a measured flight path. For typical flight paths ($L < 10$ m), a precision of a few nanoseconds is required.

The fundamental limitation of counters of this type is that the pulse height in the detector is not uniquely related to the neutron energy. If the detector is not used in a time-of-flight spectrometer, the energy of the neutron is difficult to determine. If a detailed study is done to characterize the pulse height distribution as a function of neutron energy, an estimate of the spectral shape can be obtained [1]. A related problem is that the efficiency is determined by an electronic cutoff, which is difficult to determine accurately. Calculations of the efficiency require a detailed knowledge of the pulse height as described earlier; both calculations and experimental determinations will be described in 83.3. A further complication to calculation of the efficiency is the role of carbon. Above 6 MeV, the (n,α) channel becomes possible, and above 7.6 MeV, the (n,3α) channel opens. Since the (n,p) cross section drops at higher energies, these reactions, plus ^{12}C(n,p) and ^{12}C(n,α), provide significant contributions to the efficiency. The database for

carbon cross sections is not as extensive as that for (n,p). Finally, detectors that include the entire angular distribution do not have directional sensitivity.

All of these drawbacks can be removed if only a limited portion of the angular range is sampled. This is the case in a recoil counter telescope. The spectrometer consists of a polyethylene radiator at one end of an evacuated cylinder. Near the other end are either one or, more typically, two detectors. Neutrons enter the telescope from the end holding the radiator. A small fraction of them produce recoil protons at small angles to the beam. A two-detector system is arranged so that the particles traverse both of the detectors. Requiring a coincidence in pulses reduces backgrounds. These detectors register protons scattered by the neutron beam at angles near 0° to the original beam. The recoil energy of such protons is $E_n \cos^2 \theta$, where E_n is the neutron energy and θ is the angle of recoil. This implies that for a telescope that subtends an angle θ around 0°, the energy range of recoil protons will be E_n to $E_n \cos^2 \theta$. It can be shown that the average energy is simply (1/2) E_n (1 + $\cos^2 \theta$). Obviously, as θ becomes small, the proton energy reflects the neutron energy. The smaller θ becomes, however, the lower the efficiency of the telescope. Typical telescopes of this type have efficiencies of 10^{-5} to 10^{-6}. It is obvious that only neutrons arriving along the axis will give protons of the right energy. Neutrons coming from the backward hemisphere will not be able to produce protons that traverse the spectrometer. The drawback of a low efficiency is partly compensated by the fact that the efficiency can be accurately calculated and the fact that the energy spectrum can be deduced.

A further drawback is that because the (n,p) cross section drops with neutron energy and the proton energy loss drops with proton energy, the optimal radiator thickness varies with neutron energy.

83.2.2 Detectors Based on Capture Reactions

Capture reactions have small cross sections at energies above 1 MeV. At lower energies, the cross section typically varies as $1/\sqrt{E}$, so it can be very large at low energies. Further, the capture process is exoergic, typically yielding about 7 MeV. This energy is usually divided among two to four gamma rays. The efficiency of gamma-ray detectors is usually much less than 100%, and detecting the gamma rays in coincidence requires large solid angles. It is essentially impossible to use the capture process to determine the energy of individual neutrons in a flux of neutrons of varying energies.

An alternative use of the capture process is in a passive detector. A foil of the appropriate material can be placed at the location where the flux is to be determined. The target must be one that produces a radioactive residual nucleus when a neutron is captured. Thus, ^{28}Si would not be appropriate, since ^{29}Si is stable, but ^{27}Al would be a possibility, since ^{28}A1 is radioactive. After the measurement is complete, the foil is removed and placed in a counting area. A detector then counts the beta particles or gamma rays emitted. Of particular importance is the choice of target to ensure an appropriate half-life. A very short half-life would allow the decay of the radioactive nuclei before they could be counted, while a very long half-life (>1 month) might require an excessive length of time to be counted.

The complicated energy dependence of the capture cross section makes use of the capture process mostly useful for relative measurements. If two measurements are made of spectra with the same relative distribution of neutrons with energy, a measurement of the ratio of activities will give the ratio of neutron intensities. An important advantage to this technique is that small foils can be used, allowing the determination of the flux in a small volume. Other radiation, such as gamma rays, usually does not interfere in such a measurement. An important disadvantage is that the flux is not obtained until after the counting is completed; this might be as long as a month after the measurement. An additional disadvantage has already been mentioned. If the relative energy dependence of the two spectra is not the same, the relative ratio of neutrons will not be correct.

Because the cross section for capture typically varies as $1/\sqrt{E}$ at low energy, these detectors are very effective at detecting neutrons below 1 keV, including thermal neutrons.

83.2.3 Detectors Based on Fission

The fission process is relatively unique, since an energy release of over 100 MeV occurs even when the incoming neutron has energy below 1 eV. Fission cross sections for gamma rays are small, so most of the pulses will be due to neutrons. There are two types of fission targets. Those with even proton number (Z) and even neutron number (N) have an energy threshold for fission. The cross section for fission is normally negligible below 2–3 MeV. Other targets for which either N or Z or both are odd have no threshold energy. These isotopes (e.g., ^{235}U) normally have a low-energy cross-sectional dependence of $1/\sqrt{E}$, giving a very large efficiency for neutrons below 100 keV. Fission detectors are designed either for $E_n > 2$ MeV (in which case ^{238}U would be an appropriate choice) or for the entire energy range, in which case the efficiency is highest at low energies.

Fission foils must be fairly thin to allow the escape of the fission fragments into the counting volume. This typically results in low efficiencies ($<10^{-6}$). Fission chambers also require the license to possess radioactive isotopes. The energy release in the fission process is so large that the pulse height in the counter cannot be used to infer the neutron energy. Reasonably good timing information ($\Delta t < 5$ ns) allows the use of fission chambers in time-of-flight spectrometers.

83.2.4 Counters Based on Inelastic Scattering

Inelastic scattering results in a transfer of energy from the neutron to the target nucleus through excitation of an excited energy state. This will then result in the production of one or more gamma rays. Gamma rays can be detected in germanium detectors with good energy resolution. Unfortunately, the efficiency of a spectrometer utilizing a germanium detector and inelastic scattering is quite small. This technique is being examined as a possible approach to identifying contraband materials in luggage, since both the neutrons and the gamma rays they produce are highly penetrating.

Although most gamma-ray decays occur rapidly ($<10^{-6}$ s), some gamma-ray decays are inhibited by angular momentum coupling and have half-lives longer than 1 s. These could be used in a detector based on inelastic scattering just as the capture reaction is used. Two nuclei with such states are ^{89}Y and ^{103}Rh. Unlike the capture reaction, such isomeric transitions will not be effective in detecting low-energy neutrons, since there is a threshold for populating such states (approximately equal to the energy of excitation, ignoring recoil effects).

83.2.5 Detectors Based on (n,p) or (n,α) Reactions

Detectors utilizing (n,p) or (n,α) reactions can be passive, in which a final product nucleus is radioactive, or active, in which a proton or alpha is detected. For a discussion of the passive type, see the section on detectors based on capture reactions (aforementioned text). Direct detection of the charged particle can occur in a proportional counter or a scintillator. As in the case of fission, many (n,p) and (n,α) reactions have an energy threshold, which prevents their occurring until this energy is exceeded. Others have no threshold (they are exoergic) and have low-energy cross sections that vary as $1/\sqrt{E}$. These targets and the reactions that occur are

1. ^{3}He(n,p)^{3}H
2. ^{6}Li(n,α)^{3}H
3. ^{10}B(n,α)^{7}Li

Reaction (1) is often utilized in a proportional counter filled with ^{3}He. The second reaction can be used in a scintillator made of glass that has a loading of up to 9% ^{6}Li. For Reaction (3), the gas BF_3 can be used in a proportional counter.

Each of these reactions is particularly effective in detecting neutrons below 1 MeV and especially thermal neutrons ($E \sim 0.025$ eV). Few charged particles are produced by neutrons at these energies from the counter housing or glass constituents, so the spectra are characterized by minimal interference from other materials. A common problem is that the low-energy neutrons that one seeks to measure are accompanied by a flux of thermal neutrons, which one does not wish to measure. In this case, the counter efficiency will be so high for the thermal neutrons that they may swamp the events one wishes to detect. An obvious solution to this problem is to use a thin layer of cadmium or boron to shield the detector, since these materials have very high thermal absorption cross sections. Each of these reactions is much less effective at energies above 10 MeV, since backgrounds due to charged particles in the housing or in the glass can be comparably large. Although counters utilizing these reactions can be approximately 100% efficient at thermal energies, they are often less than 5% efficient at energies above 3 MeV.

There is an interest in having a counter that can detect neutrons with virtually constant efficiency independent of energy. A counter that approximately meets this characteristic is the long counter. This counter consists of a BF_3-filled proportional counter embedded in a moderating cylinder of polyethylene, paraffin, or similar material. Those neutrons directly incident on the BF_3 chamber with very low energies will be detected with approximately 100% efficiency, while neutrons with $E_n > 1$ MeV will be detected with much smaller efficiency (<10%). By surrounding the BF_3 counter with polyethylene or paraffin, the higher-energy neutrons can be slowed down and the net efficiency increased, although some may be captured in the moderating material. Very low-energy neutrons will be more likely to be captured in the polyethylene or paraffin, so the detection efficiency for these neutron energies is reduced. Study of the detailed energy dependence of the efficiency of such counters has resulted in a design that has certain cavities cut in the shielding on the side from which the neutrons enter. This allows some of the lower-energy neutrons a shorter path to the counter. The fact that neutrons are detected after moderation means that the counter cannot give precise timing information, nor can it be used to determine neutron energy spectra. It also cannot determine the neutron flux over a small cross-sectional area but gives an average over 40–100 cm^2. Further, it will have a slightly different efficiency for neutrons coming from different directions. A sample design is shown in Figure 83.1. Note that since the counting element needs only to have an efficiency that rises sharply at low energy, a long counter can be designed using a ^3He ion chamber instead of a BF_3 tube.

Other (n,z) reactions do not have positive Q values (are not exoergic). These have thresholds and, in fact, usually have small cross sections for a few MeV beyond the threshold. An example is $^{27}Al(n,\alpha)^{24}Na$. This reaction cannot occur for neutrons of less than 3 MeV and has a very small cross section below 4.5 MeV. It does lead to a radioactive final nucleus and can be used as a passive detector. Unlike the detectors based on the (n,γ) reaction, it will not detect low-energy neutrons.

An alternative reaction that can be used in a passive detector is the (n,2n) reaction. This typically has a threshold of about 8 MeV but has a large cross section once the threshold is exceeded. Note that just as is found for (n,γ) or (n,z) reactions, not all nuclei reached in (n,2n) reactions are unstable.

A passive foil can be placed adjacent or behind another foil to obtain more spectral information. If a first foil is counted for the (n,γ) events, a second, which utilizes (n,p) or (n,2n), could provide an estimate of the neutron flux at energies above the region where the (n,γ) reaction is effective.

83.2.6 Standards

The National Institute of Standards and Technology (Gaithersburg, MD) maintains two standard neutron sources [2]. NBS-I is based on ^{226}Ra-Be(γ,n) and NBS-II is a similar backup source. The activities of NBS-I and NBS-II are known to about 0.85%. These can be used with a flowing bath containing ^{55}Mn to compare the production of ^{56}Mn in the $^{55}Mn(n,\gamma)$ reaction between NBS-I or NBS-II and an external source. The estimated error on such a determination of source strength for an external source is about 3.5% (2σ) [2].

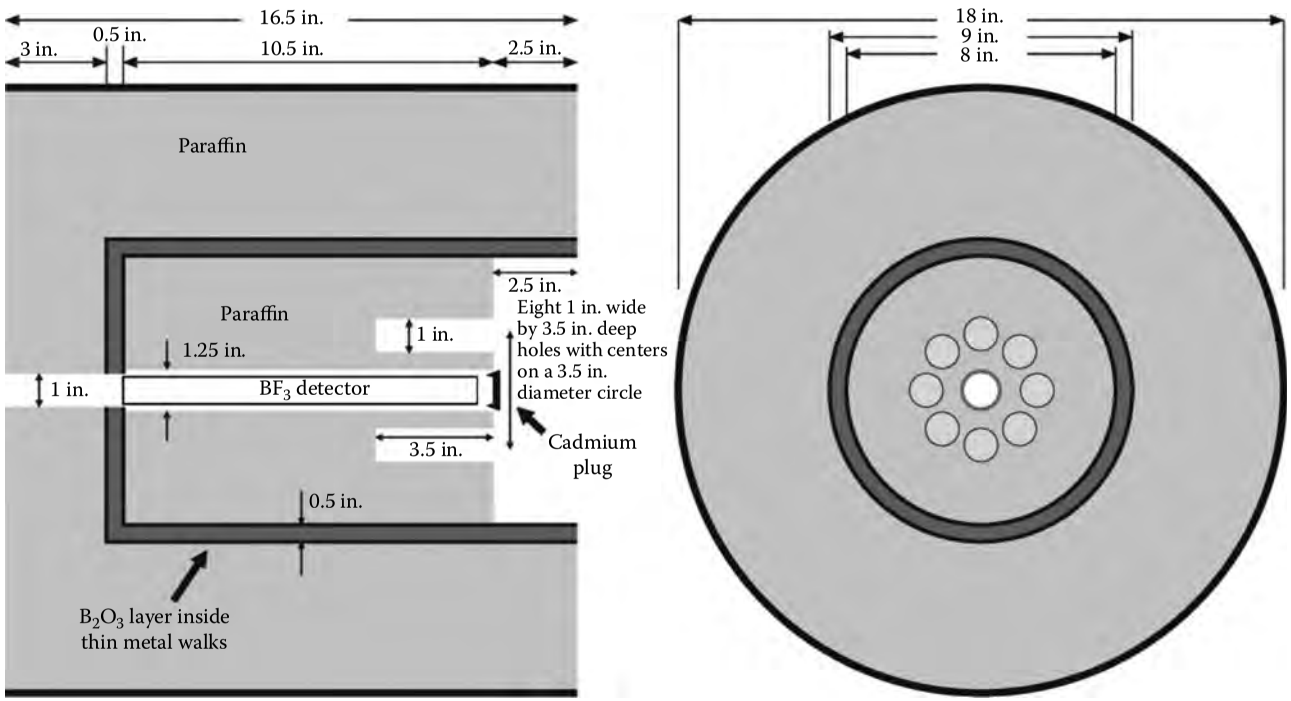


FIGURE 83.1 Cross section of a typical long counter. Neutrons incident from the right are scattered by the paraffin, causing them to lose energy. The BF_3 counter is efficient primarily for neutrons below 1 keV. The holes and the B_2O sections balance the efficiency as a function of energy. Cadmium has a very large $1/E$ capture cross section. Use of the cadmium filter suppresses efficiency for thermal neutrons, while removing it adds the contribution due to thermal neutrons. This counter is at the Edwards Accelerator Laboratory of Ohio University.

83.3 Efficiency Calculations and Measurements

Detectors based on n-p scattering have an efficiency that depends on the n-p elastic cross section. For a proportional counter or scintillator, an approximate expression for the efficiency is

$$\varepsilon = \left(1 - e^{-n\sigma_H L}\right)\left(1 - \frac{B}{E_n}\right) \tag{83.1}$$

This expression involves the parameters n, the number of hydrogen atoms per cm^3; σ_H, the n-p elastic cross section; B, the electronic cutoff (in energy equivalent units); and E_n, the neutron energy. This expression ignores multiple scattering, so it would not include pulses produced by two scattering events in succession. It also does not include carbon interactions. For energies below 4.4 MeV, carbon interactions are almost entirely elastic and do not result in absorption or pulses large enough to detect. Particularly at energies above 10 MeV, carbon interactions not only remove neutrons from the beam but also, through (n,α) or (n,3α) events, can produce pulses that can be detected. Calculating the efficiency in this energy range can be done with Monte Carlo codes, but the efficiency will depend on knowledge of the ^{12}C(n,α) and ^{12}C(n,3α) cross sections as well as the relation between alpha pulse heights and proton pulse heights as a function of energy.

Recoil telescopes have an efficiency that has the form:

$$\varepsilon = \frac{n\sigma(0°)LA}{d^2} \tag{83.2}$$

Here

n is the number of hydrogen atoms per cubic centimeter
$\sigma(0°)$ is the n-p elastic cross section at 0° (in units of cm^2/steradian)
L is the thickness of the radiator
A is the area of the detector
d is the distance between radiator and detector

Note that this assumes that $n\sigma L \ll 1$ and that θ is small. A corrected version based on the assumption that the n-p cross section is isotropic in the center of mass has

$$\sigma(0°) = \frac{\sigma_{np}}{\pi} \tag{83.3}$$

where σ_{np} is the total elastic cross section and, if θ is not sufficiently small, $\sigma(0°)$ is replaced by

$$\frac{\sigma_{np}}{2\pi}(1 + \cos\theta) \tag{83.4}$$

Note that the efficiency can be increased by increasing L or A or reducing d. All of these changes degrade the energy resolution.

Efficiency of passive detectors is a function of many variables. The total activity of a sample after being struck by N_n neutrons will be

$$n\sigma L N_n \frac{\tau}{T}\left[1 - e^{-T/\tau}\right] \tag{83.5}$$

where

n is the number of atoms per cm^3
σ is the appropriate cross section (e.g., for n, γ)
L is the thickness of the foil
T is the time over which the irradiation took place
τ is the average lifetime of the radioactive nuclei produced

This expression assumes that $n\sigma L \ll 1$ and that the source strength was uniform over the time T.

If the counter has an area A, is at a distance d from the activated foil, has an efficiency ε_r, and measures the activity from a time t_1 after the bombardment to a time t_2, the efficiency ε_n will be

$$\varepsilon_n = (n\sigma L)\frac{\tau}{T}\left(1-e^{-T/\tau}\right)\left(\frac{A\varepsilon_r}{4\pi d^2}\right)\left(e^{-t_1/\tau}-e^{-t_2/\tau}\right) \tag{83.6}$$

Each of the factors in parenthesis is less than one, and the typical efficiency of such a monitor is 10^{-4} to 10^{-6}.

Fission chambers have an efficiency of

$$\varepsilon_n = n\sigma_f L \tag{83.7}$$

where

n is the number of atoms per cm^3 in the foil
σ_f is the fission cross section
L is the thickness of the foil

At low energies, σ_f may be large enough so that the aforementioned expression is greater than 0.1. In that case,

$$\varepsilon_n = \left(1-e^{-n\sigma L}\right)\frac{\sigma_f}{\sigma} \tag{83.8}$$

where σ is the total cross section for neutrons.

The efficiency for a lithium- or boron-containing scintillator is

$$\varepsilon_n = \left(1-e^{-(n_1\sigma_1+n_2\sigma_2)L}\right)\left[\frac{n_1\sigma_1}{n_1\sigma_1+n_2\sigma_2}\right] \tag{83.9}$$

where

n_1 is the density of lithium or boron atoms per cm^3
σ_1 is the (n,α) cross section
n_2 is the density of atoms other than lithium or boron in the scintillator
σ_2 is the average absorption cross section for these atoms

If $\left(n_1\sigma_1+n_2\sigma_2\right)L \ll 1$, this becomes

$$\varepsilon_n = n_1\sigma_1 L \tag{83.10}$$

Note that in each case where a term of the form $(1 - e^{-n\sigma L})$ appears, the average of this quantity is not $(1 - e^{-n<\sigma>L})$ if σ has energy fluctuations over an energy bin that are substantial. Here, $<\sigma>$ is the average cross section over the energy bin. This limit is particularly important at low energies, where isolated resonances cause large fluctuations in σ and where the limit of small $n\sigma L$ may not be appropriate. In this case, the expression for the efficiency should be evaluated separately for individual energies in the bin and the resultant efficiencies averaged.

Efficiencies calculated using the aforementioned formulas are likely to be reliable for proton recoil telescopes and fission chambers. This is because the detector as normally used is insensitive to gamma rays. The fission process produces a large pulse, and background pulses from counter constituents are likely much smaller. Further, the thin fission foil makes multiple scattering negligible.

For counters involving the use of (n,p) and (n,α) reactions or elastic scattering, multiple scattering effects may not be negligible. Further, constituent atoms of the counter or support may contribute background.

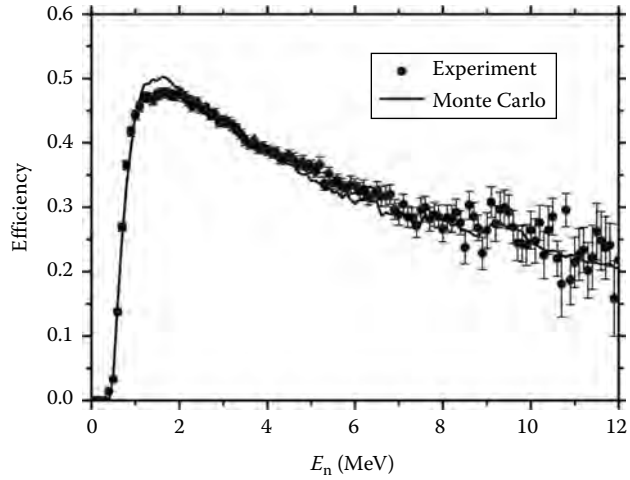


FIGURE 83.2 A calibration of a hydrogenous scintillator using a fission chamber containing ^{252}Cf. The solid line is a calculation based on Monte Carlo. (Reprinted from *Nucl. Instrum. Meth. Phys. Res. A,* 399, Total characterization of neutron detectors with a ^{252}Cf source and a new light output determination, Kornilov, N.V., Fabry, I., Oberstedt, S., and Hambsch, F.J., Copyright 2009, with permission from Elsevier.)

Figure 83.2 shows a calibration [1] of a hydrogenous scintillator using a fission chamber containing ^{252}Cf. This isotope fissions spontaneously. Because its fission spectrum has been previously characterized, the fission chamber in detecting a fission pulse serves as a calibrated neutron source. A detector that is capable of giving an accurate time for a pulse can be used to determine the efficiency as a function of energy. At low energies, the efficiency of the counter is smooth and has the shape expected for

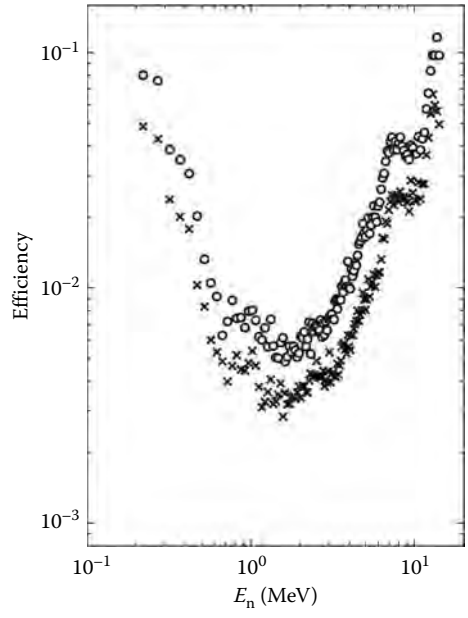


FIGURE 83.3 A calibration for a lithium glass detector using the continuous neutron spectrum produced by deuterons incident on a stopping target of aluminum. At low energy ($E < 1.5$ MeV), the efficiency is largely due to ^{6}Li(n,α) reactions, while above this energy, the observed spectrum and efficiency suggest that the (n,n$'$ γ) reaction on the constituents of glass provides most of the efficiency.

proton recoil pulse. At energies above 8 MeV, the statistical precision of the measurement gets worse because of the small fission neutron fraction above this energy. It appears that there is some structure in the efficiency. This may be due to carbon-produced pulses or absorption by air between the chamber and detector.

A similar calibration for a lithium glass detector has been carried out [3] using the continuous neutron spectrum produced by deuterons incident on a stopping target of aluminum and is shown in Figure 83.3. This spectrum was measured with a fission chamber. The observed efficiency is striking for two reasons. The different efficiency values are found for detectors that are similar in size and lithium content. Thus, their calculated efficiencies would be identical. Furthermore, the efficiency has the behavior expected for a detector based on the ^{6}Li(n,α) reaction only at energies below 1 MeV. Above this energy, it appears that neutron inelastic scattering from silicon and oxygen in the glass yields gamma rays that are detected and enhance the detection efficiency. This illustrates the value of experimental measurements of detector efficiency.

83.4 Summary

A summary of the features of various counters is presented in Table 83.1. More exhaustive treatment of the subject of neutron detectors can be found in Refs. [4,5]. Table 83.2 lists some suppliers of neutron detectors.

TABLE 83.1 Characteristics of Various Neutron Detectors

Detector Type	Gives Energy Spectrum	Small Size	Directional Information	Insensitive to Gamma Rays	Detects Neutrons below 500 keV	Timing Information[a]	High Efficiency[b]
Hydrogen-containing counters							
Proportional counter	No	Yes	No	Somewhat	No	No	Yes
Scintillator	No[c]	Yes	No	Liquid scintillator can reject gamma rays	No	Yes	Yes
Proton recoil telescope	Yes	Yes	Yes	Yes	No	Yes	No
Non-hydrogen-based counters							
Passive (n,γ) detector	No	Yes (smaller than others)	No	Yes	Yes	No	No
Fission detector	No[c]	Yes	No	Yes	Yes	Yes	No
Passive inelastic-scattering detector	No	Yes	No	Largely	No	No	No
Detectors based on (n,z) reactions							
Passive detector	No	Yes	No	Yes	Yes	No	No
Scintillator (lithium glass)	No[c]	Yes	No	No	Yes	Yes	Yes
Long counter	No	No	No	Yes	Yes	No	Yes

[a] Timing information is considered to be available if the time is determined to within less than 10^{-6} s.

[b] High efficiency means $\geq$1%.

[c] Can give energy information when used in time-of-flight spectrometer.

TABLE 83.2 Commercial Suppliers of Neutron Detectors

Supplier	Product
Saint-Gobain Crystals (Bicron), 17900 Great Lakes Parkway, Hiram, OH 44234-9681, Tel.: (440) 834-5600, Fax: (440) 834-7680, scintillation@saint-gobain.com, www.detectors.saint-gobain.com	Plastic scintillators Liquid scintillators with pulse shape discrimination Assemblies, phototubes
Advanced Measurement Technology, Inc. (ORTEC), 801 South Illinois Avenue, Oak Ridge, TN 37831-0895, Tel: (865) 482-4411, Fax: (865) 483-0396, ortec-info@ametek.com, www.ortec-online.com	He-3 detector and electronics Li-6 detector and electronics
Reuter Stokes, GE Energy, 8499 Darrow Road, Twinsburg, OH 44087, Tel.: (330) 425-3755, www.gepower.com	He-3 proportional counters Fission chambers
REXON, 24500 Highland Point, Beachwood, OH 44122, Tel.: (216) 292-7373, Fax: (216) 292-7714, www.rexon.com	Plastic scintillators Liquid scintillators with pulse shape discrimination Assemblies, phototubes
N. Wood Counter Laboratory, Inc., P.O. Box 509, Chesterton, IN 46304, Tel.: (219) 926-3571, Fax: (219) 926-3571, www.nwooddetectors.com	Boron triflouride (BF_3) gas tubes

Defining Terms

Compton effect: A process by which a gamma ray scatters from an electron on an atom. As a result, the electron is given a recoil velocity, producing ionization and yielding a pulse in an ionization chamber or proportional counter.

Efficiency: The efficiency of a neutron detector is its probability of detecting a neutron incident on the detector. Efficiency is obviously a number between 0 and 1 and generally depends on neutron energy.

Ionization chamber: A detector that collects the charge produced by ionization as a charged particle passes through the chamber.

Isomeric transition: Gamma-ray decay of a nucleus in which the rate of decay is greatly slowed by restriction imposed by angular momentum coupling.

Proportional counter: A detector that collects the charge produced by ionization with linear amplification as a charged particle passes through the chamber.

***Q* value**: A parameter expressing whether a nuclear reaction leads to a final state of higher final mass ($Q < 0$) or of lower final mass ($Q > 0$) than the initial state. Elastic scattering reactions have $Q = 0$.

Scintillator: A detector that converts ionization energy to a short light pulse.

Time-of-flight spectrometer: A detection device for neutrons in which a neutron detector is placed a known distance from the neutron source and the energy of the neutron is deduced from the time the neutron takes to traverse this path.

References

1. N. V. Kornilov, I. Fabry, S. Oberstedt, and F. J. Hambsch, Total characterization of neutron detectors with a ^{252}Cf source and a new light output determination, *Nucl. Instrum. Meth. Phys. Res. A,* **399**, 226 (2009).
2. J. M. Adams, *Nucl. Instr. Meth. B* **213**, 218 (2004).
3. T. N. Massey, S. Al Quraishi, C. E. Brient, J. F. Guillemette, S. M. Grimes, D. Jacobs, J. E. O'Donnell, J. Oldendick, and R. Wheeler, *Nucl. Sci. Eng.* **129**, 175 (1998).
4. J. B. Marion and J. L. Fowler (eds.), *Fast Neutron Physics, Part I,* Interscience Publishers (New York, 1960).
5. G. F. Knoll, *Radiation Detection and Measurement,* John Wiley & Sons (New York, 1979).



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

84

Dosimetry Measurement

84.1 Radiation Dosimetry Quantities and Units **84**-1
84.2 Thermoluminescence Dosimetry **84**-3
84.3 Ionization Chamber Dosimeters **84**-7
84.4 Film Dosimetry **84**-8
84.5 Track-Etch Dosimetry **84**-10
Alpha Track Dosimetry • Fast Neutron Track Dosimetry
84.6 Bubble Dosimetry **84**-11
84.7 Electronic Personal Dosimeters **84**-12
Defining Terms **84**-12
References **84**-13
Further Information **84**-13

Brian L. Justus
Naval Research Laboratory

Mark A. Miller
Naval Research Laboratory

Alan L. Huston
Naval Research Laboratory

Radiation dosimetry is a field of radiation detection devoted to the quantitative measurement of the physical changes that occur in matter upon exposure to ionizing radiation. Radiation dosimetry is performed on a routine basis to ensure the occupational safety of workers when there exists a risk of radiation exposure. Such routine personal dosimetry is used to monitor and limit the long-term occupational exposure to workers and to assist with the assessment of the dose received in the event of an accidental exposure. Some occupations that have an associated risk of radiation exposure include medical personnel performing clinical x-ray diagnostic and radiotherapy procedures and military and civilian personnel involved in power plant maintenance and operation. Radiation dosimetry is also an important tool in a wide range of environmental monitoring and industrial processing applications. The scope of this chapter is limited to discussion of the most widely used and commercially available technologies for dosimetry of ionizing radiation for personal protection; however, it should be recognized that many of the same technologies are used with equal effectiveness for other dosimetry applications.

84.1 Radiation Dosimetry Quantities and Units

The goal of radiation dosimetry is to quantify the amount of energy that is deposited in matter upon interaction with ionizing radiation. Ionizing radiations fall into three categories: charged particles such as beta and alpha particles, neutral particles such as neutrons, and electromagnetic radiation such as gamma rays and x-rays. The mechanism and the efficiency of the energy deposition depend on the type and the energy of the radiation as well as on the composition of the absorbing material. In particular, the biological consequences to living tissue following radiation exposure are quite dependent on the type of radiation. In order to provide a useful measure of the biological damage that might be expected to occur upon energy deposition in tissue, a system of units and standards has been developed that takes into account the differing biological effectiveness of different types of radiation. Over the years, these units and standards have undergone extensive revision, and this process is still evolving. However, the units and standards in current use have been defined in detail in recent reports [1,2] published by

the International Commission on Radiation Units and Measurements (ICRU) and the International Commission on Radiological Protection (ICRP) and are the subject of a recent review article [3]. A brief summary of the fundamental units is provided as follows.

The quantity of primary interest in radiation dosimetry is the amount of energy that is absorbed per unit mass. The amount of energy imparted to a volume of matter is the difference between the sum of the energies of all the charged and uncharged particles entering the volume and the sum of the energies of all of the charged and uncharged particles leaving the volume, plus any change in the rest mass of the matter. The mean energy imparted per unit mass is defined as the absorbed dose, D, expressed in (SI) units as J kg^{-1}. The special name for the unit of absorbed dose is the gray (Gy). If the mass exposed to radiation is at a point in tissue, knowledge of the amount of absorbed dose alone is not sufficient to judge the biological effectiveness of the charged particles producing the absorbed dose. A quality factor, Q, is used to weight the absorbed dose to provide an estimate of the relative hazard of different energies and types of ionizing radiation. Q is a dimensionless factor that converts absorbed dose to dose equivalent, H, at a point in tissue. The dose equivalent has the same (SI) units, J kg^{-1}, as the absorbed dose, but the special name for the unit of dose equivalent is the sievert (Sv). If the amount of absorbed dose received on average over a specific tissue or organ is of interest, the quality factor is called the radiation weighting factor, w_R [1]. Table 84.1 lists values of w_R for various types and energies of radiation. The equivalent dose, H_T, in a tissue or organ is defined as the average absorbed dose in the given organ due to a particular type of radiation and is scaled by w_R. Thus, the total equivalent dose in a specific organ, T, due to exposure by each radiation type, R, is given by:

$$H_T = \sum_R w_R D_{T,R} \tag{84.1}$$

The equivalent dose also uses the special unit of sievert. For example, if an organ is exposed to neutrons of energy less than 10 keV, and receives an absorbed dose of 10 Gy (10 J kg^{-1}), the equivalent dose is determined, using Table 84.1 and Equation 84.1, to be 50 Sv (50 J kg^{-1}). Since the biological effectiveness of the exposure also depends on the specific organs targeted, dimensionless tissue weighting factors, w_T, have been defined that reflect the probability of damage resulting from a given equivalent dose. The effective dose, E, for an exposed individual is then the sum of the weighted equivalent doses for all tissues and is given by the expression

$$E = \sum_T w_T H_T \tag{84.2}$$

TABLE 84.1 Radiation Weighting Factors for Various Types of Radiation

Type of Radiation	Quality Factor
Photons, all energies	1
Electrons and muons, all energies	1
Neutrons, <10 keV	5
Neutrons, 10–100 keV	10
Neutrons, >100 keV–2 MeV	20
Neutrons, >2–20 MeV	10
Neutrons, >20 MeV	5
Protons, other than recoil protons, >2 MeV	5
Alpha particles, fission fragments, and heavy nuclei	20

Source: From ICRP Publication 60, Ann. ICRP 21(1–3), 1990 *Recommendations of the International Commission on Radiological Protection*, International Commission on Radiological Protection, Pergamon Press, Oxford, U.K., 1991.

The effective dose also uses the special unit of sievert. It should be pointed out that despite the adoption of nomenclature and definitions for equivalent dose and effective dose, for all practical purposes, the quantities measured and reported for dosimetry applications are the dose equivalents defined in terms of the quality factor, Q. A further quantity for reporting purposes is the personal dose equivalent, $H_p(d)$, the dose equivalent in soft tissue, at an appropriate depth, d, below a specified point in the body. For weakly penetrating radiation, such as beta particles, a depth of 0.07 mm, or $H_p(0.07)$, is used. Eye doses are reported for 3 mm depth ($H_p(3)$), and deep doses are reported for 10 mm depth ($H_p(10)$). The personal dose equivalent is typically measured with a dosimeter that is covered with the appropriate thickness of tissue equivalent material. The dosimeter is worn on the surface of the body and is calibrated in a phantom. A phantom is a medium, such as poly(methyl methacrylate), that mimics the attenuating characteristics of the human body.

Another unit for absorbed dose, found in older publications and in current use in the United States, is the rad. One rad equals 100 erg g^{-1}. Similarly, another unit for dose equivalent is the rem. To convert from this system to the international standard (SI) units, the following conversions are used: 1 Gy = 100 rad and 1 Sv = 100 rem.

When an ionization chamber is exposed to x-ray or gamma-ray photons, the amount of ionization produced in the air inside the chamber is called the exposure, X. The (SI) unit of exposure is C kg^{-1} and the special unit is called the roentgen, R. Originally, R was defined as the amount of x-ray or gamma radiation that produces 1 esu of charge of either sign in 0.001293 g of air (the mass of 1 cm^3 of air at STP). The conversion from R to C kg^{-1} is $1\ R = 2.58 \times 10^{-4}$ C kg^{-1}. When air is exposed to radiation, the average energy required to form an ion pair is 33.97 J C^{-1}. For an exposure of 1 R, the absorbed dose in air is 8.77×10^{-3} J kg^{-1}.

84.2 Thermoluminescence Dosimetry

Thermoluminescence (TL) dosimetry is perhaps the most widely used and cost-effective technique for radiation dosimetry. It is for many organizations the technique of choice for routine monitoring of occupational radiation exposure. TL dosimetry is also widely used in medicine to determine patient exposure as a result of x-ray diagnostic procedures and cancer radiotherapy treatments. The dose ranges of interest for these applications can be roughly defined as 0.1–1 mGy for personal dosimetry applications, 1–100 mGy for clinical x-ray diagnostics, and 1–10 Gy for medical radiotherapy applications. TL dosimeter (TLD) phosphors are commercially available that exhibit a linear dose response and are capable of accurately measuring the absorbed dose for all the aforementioned applications and doses. The radiation-sensitive element of a TLD is a small quantity, typically less than 100 mg, of an inorganic crystal doped with metal impurities, called activators. The activators provide the crystal with the energy storage capacity as well as the luminescent properties that are required for the crystal to function as a thermoluminescent phosphor upon exposure to ionizing radiation. The details of the energy storage and TL mechanisms, despite intense research over many decades, are not well understood. Complex, long-range, many-body interactions between activator sites and their immediate environment may be involved. Despite this mechanistic uncertainty, the TL phenomenon has been used reliably by following well-established heuristic procedures and by referring to a simple and intuitive model. This model assumes that the activators provide point defects, known as traps and luminescence centers, in the crystal lattice. Upon exposure to ionizing radiation, electrons and holes are captured in metastable states at the trap centers by local potential minima until such time that the electrons and holes are thermally stimulated to overcome the electric potential. The electrons and holes can then recombine, with the emission of photons, at the luminescence centers. It is apparent from this model that TL dosimetry is a passive, integrating technique. Further, the technique shows no dose rate dependence, and for most applications, the TLD phosphors can be reused many times.

A comprehensive review of the properties and applications of a wide range of TLD phosphors has been published recently [4]. The characteristics of the phosphors that are of most interest for dosimetry applications include the sensitivity and the dynamic range of the TL response, the response to different

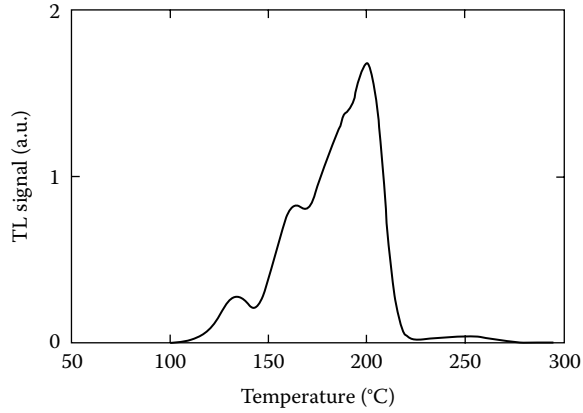


FIGURE 84.1 TL glow curve of LiF phosphor activated with trace amounts of Mg^{2+} and Ti^{4+} (TLD-100). The phosphor was exposed to a 2 Gy dose of ^{60}Co gamma rays and the glow curve was read using a Harshaw 2000 A/B TLD reader.

types of radiation over a broad energy spectrum, and the fading characteristics, reusability, and the reproducibility of the response. The TL response of the phosphor to ionizing radiation is indicated by measurement of the glow curve. A glow curve is a plot of the TL intensity versus temperature, measured as the phosphor is heated. A representative glow curve of a commercially available LiF TLD is shown in Figure 84.1. The glow curve usually exhibits one or more glow peaks that correspond to the release and recombination of electron and hole traps of differing trap depth and stability. TL dosimetry is best accomplished using phosphors having glow peaks in the range 200 °C–250 °C. Peaks in this temperature range are necessary to provide room-temperature trap storage times of several months. Although all traps fade with time, rapid fading of the stored information occurs from traps having glow peak temperatures below 200 °C, effectively precluding practical dosimetry. Glow peaks having temperatures greater than 250 °C require quite high temperatures (>300 °C) to release all the filled traps. In this case, blackbody radiation from the hot phosphor and other ancillary hot matter within the sample readout chamber (sample planchet, hot gas) interferes with and can easily overwhelm the TL signal due to actual trap recombination. The response of TLD phosphors as a function of absorbed dose must be carefully calibrated for all doses of interest. This is necessary because the response deviates from linear above a threshold dose, different for each phosphor, and, ultimately, the TL signal saturates and the phosphors suffer irreversible damage at very high doses. The TL response per absorbed dose as a function of energy, for low-LET electromagnetic radiation (gamma ray and x-ray), is generally fairly uniform. The efficiency for higher LET radiation can vary and should be calibrated as a function of energy. Since personal dosimetry applications require the measurement of doses absorbed by human tissue, tissue equivalent TLDs (effective atomic number closely matches that of tissue) are attractive because the energy dependence of the TL response closely matches that of tissue.

The most commonly used TLD phosphors are lithium fluoride (LiF), calcium fluoride (CaF_2), lithium borate ($Li_2B_4O_7$), calcium sulfate ($CaSO_4$), and aluminum oxide (Al_2O_3) activated with trace quantities of transition metal or rare earth metal ions. TLD phosphors are available in a variety of forms, including powders, compressed chips, Teflon-impregnated disks, single crystals, extruded rods, and thin films. TLD phosphors have been developed that respond primarily to gamma radiation and x-radiation, while others have been developed that, in addition, respond strongly to thermal neutrons. Unfortunately, there are no TLD phosphors having neutron sensitivity without gamma sensitivity. Therefore, in a mixed field of neutrons and gamma rays, TLDs are used in pairs, with one TLD having primarily gamma-ray sensitivity and the other having gamma-ray and thermal neutron sensitivity. The difference in the TL signal between the two TLDs of such a pair is assumed to represent the thermal neutron dose. Neutron sensitivity is enhanced, for example, in LiF phosphors by incorporating a high concentration of ^6Li since ^6Li

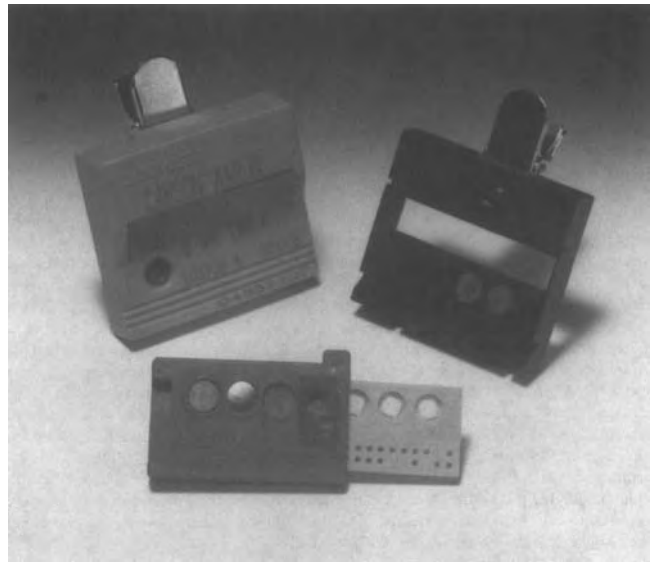


FIGURE 84.2 TLD body badge. A standard TLD badge contains three LiF phosphor chips. A fourth chip and a fast neutron track-etch foil can be added if required. (Courtesy of Landauer, Inc., Glenwood, IL.)

has a much higher thermal neutron cross section than ^{7}Li (TLD-600 contains 96% ^{6}LiF while TLD-700 has essentially no ^{6}Li). The manufacturers of TLDs exert considerable effort to maintain the uniformity of their products from batch to batch, a task of some difficulty since the characteristics of phosphors are a sensitive function of not only the composition but also the details of the manufacturing process. Thus, the properties of TLD phosphors of similar composition can vary considerably, and both the manufacturers of phosphors as well as the dosimetry service industries typically test and sort chips into groups based on their measured parameters. A representative commercially available TLD badge is shown in Figure 84.2. This standard badge contains three LiF TLDs and is capable of measuring gamma- and x-ray doses from 0.1 mGy to 10 Gy and energetic beta radiation doses from 0.4 mGy to 10 Gy. These badges are typically returned to the manufacturer on a monthly or quarterly basis for readout.

The actual practice of TL dosimetry requires some care to achieve accurate and reproducible results. For example, TLD-100 traditionally has been a very widely used phosphor, due to its tissue equivalence, that can yield accurate and fairly sensitive dose information. However, the energetics of the trap population in TLD-100, as well as the trap dynamics and intertrap communication with time, are extremely complex phenomena. Reproducible dosimetry requires that the sensitivity remain unchanged after each use. This is accomplished by careful adherence to a detailed protocol that may include a pre-irradiation anneal (e.g., 400 °C for 1 h followed by 80 °C for 10–24 h), a postirradiation anneal to 125 °C, and finally heating to readout the TL signal. It is common for individual TLD users to develop their own unique protocols for annealing and reading TLD chips. Excellent dose reproducibility (<5%) can be expected with care by following a well-defined protocol and using a set of individually calibrated TLD chips on a single, well-maintained TLD reader.

TLD instrumentation is commercially available from a number of suppliers. The choice of instrumentation depends largely on the form of the dosimeter phosphor chosen (compressed chip, Teflon disk, etc.), the desired sensitivity, and the degree of automation desired. The key element in any TLD reader is the heating method. The heating method should provide for controlled, reproducible heating and, if speed of processing is important, rapid cooling. Contact heating of the TLD on a planchet through which an electric current flows is a widely used heating technique. Because the surface of the TLD is not perfectly smooth, more uniform heating can be performed using noncontact methods, such as heating with a stream of hot gas or heating with infrared light. Laser heating is another noncontact

method that permits rapid heating (>1000 °C s^{-1}) using small quantities of TLD phosphor, yet permits very high signal-to-noise ratios [5]. Regardless of the heating method, blackbody radiation from the metal planchet and/or the TLD chip must be attenuated by colored glass and dielectric coated optical filters. The signal-to-noise ratio is also improved by the elimination of spurious TL signals that arise due to static charges or contamination on the surface of the TLD. Spurious TL signals are reduced by purging the sample compartment with a low flow of inert gas, such as dry nitrogen. Light detection is typically accomplished with a photomultiplier tube, using either a DC or photon-counting mode of operation. The glow curve is collected using any of the heating methods described, and the TL signal can be reported as either a glow peak height or as the integrated area of a glow peak.

The conventional TLD technology described previously requires bulk heating of the TLD to high temperatures (400 °C–600 °C) in order to read the stored dose information. Once the TLD has been read, the dose information is permanently erased. If any malfunction in the equipment occurs during readout, the data cannot be recovered. Alternative TLD technologies have been developed that do not completely erase the dose information and permit additional, successive readouts of the absorbed dose. This characteristic can eliminate accidental loss of dose information and can also provide, if desired, an archival record of exposure. Two such dosimeters, utilizing laser readout methods, are commercially available. These dosimeters function based on optical phenomena known as phototransferred TL (PTTL) [6] and

TABLE 84.2 Vendors of Personal Dosimetry Products and Services

Vendor	Product or Service
Bicron NE, 6801 Cochran Road, Solon, OH 44139, Tel.: (216) 248-7400, http://bicronusa.com/	TLD phosphors, TLD readers
Panasonic Industrial Comp., Applied Technologies Group, 2 Panasonic Way/7E-4, Secaucus, NJ 07094, Tel.: (201) 348-5339, http://www.panasonic.com/	TLD phosphors, optically heated TLD readers, electronic dosimeters
Landauer, Inc., 2 Science Road, Glenwood, IL 60425-1568, Tel.: (708) 755-7000, http://www.landauer.com/	TLD, film, track-etch services, COSL dosimeter
ICN Dosimetry Service, P.O. Box 19536, Irvine, CA 92713, Tel.: (714) 545-0100	TLD, film services
Eberline Dosimetry Service, 5635 Jefferson St, NE, Albuquerque, NM 87109-3412, Tel.: (505) 345-9931, http://www.eberlineservices.com/rad.htm	TLD readers, TLD service
Dosimeter Corporation of America, 5 Eastmans Road, Parsippany, NJ 07054, Tel.: (800) 322-8258	Pocket dosimeters, electronic dosimeters
Toshiba Glass Co., Ltd., 3583-5 Kawashira Yoshida-Cho Haibara-Gun, Shizuoka-Ken, Japan, Tel.: 0548-32-1217, http://www.toshiba.com/	RPL glass dosimeter
Siemens Environmental Systems Limited, Sopers Lane, Poole, Dorset, BH17 7ER, United Kingdom, Tel.: 44 1202 782779, http://www.siemens.com/entry/cc/en/	Electronic dosimeter
SAIC Commercial Products, 4161 Campus Point Court, San Diego, CA 92121, Tel.: (619) 458-3846, https://www.saic.com/	Electronic dosimeters
RADOS Technology, Inc., 6460 Dobbin Road, Columbia, MD 21045, Tel.: (410) 740-1440	TLD dosimeters, electronic dosimeters
SE International, Inc., 436 Farm Road, Summertown, TN 38483-0039, Tel.: (615) 964-3561, http://www.seintl.com/	Pocket dosimeters
Bubble Technology Industries, Hwy. 17, Chalk River, Ontario, Canada, K0J 1J0, Tel.: (613) 589-2456, http://www.bubbletech.ca/	Bubble dosimeters
Apfel Enterprises, 25 Science Park, New Haven, CT 06511, Tel.: (203) 786-5599	Bubble dosimeters

radiophotoluminescence (RPL) [7]. A commercial dosimeter (Landauer, Inc.) based on the phenomenon of PTTL is also known as a cooled optically stimulated luminescence (COSL) dosimeter. In preparation for readout of the dose information, the dosimeter is first cooled to liquid nitrogen temperature. It is then irradiated with ultraviolet laser light to phototransfer electrons from deep traps to shallow traps. The TL that results from the release of electrons from these very shallow traps, as the dosimeter is allowed to warm to room temperature, provides a very sensitive measure of the absorbed dose. Since it is not necessary to heat the dosimeter to high temperatures, the trapped charges are not fully annealed and the process can be repeated several times. A commercial RPL dosimeter, also known as the flat glass dosimeter (Toshiba Glass Co., Ltd.) uses a silver-activated phosphate glass that is read out by a pulsed ultraviolet laser. The manufacturer employs a special luminescence analysis protocol since the glass exhibits a prompt luminescence even in the absence of ionizing radiation exposure. RPL dosimeters are nevertheless capable of performing sensitive and reproducible dose measurements. New optical dosimetry technologies, based on the phenomenon of room-temperature optically stimulated luminescence (OSL), are being developed [8,9] and promise to provide both superior sensitivity and multiple readout capability.

Companies that provide TLD and other personal dosimetry products and services are listed in Table 84.2. This list is not intended to be comprehensive, nor is it intended to represent an endorsement by the U.S. government.

84.3 Ionization Chamber Dosimeters

Ionization chamber-based radiation dosimeters are among the most widely used and the most accurate instruments available for determining radiation exposure and absorbed dose. The operational principle of the ionization chamber is based on the formation and the collection of ion pairs that result from the interaction of energetic charged particles that pass through gases contained in a well-defined volume and electric field. Several types of ionization chambers are available, including real-time radiation field monitors and integrating devices that accumulate dose information for extended periods of time. Real-time devices monitor the radiation-induced currents, while integrating devices record changes in a static electric field. A wide array of ionization chambers, intended for applications in medical physics, equipment calibration, and personal dosimetry, are available. Small volume (<1 mL), precision ionization chambers are used by medical physicists for radiotherapy applications. These chambers are generally tissue or air equivalent and, while very accurate, have low sensitivity. Larger volume (<1000 mL) ionization chambers are used for applications such as field standardization and diagnostic x-ray machine calibration that require high sensitivity in addition to excellent accuracy. For personal dosimetry applications, compact integrating ionization chambers, about the size of a ballpoint pen, termed "pocket chambers" are used. Some pocket chamber dosimeters provide a method for direct visual readout based on the electrostatic deflection of a charged quartz fiber relative to a calibrated scale. Initially, the dosimeter is fully charged, representing the zero dose setting. Exposure to a radiation source causes a loss of charge on the fiber, and spring tension causes the fiber to partially return to its uncharged position. A lens system incorporated into the dosimeter in conjunction with a scale can be used to determine the absorbed dose. A commercially available direct-reading dosimeter is illustrated schematically in Figure 84.3. These instruments can measure gamma- and x-ray doses from a fraction of a milligray up to several gray. Greater accuracy is possible using indirect readout, pocket chambers that utilize precision electrometers to measure the change in the potential between the cathode and anode following exposure to radiation. These dosimeters are particularly useful for measuring low-level (<mGy) exposures. Another type of ionization chamber, used primarily for radon detection, utilizes the voltage drop across an electret element to determine radiation dose (Rad Elec, Inc.). An electret is an electrically poled insulator material, such as Teflon, that retains surface charges for long periods of time. Ion pairs produced by radiation exposure are separated and collected, resulting in a net decrease in the electret charge. Changes in the electret voltage can then be calibrated in terms of radiation exposure.

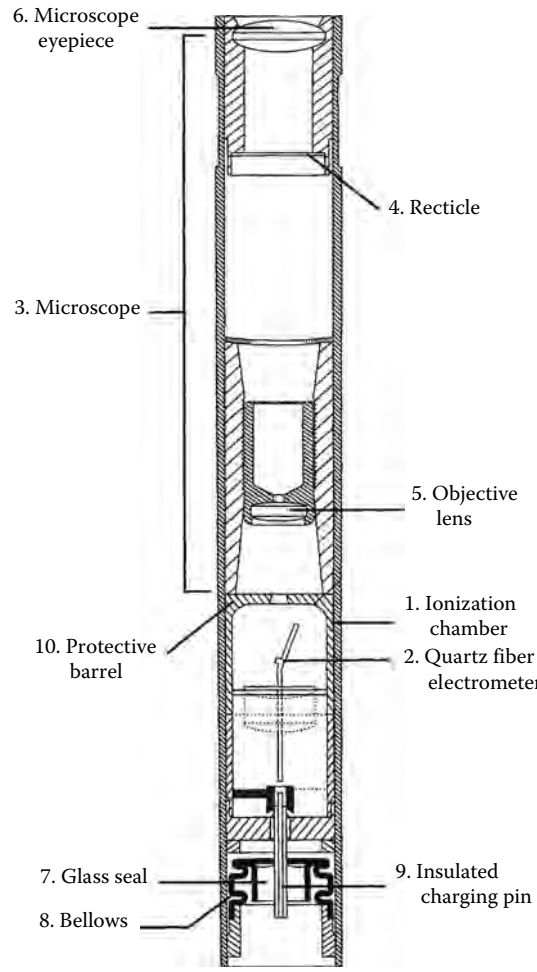


FIGURE 84.3 Schematic of direct-reading pocket ionization chamber. The radiation dose is determined by measuring the deflection of a movable quartz fiber electrode. The deflection of the fiber is determined optically by projecting the image of the fiber onto a reticle using a microscope objective. (Courtesy of Dosimeter Corp. of America , Rolla, ND.)

84.4 Film Dosimetry

Photographic films have been used for many years for radiography and personal dosimetry applications. However, due to the greater convenience, superior sensitivity, faster turnaround, and reusability of TLD personal dosimeters, film emulsion personal dosimeters have in many cases been displaced by TLD dosimeters. A notable example of this trend is the dosimetry program of the U.S. Navy, which has turned exclusively to TLD dosimeters. Despite this, film remains useful for specific applications that require analysis of an image or a charged particle track, and they remain the most convenient dosimetry method available if a permanent record of the exposure is desired. Radiographic films are sensitive to a variety of radiation sources, including photons (x-ray and gamma ray), charged particles (electrons, protons, and alpha particles), and neutrons (slow and fast). A major problem associated with the use of radiographic films for dosimetry purposes is that the energy response is not at all flat. A number of approaches have been used to flatten the response, including the incorporation of a scintillator dye such as *p*-terphenyl in the emulsion [10]. This approach has the effect of flattening



FIGURE 84.4 Film badge for personal dosimetry. (Courtesy of Landauer, Inc., Glenwood, IL.)

the response curve dramatically between 0.1 and 1.0 MeV, as well as increasing the sensitivity of the film. More commonly, in a manner similar to that used in multiple chip TLD badges, a series of metal filters, applied to different portions of the film, is used to provide energy discrimination. The film is then analyzed according to the sensitivity of each film segment for a particular radiation energy range. Figure 84.4 is a photograph of a commercially available film dosimeter badge that incorporates six absorbing filters to discriminate beta, gamma, and x-radiation. The dose range that can be measured is 0.1 mGy–5 Gy for gamma radiation and x-radiation and 0.4 mGy–10 Gy for energetic beta particles. Radiographic films consist of a thin layer of gelatin, approximately 10–20 μm thick, containing silver halide microcrystals, coated onto a polymer or glass substrate. Exposure to a source of radiation sensitizes the microcrystallites, creating a latent image that darkens visibly following development of the film. The radiation dose is determined by measuring the extent of the darkening with a densitometer. The densitometer measures the optical density (OD) of the film by comparing the transmission, T, of light through the exposed film with the transmission, T_0, of an identical film that has not been exposed to radiation. The optical density is a logarithmic function, $\mathrm{OD} = \log(T_0/T)$, and can be measured accurately for values up to approximately 3. Developed film typically has a background OD of about 0.1 due to scattered light. Undeveloped film is quite sensitive to environmental factors such as light, temperature, and, in particular, moisture and must be carefully protected while in use. Film dosimeters are available that can also perform thermal and fast neutron dosimetry. The response of film to thermal neutrons is enhanced by the use of a film converter such as a 0.5–1 mm thick Cd foil. The foil absorbs thermal neutrons, yielding gamma rays that interact with the emulsion. Nuclear track emulsions are films that are used for dosimetry of fast neutrons in a manner quite similar to the track-etch detectors described in the next section. Nuclear track emulsions are generally much thicker than radiographic film emulsions and contain higher concentrations of silver halide microcrystals. Recoil protons transfer energy to the microcrystallites in the emulsion, creating latent images of the tracks of the protons. The range of charged particles in the emulsion is generally quite small, compared to photons, so that practically all of the energy is deposited in the film along the tracks. Nuclear track emulsions are no longer widely used due largely to the extremely high threshold, ~0.7 MeV, required for track visualization. Fast neutron dosimetry with film badges is most often accomplished by using a track-etch detector in conjunction with the radiographic film emulsion.

84.5 Track-Etch Dosimetry

Solid-state track detectors exploit the damage that occurs in dielectric materials upon exposure to ionizing charged particles. Energy is deposited in the material along a track defined by the trajectory of the ionizing particle. The damage is manifested as pits that develop on the surface of the material upon etching by chemical and electrochemical techniques. The pits are clearly evident using a simple optical microscope, and the track density can be estimated using either manual or automated pit counting methods. The track density is a direct measure of the radiation dose (for many particles, there is one track formed per particle), and track-etch detectors are very effective and sensitive dosimeters. At present the most important track-etch dosimetry applications utilize organic polymer films for alpha track detection (radon monitoring) and fast neutron personal dosimetry. The principal attractive features of track-etch detectors are their extreme simplicity and low cost, good sensitivity, near tissue equivalence, small size, passive operation, and insensitivity to fast electrons and gamma rays. Track-etch detection is a favorite dosimetry technique of graduate students worldwide because of its simplicity; however, commercially available track-etch detectors are manufactured with extreme care to provide users with uniform polymer films and reproducible dose response.

84.5.1 Alpha Track Dosimetry

Commonly used solid-state track detectors that possess sensitivity suitable for alpha particle dosimetry include plastics such as cellulose nitrate, polycarbonate, and polyallyl diglycol carbonate [11]. Since track-etch detection is a surface phenomenon, care must be exercised to avoid scratching the surface of the polymer film, as any surface imperfections can result in false-positive signals. The surfaces of the polymer film should also be protected from soiling since dust, dirt, grease, water, or any other foreign material can attenuate the alpha particles and prevent track registration in the detector. Polymer track-etch detectors do not fade significantly at room temperature; however, if the temperature is elevated such that the polymer softens, then the track damage can be annealed and etching will not yield a pit. Typically, radon track dosimeters are allowed to accumulate the alpha dose due to radon and its daughters for extended periods (up to a year) before being returned to the manufacturer for determination of the track density and the corresponding alpha particle activity. Dosimetry of other radioisotopes that emit alpha particles can also be performed; however, the response of the detector is dependent on the isotope detected. For example, alpha track detectors have been used for monitoring plutonium and americium contamination in soil and provide a convenient and cost-effective radiation survey method for decontamination and decommissioning activities [12].

84.5.2 Fast Neutron Track Dosimetry

The performance of polymer track-etch detectors used for personal dosimetry of fast neutrons is typically enhanced by the use of adjacent layers of a hydrogen-rich polymer, known as a proton radiator [13]. Fast neutron dosimetry is made possible by the neutron–proton elastic scattering process that occurs in the track-etch detector and the radiator. The recoil protons produced by this process are responsible for generating the damage tracks in the polymer film and thus provide an indirect measurement of the neutron flux. Although the polymers used for track-etch detection of neutrons, such as CR-39, contain hydrogen and provide for neutron–proton elastic scattering, the use of a proton radiator, such as polyethylene (having a much higher concentration of hydrogen), significantly improves the performance of the dosimeter [14]. The energy spectrum of the protons that leave the radiator and reach the detector depends on both the neutron energy and the thickness of the radiator. In addition, the total number of protons emitted depends on the neutron energy and the radiator thickness. When the thickness of the radiator equals the range of the most energetic proton emitted, then the condition of protonic equilibrium is met. Under this condition, the greatest number of protons are emitted, and the detector has the greatest sensitivity. However, at protonic equilibrium, the sensitivity also displays the greatest variation in the energy response. A flatter energy response can be obtained using thinner radiators that avoid the condition of protonic equilibrium. This is obtained, however, at the expense of the sensitivity. Track-etch detection method can be used

for dosimetry of fast neutrons at nuclear power plants and particle accelerators and due to exposure from unmoderated isotope sources such as californium-252 and americium-241 beryllium.

84.6 Bubble Dosimetry

Another method for fast neutron personal dosimetry is bubble detection. Bubble dosimeters are manufactured using a clear, elastic polymer gel as a host matrix, supporting a dispersion of nanometer-scale superheated droplets of a liquid, such as Freon. Neutrons incident on the polymer gel generate recoil protons in an elastic scattering process, followed by deposition of the proton energy along a track, in a process similar to that described in the discussion of track-etch detectors. However, in the track-etch detectors, the tracks must be visualized by a separate etching procedure. The charged particles in a bubble dosimeter deposit their energy directly into the superheated droplets, causing the droplets to effectively explode into much larger gas bubbles. This process, which can be readily heard, is immediate, generating bubbles that can easily be seen by the naked eye, provided the bubble diameter exceeds a critical diameter required in order for the bubble to persist stably in the gel. The fast neutron dose is determined by counting the total number of bubbles generated. Bubble counting can be performed manually, by visual inspection of the dosimeter, or by automated means. Commercial bubble dosimeters are available that use either optical or acoustic automated counting techniques. Dosimeters have been developed using droplet materials that have differing energy thresholds, thereby providing the dosimeters with limited spectroscopic capabilities. Bubble dosimeters have been developed that are compensated to provide a temperature-independent response. Fast neutron bubble dosimeters have several advantages in common with track-etch detectors, including good sensitivity, near tissue equivalence, passive operation, small size, and insensitivity to thermal neutrons and gamma radiation. Additional advantages include immediate bubble visualization (no need for a separate track developing step), limited reusability, sensitivity adjustability, a flat dose response over a wide energy range, utilization of the entire volume of the dosimeter, and angular response superior to that of track-etch detectors. Bubble dosimeters are more expensive than track-etch dosimeters and are best utilized when the energy of the neutron field is known. Figure 84.5 is a photograph of commercially available bubble dosimeters, before and after exposure to neutrons. A comprehensive review of bubble dosimeters, including new developments, has been published recently [15].



FIGURE 84.5 Bubble dosimeters for personal dosimetry of fast neutrons. The dosimeter on the left is shown before exposure to fast neutrons. For comparison, a dosimeter exposed to fast neutrons is shown on the right. Bubbles are clearly evident throughout the volume of the dosimeter. (Courtesy of Siemens Medical Systems, Inc., Concord, CA.)



FIGURE 84.6 Electronic personal dosimeter. (Courtesy of Siemens Medical Systems, Inc.)

84.7 Electronic Personal Dosimeters

All of the personal dosimeters discussed to this point are passive devices that require no external power source. The radiation dose is determined upon completion of the sampling period by subjecting the dosimeter to a separate processing step. The nature of the processing depends on the dosimeter, but examples include heating of a TLD, chemical etching of a foil, development of an emulsion, or simply counting bubbles. The electronic personal dosimeter (EPD), by contrast, is an active device that uses silicon diode detector technology to provide real-time measurements of radiation dose. The EPD has gained popularity in recent years and has mounted a serious challenge to TL dosimetry as the preferred official dosimeter of record. EPDs can provide extremely sensitive real-time measurements of both the total dose and the dose rate. In addition, the EPD can be designed to provide an audible alarm if the total dose or dose rate exceeds user-defined settings. While the quality and characteristics of EPDs from different manufacturers can vary significantly, some EPDs offer accurate dose measurements over a wide range of doses, dose rates, and energies and perform on a par with TLD dosimeters. A commercially available EPD is shown in Figure 84.6. This unit uses three silicon diodes to provide deep and shallow dose information for a wide range of gamma, beta, and x-radiation energies. EPDs typically have a display for manual readout of dose data but can be used with an automated reader for convenient archiving of data and for rezeroing the dosimeter for repeat use. Personal dosimeters are envisioned that would combine active and passive technologies, providing the accuracy and convenience of real-time dose monitoring in addition to the reliability of passive, cumulative dose measurements.

Defining Terms

Absorbed dose (*D*): The mean energy absorbed per unit mass.

Dose equivalent (*H*): The mean energy absorbed per unit mass, scaled with a quality factor, *Q*, that provides an estimate of the relative biological hazard of the radiation.

Effective dose (*E*): The total mean absorbed dose for an exposed individual. Obtained by summing the equivalent doses for all tissues, each scaled by a tissue weighting factor, w_T.

Equivalent dose (H_T): The mean absorbed dose in an organ, *T*, scaled by a radiation weighting factor, w_R.

Personal dose equivalent $H_p(d)$: The dose equivalent in soft tissue at a depth, d, below a specified point in the body. Commonly reported as shallow dose ($d = 3$ mm) and deep dose ($d = 10$ mm).

Radiation dosimetry: The quantitative measurement of the physical changes that occur in matter upon exposure to ionizing radiation.

References

1. ICRP Publication 60, Ann. ICRP 21(1–3), 1990 *Recommendations of the International Commission on Radiological Protection*, Oxford, U.K.: Pergamon Press; International Commission on Radiological Protection, 1991.
2. ICRU Report 51, *Quantities and Units in Radiation Protection Dosimetry*, Washington, DC: International Commission on Radiation Units and Measurements, 1993.
3. W. K. Sinclair, The present system of quantities and units for radiation protection, *Health Physics*, 70, 781–786, 1996.
4. S. W. S. McKeever, M. Moscovitch, and P. D. Townsend, *Thermoluminescence Dosimetry Materials: Properties and Uses*. Ashford, U.K.: Nuclear Technology Publ., 1995.
5. P. Braunlich, Present state and future of TLD laser heating, *Radiation Protection Dosimetry*, 34, 345–351, 1990.
6. S. D. Miller and P. A. Eschbach, Optimized readout system for cooled optically stimulated luminescence, *Radiation Effects and Defects in Solids*, 119–121, 15–20, 1991.
7. E. Piesch, B. Burgkhardt, and M. Vilgis, Progress in phosphate glass dosimetry: Experiences and routine monitoring with a modern dosimetry system, *Radiation Protection Dosimetry*, 47, 409–414, 1993.
8. S. W. S. McKeever, M. S. Akselrod, and B. G. Markey, Pulsed optically stimulated luminescence dosimetry using α-Al_2O_3:C, *Radiation Protection Dosimetry*, 65, 267–272, 1996.
9. B. L. Justus, S. Rychnovsky, M. A. Miller, K. J. Pawlovich, and A. L. Huston, Optically stimulated luminescence radiation dosimetry using doped silica glass, *Radiation Protection Dosimetry*, 74, 151–154, 1997.
10. R. A. Dudley, Dosimetry with photographic emulsions. In: Frank H. Attix and William C. Roesch (Eds.), *Radiation Dosimetry*, Vol. II, pp. 117–129. New York: Academic Press, 1966.
11. G. Espinosa, R. B. Gammage, and K. E. Meyer, Response of different polyallyl di-glycol carbonate materials to alpha particles, *Radiation Measurements*, 26, 173–177,1996.
12. K. E. Meyer, R. B. Gammage, C. S. Dudney, S. Reed-Walker, P. Kotrappa, R. V. Wheeler, and M. Salasky, Field measurements of plutonium and americium contamination in soils at the nevada test site using passive alpha detectors, *Radioactivity and Radiochemistry*, 5, 26–41, 1994.
13. J. L. Decossas, J. C. Vareille, J. P. Moliton, and J. L. Teyssier, Theoretical study and calculation of the response of a fast neutron dosemeter based on track detection, *Radiation Protection Dosimetry*, 5, 163–170, 1984.
14. S. Sadaka, L. Makovicka, J. C. Vareille, J. L. Decossas, and J. L. Teyssier, Study of a polyethylene and CR 39 fast neutron dosemeter II: Dosimetric efficiency of the device, *Radiation Protection Dosimetry*, 16, 281–287, 1986.
15. H. Ing, R. A. Noulty, and T. D. McLean, Bubble Detectors—A maturing technology, *Radiation Measurements*, 27, 1–11, 1997.

Further Information

Attix, F. H., *Introduction to Radiological Physics and Radiation Dosimetry*. New York: John Wiley & Sons, 1986.

Dosimetry, Radiation, EHS. http://www.ehs.uci.edu/programs/radiation/radfaq.html (accessed on July 27, 2012).

Dosimetry Systems for Radiation Processing, GEX Corporation. http://gexcorporation.com/ProductsType/Dosimetry-Systems.html (accessed on July 28, 2012).

Gad Shani, *Radiation Dosimetry: Instrumentation and Methods*, 2nd edn. Boca Raton, FL: CRC Press, 2000.
Kase, K. R., Bjarngard, B. E., and Attix, F. H. (Eds.), *The Dosimetry of Ionizing Radiation*, Vol. III. New York: Academic Press, 1990.
Kiff, S., *Advanced Ionization Chambers for Radiation Detection: Electrode Design and Performance Analysis of High Pressure Xenon Ionization Chambers*. Saarbrucken, Germany: VDM Verlag, 2008.
Knoll, G. F., *Radiation Detection and Measurement*. New York: John Wiley & Sons, 1989.
Stabin, M.G., *Fundamentals of Nuclear Medicine Dosimetry*. New York: Springer, 2008.

IX

Wireless Instrumentation

85 **Wireless Instrumentation** *J.P. Carmo and J.H. Correria* **85**-1
Introduction • Instruments and Instrumentation • Technology for Wireless Systems • Biomedical Applications • References • Partial List of Manufacturers

86 **Wireless Sensor Node Hardware** *Michael Healy, Thomas Newe, and Elfed Lewis* **86**-1
Introduction • Development History of Sensor Node Hardware • Sensor Node Components • Future Directions • Summary • References • Partial List of Vendors and Suppliers

87 **Mobile Instrumentation with Wireless Design and Implementation** *Frederick Fortson and Kenneth Johnson* **87**-1
Introduction • Wireless Hub Design Considerations • Implementing I/O for Vehicle and Mobile Assets • Embedded Vehicle Area Network Hub Architectural Overview • EVAN Power Supply Design Considerations • Physical and Electric Footprint Advantages • Bibliography

88 **Powering Autonomous Sensors** *Manel Gasulla, Maria Teresa Penella, and Oscar Lopez-Lapeña* **88**-1
Autonomous Sensors • Ambient Energy Sources and Transducers • Energy Storage Units • References

89 **Wireless Sensing Technology** *Gregory C. Willden, Ben A. Abbott, and Ronald T. Green* **89**-1
Introduction and General Definition • Wireless Sensor Networks • Choosing Reasonable Wireless Sensor Technology Solutions • Example System: Perch and Move • Example System: Localization for Landslide Measurements • Example System: Neutrally Buoyant Sensor Cave Mapping • Example System: NBS Variant for Borehole-Accessed Caves • Example System: NBS Variant for Culvert Inspection • Conclusions and General Design Considerations • References

90 **Telemetry** *Albert Lozano-Nieto* **90**-1
Introduction • Multiple-Channel Telemetry • Defining Terms • References • Relevant Websites about Telemetry



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

85

Wireless Instrumentation

85.1 Introduction **85**-1
85.2 Instruments and Instrumentation **85**-2
Measurement Systems • Multiplexing Structures • Wireless Instruments and Communication Protocols
85.3 Technology for Wireless Systems **85**-8
Operational Issues • RF Interfaces and Examples • Power Consumption • Networks of Wireless Instruments
85.4 Biomedical Applications **85**-19
COTS and Customized Applications • New Concepts for Wireless Instruments
References **85**-22
Partial List of Manufacturers **85**-25

J.P. Carmo
University of Minho

J.H. Correria
University of Minho

85.1 Introduction

Almost anyone knows that a measurement is the process of comparing a quantity with another one of the same type (e.g., length, volume, area) whose result is a number. Such a measurement can be done without ambiguities in a straightforward manner with the help of a measurement system, which requires specific instrumentation for achieving such task [1]. This instrumentation can be a simple instrument for directly measuring a physical quantity (e.g., a voltmeter for measuring an electrical potential difference between two points of a circuit, an ampere meter for measuring a current flowing in a branch of a circuit, a thermometer for measuring temperatures) and can take one of these following forms regarding their internal working and signal processing: analog or digital [2]. The ability to connect and communicate with external devices [3] (using dedicated cables and/or communication networks) as well as the inherent flexibility [4] (their easiness for adding new functions and/or reconfiguration of the existing ones) makes digital instruments have major potential for use in several fields of human activity (heavy industry, medicine, transportation systems, domestic, agriculture, and food industry are some application examples). The next evolutionary step of measurement instruments is integrating functions to provide wireless transfer of data. In this sequence of ideas, the developments of microelectronics and microsystems allowed engineers to successfully develop this new measuring method [5]. This resulted in new possibilities for measuring, acquiring, transferring, storing, and analyzing the physical world: embedded systems [6] and wireless sensors networks [7] are two new possibilities for achieving such a goal, with wireless being the *major attractive technology*. This leads to the wireless instrument concept, which by its nature requires multidisciplinary concepts such as measurement science, electronic circuits design, microelectronics and microsystems fabrication, wireless communication systems, and networking [8]. Figure 85.1 reinforces this idea by showing the different disciplines that must be employed for designing a wireless instrument. The primary focus of this chapter is the presentation and integration of these concepts. This chapter also presents biomedical applications based on wireless instruments and new application concepts.

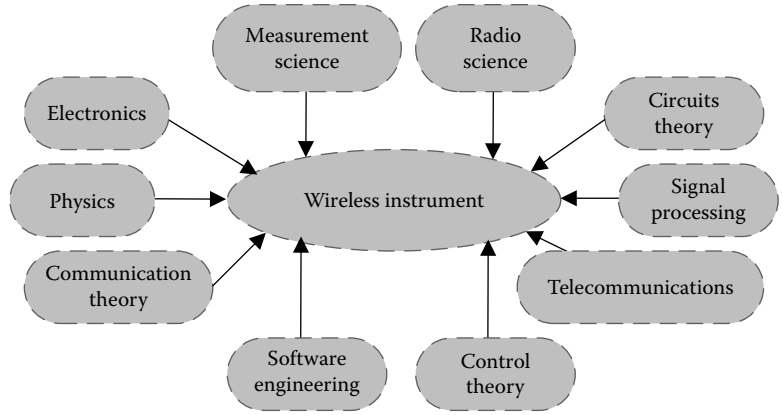


FIGURE 85.1 The multidisciplinary areas to take in account when designing a wireless instrument.

85.2 Instruments and Instrumentation

85.2.1 Measurement Systems

Figure 85.2 shows a block diagram of a generic measurement instrument. The blocks of a measurement system can be grouped into three major types: the real world (representing the physical quantity to be acquired), the interface block (with the sensor), and the core (e.g., the instrumentation itself).

There are situations where the interface block can be part of the core, e.g., a voltmeter doesn't require any external sensor because this one is already embedded inside the measuring instrument, thus the sensing tips can directly touch the electrical potentials. In this context, it must be clarified that a sensor cannot be confused with a transducer, because the transducer can perform the same function as the sensor, but if the sensor is passive (e.g., a physical quantity-dependent resistor mounted in a Wheatstone bridge), then additional circuits must be provided for obtaining the signal from the sensor. This means that the set composed of the sensor and powering system makes a transducer, confirming that certain sensors are simultaneously transducers.

The core blocks can include electronics of control for acquisition from the transducer. The core also provides signal-processing functions for signal conditioning purposes. These last functions include amplification (with the possibility to adjust the gain), filtering (either low-pass or band-pass or even high-pass filtering), and analog-to-digital conversion. Then, the user can read the acquired values in a dedicated display. A more sophisticated core system can interface with the external world either to connect several measurement instruments or to send data to a central unit for further processing. These communications can be done using wired buses (e.g., I^2E bus [9], GPIB [10], RS232 [11,12], parallel ports [13], or even USB [14]) or wirelessly (e.g., IEEE 802.11 [15], ZigBee [16], Bluetooth [17], or using a customized solution [18]).

The core blocks of the measurement instruments can be analog or digital. Analog is the less versatile core because it requires the presence of a person to read the measurements. This type of instrument is very limited

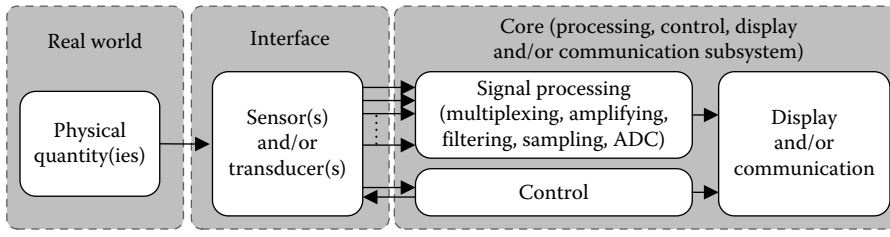


FIGURE 85.2 Block diagram of a generic measurement instrument.

and very difficult to adapt to a wide range of signals to measure. Furthermore, it is not possible to wirelessly send the physical quantities, unless a specific interface with an analog modulation scheme is provided.

A digital core can be used for connecting transducers (whose output can provide signals in the analog or in the digital domain). The difference from their analog counterparts resides in the conversion component used in the final processing stage, e.g., the sampler and the analog-to-digital converter (ADC) block. The inclusion of multiplexers enables the acquisition of multiple channels with a single measurement instrument. This topic will be the focus of discussion in the next section. After the ADC conversion, the acquired measurements can be presented in a numerical display. These cores can also be built with internal memory for storing the ADC-converted samples for rendering in a more complete display system (e.g., a flat screen) or for remote transmission through a communication interface. This core also allows for changing and/or for programming the amplifier's gain, thus allowing adaptation to wide variations of physical quantities. The latest developments of microelectronics make available transducers with digital outputs (e.g., integrated monolithic temperature transducers [19], Hall effect magnetometers [20], and accelerometers [21], among others) it is possible to have full digital and reusable cores.

The judicious selection of transducers and cores can be decisive points for fabricating wireless instruments with low power, reduced sizes, and low prices. This is especially evident on measurement instruments composed by reusable cores (for controlling and displaying/communicating), monolithic transducers (for signal acquisition), and on-chip signal conditioning circuits [22] (for signal processing).

85.2.2 Multiplexing Structures

A multiplexing operation is required in measurement instruments for simultaneously acquiring more than one signal at once. The signal multiplexing is not a simple matter to treat. The first issue concerns the availability of a variety of multiplexing structures and the decision to select the most suitable. This poses tradeoff problems related to implementation costs (e.g., a more complex multiplexing system costs more) and specifications (e.g., a more general purpose instrument will be more complex and cost more). In the second place, it is mandatory to have a clear knowledge of the input characteristics. The most important issues are the input impedance, the dynamic range, the bandwidth, the balancing type of signals (e.g., single-ended or differential), power-supply interference rejection (e.g., 50 Hz in the Europe and 60 Hz in the United States), interference between input channels due to the multiplexer, and noise.

The multiplexing configuration can assume one of the following classifications: either low level or high level for analog or digital multiplexers, respectively.

The structure illustrated in the Figure 85.3 is the most simple. The signals at the outputs of transducers connect to an analog multiplexer (with single-ended or differential inputs). It is important to have a variable gain amplifier when transducers with different signals are used, in order to provide signals within the full dynamic range of inputs of the ADC. This structure poses significant restrictions for solving speed bottlenecks: First, the multiplexer must be fast enough for switching the different analog channels; in the second place, the bandwidth of the ADC and sample-and-hold (S/H) (which is actually a part of the ADC) circuits must be high enough to avoid distortion of the analog signals for conversion.

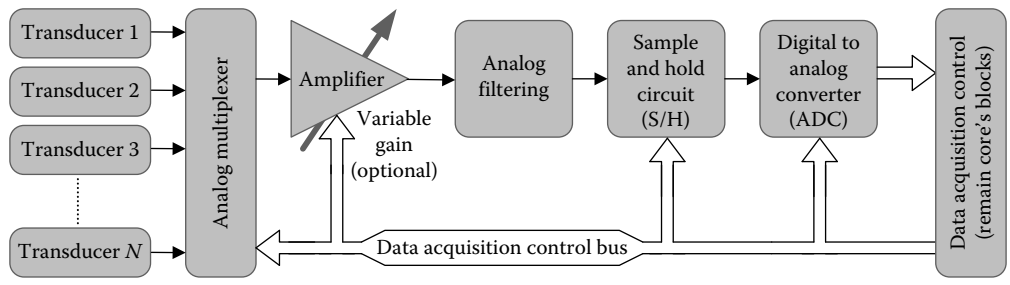


FIGURE 85.3 Low-level multiplexing configurations with shared amplifier.

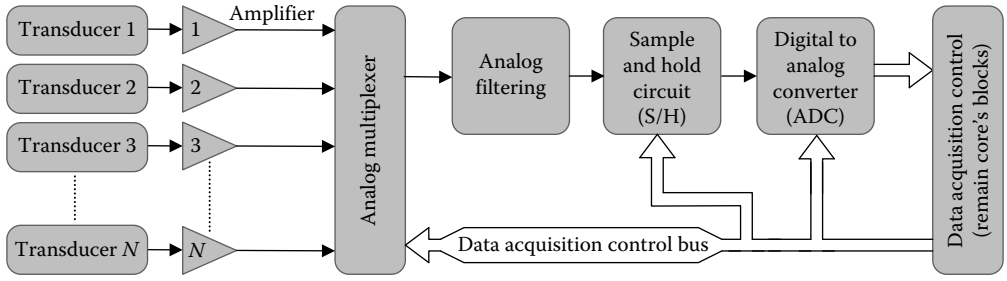


FIGURE 85.4 Low-level multiplexing configurations with dedicated amplifiers.

The low-level shared-amplifier configuration is less flexible in terms of plurality of physical quantities to measure; the ADC must be compatible with the transducer with the highest bandwidth; the direct acquisition of multiple signals can pose shielding problems because the transducers can be located very far away from each other. However, the low-level shared-amplifier configuration is the configuration with the highest potential for fabricating low-size electronic modules with high degree of integration (e.g., one multiplexer, one amplifier, one analog filter, one S/H plus ADC and control electronics).

The structure illustrated in the Figure 85.4 is very similar to that illustrated in Figure 85.3. The only exception is the use of transducers with associated amplifiers. These dedicated amplifiers guarantee signals with equal excursions to use the full dynamic range of the ADC. The bandwidth considerations are the same as those made in the previous multiplexing configuration. A high degree of integration is still possible to achieve with this configuration, but it sacrifices the compactness and low size due to the use of multiple amplifiers. Despite the acquisition of signals with equal amplitudes, the shielding problems remain in applications requiring transducers placed away from each other.

Low-level multiplexing configurations have a high integration potential, allowing the fabrication of wireless instruments with small sizes and low power consumption. However, the shared nature of the analog multiplexer makes them less than ideal. Nonidealities include the finite impedance of unselected analog channels, cross talk between channels, nonzero switching times between channels, and handling of different bandwidths between channels.

Figure 85.5 illustrates a high-level multiplexing configuration, providing a dedicated set of amplifiers, analog filters, S/H circuits, and ADCs for each transducer. After conversion to the digital domain,

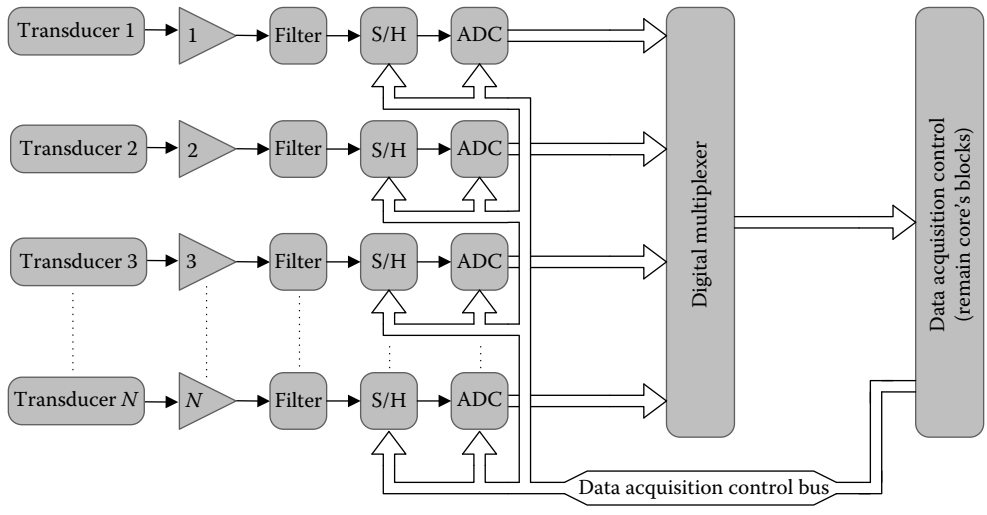


FIGURE 85.5 High-level multiplexing configuration.

the acquired physical measures are digitally multiplexed, avoiding the nonidealities of analog multiplexers. This configuration can pose integration restrictions by requiring circuits with increased sizes because each analog channel requires a dedicated signal-processing chain. However and despite these drawbacks, a high-level configuration is the most flexible of those analyzed because different amplifiers and filters can be provided for an extensive set of transducers. Moreover, this configuration offers the possibility to select the most suitable ADC for the respective analog chain. More important is the ability to accommodate channels with different sampling frequencies, since the switching speed of the analog multiplexer is not exceeded. This configuration also allows a variety of channel-selecting policies for desired channels with desired sampling frequencies.

85.2.3 Wireless Instruments and Communication Protocols

The design of a wireless instrument (as well as a generic measurement instrument) can't be done without taking in account the communication protocol. Table 85.1 shows the layer structure proposed by the OSI (Open Systems Interconnection) model [23].

Layer 1 specifies the modulation in conjunction with the line coding m (*codulation* [24]), the DC balancing, single-ended/differential-ended balancing.

Normally, layer 2 is divided in two sublayers, e.g., the medium access control (MAC) and the logic link control (LLC). The MAC sublayer is on bottom and defines mechanisms (or rules), in which a given transmitter is allowed (or not!) to access the physical medium for signal injection (corresponding to the data intended for the target receiver/s). The top LLC sublayer specifies the types of frames: data or control frames. This sublayer also provides a clear definition of frame formats in terms of their contents (their fields). Basically, the data frames are used to transport useful information in a field known as payload, while the other frames (e.g., the control frames) are used to make the transmitter know if the previous transmissions were received and processed by the receiver with or without errors. The control frames can also be used for performing flow control, in order to avoid data congestion in the receiver or across the network (with possible loss of data).

The network layer specifies a set of procedures for guaranteeing a reliable transmission between consecutive nodes along the network. Examples of procedures include the detection/correction of transmission errors and flow control. This layer also establishes the routing paths for the messages.

The transport layer ensures a reliable communication between terminals, e.g., between end-to-end users. This layer also provides error control procedures to verify the correct reception of all packets that form the messages (e.g., error-free packets). Another important procedure provided by the transport layer is grouping the packets in the correct sequence order to obtain a correct reassembled segment.

The session layer provides mechanisms for allowing the hosts to establish a communication. This layer also provides recovery mechanisms when an interruption occurs during the communication.

TABLE 85.1 Layers in the OSI Model

	Data Unity	Layer	Function
Host layers	Data	Application	Final communicating application
		Presentation	Data formatting and cryptography
		Session	Communication between hosts
	Segments	Transport	End-to-end communication reliability
Medium layers	Packets	Network	Node-to-node routing and communication reliability
	Frames	Data link	Physical identification (MAC and LLC sublayers)
	Bits	Physical	Signal transmission through the communication channel (*codulation*, other physical aspects of signals)

Source: Stallings, W., *Data and Computer Communications*, Prentice Hall, Englewood Cliffs, NJ, 2003.

The definition and conversion between data formats is done in the presentation layer. Normally, it is in this layer that the data are ciphered and/or converted to (and obviously converted from) the format used by the application.

The last layer (e.g., the application layer) provides interfaces between the application itself and between the protocol stack on the bottom.

Figure 85.6 provides a better understanding as how these concepts can be applied when designing a wireless instrument. This example helps to identify the blocks inside the wireless instrument that implement functions defined by the OSI model. This figure shows only the measurement part of the wireless instrument (e.g., the measurement instrument). The system is composed of the core measurement subset including the signal acquisition block, the signal-processing block, the control block, the memory block—a part of the control block—and the display control, which is actually optional in wireless instruments. Additionally and as previously stated, the interface block can be a part of the

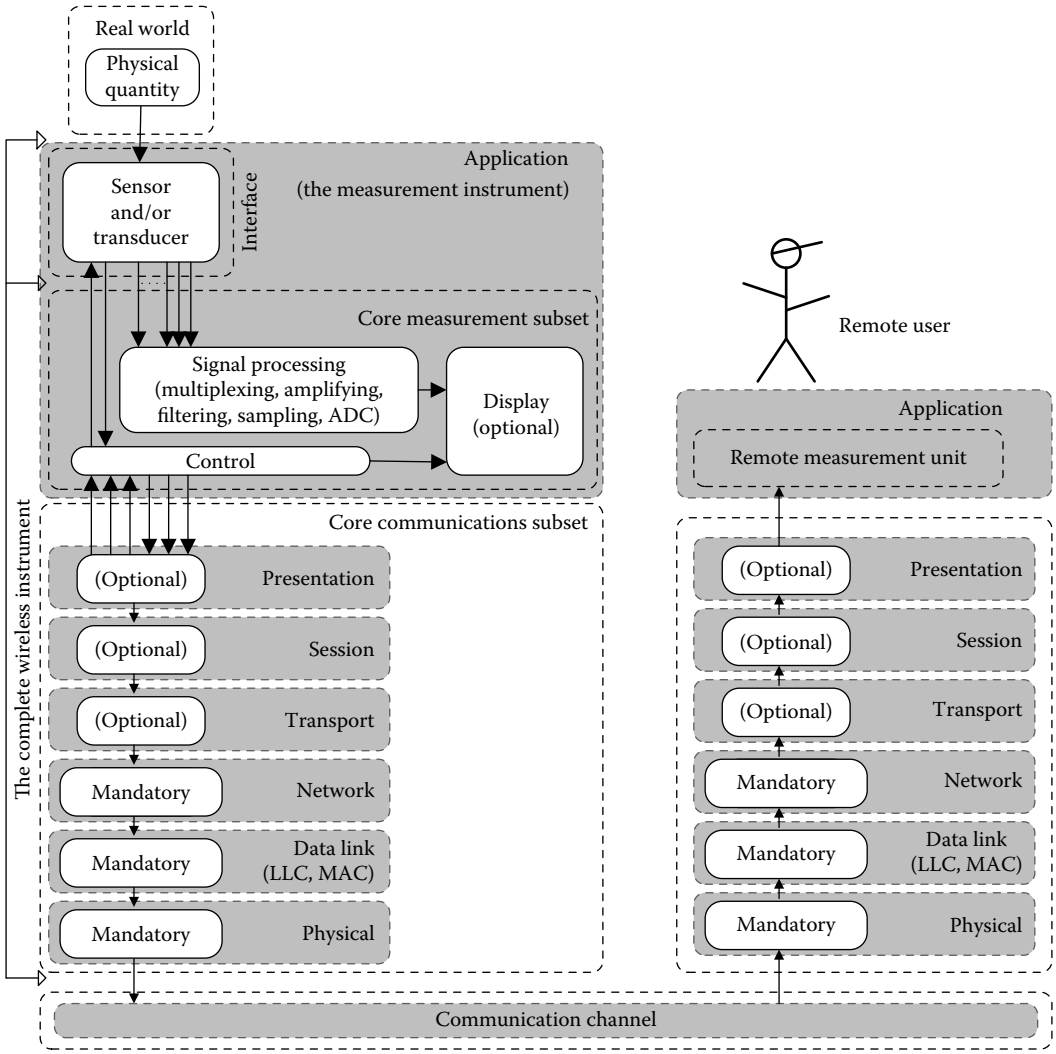


FIGURE 85.6 Relation between wireless instruments with the stacks of the OSI model.

measurement instrument. The remaining part of the wireless instrument is implemented by the core communication subset. This subset is responsible for communicating with external devices (other wireless instruments or with a central processing unit or even a remote measurement unit controlled by a remote user).

Normally and as shown in Figure 85.6, there are three optional functions inside a wireless instrument. This results from the need to simplify the transmission procedures in order to minimize the latencies (this is the case of the IEEE 802.15.4 that supports the low-level layers of the ZigBee protocol stack) and to maximize the data throughput.

The most common function of the network layer is the cyclic redundancy check (CRC) generation (by the wireless instrument on the transmitter side) and decoding (in the receiver side) for detecting transmission errors. It is possible to use forward error correction (FEC) schemes for error correcting but with the cost to send redundant bits for each bit of information (e.g., each bit of acquired data) and larger bit rates. In the case of auto-repeat request (ARQ) schemes, the CRC generation and decoding can be used by the receiver to do a request to the transmitter for retransmission of the erroneous packet.

The LLC sublayer functions can't be bypassed (or at least, a simple implementation must be provided) because at least a frame alignment word (or synchronization character) must be provided in the transmitter in order to allow the receiver to start the bits reception. The LLC implemented by Carmo et al. [18] uses a header composed of an alternate sequence (with a length equal to an integer multiple of eight) 1's and eight 0's for DC balance, while the remaining header contains a synchronization character with 8 bits (this complete header is used by both control and data frames).

The simpler way for implementing MAC sublayer functions is on point-to-point communications with only one transmitter and receiver. In this situation, the transmitter can send data without restrictions because the communication channel is always available and ready for use. Sophisticated MAC protocols are required for managing the medium access when the general scenario (with a multiplicity of transmitters and a multiplicity of receivers) is present. Silva et al. [25] proposes a MAC protocol for transmitting signals from transmitters placed on a multiplicity of paraplegic patients (doing hydrotherapy inside a swimming pool) into a base station, which stores the data for further analysis by a health professional. Note that the most common applications are based on simple point-to-point configurations as it is the case of the work done by Dias et al. [26], where a wireless instrument acquires EEG signals and transmits them to a base station connected to a personal computer for data logging. This solution uses wireless modules based on IEEE 802.15.4, but a solution based on a microcontroller with a simple radio frequency (RF) module could be used for reducing the latencies.

Taking into account the wired case (which can be directly transposed to the wireless case), the physical layer can be implemented by doing a direct connection between the transmitter and the receiver (e.g., connecting the transmitter's output to the receiver's input). A wireless instrument must provide at least one modulation scheme in order to reduce the errors introduced during the communication. A huge set of flexible wireless modules from third-party manufacturers ready for connection into the wireless instrument are available at low cost [27–30]. The modulation is an important issue in wireless instruments, but a coding scheme is recommended either for DC balancing or even for error control and for synchronization. In this context, the RF CMOS receiver at 433 MHz for integration in implantable devices found in [31] was designed taking in account the following *codulation* scheme: simultaneous on/off keying (OOK) modulation and biphase code [31]. Another example of physical layer definition is the RF transmitter proposed by Morais et al. [32] for operation at 433 MHz but with a different codulation scheme: OOK modulation and pulse-width modulation (PWM) code. This RF transmitter is compatible with the commercial receiver unit model LM RXAM2433 (from the manufacturer LPRS Inc.) and was tested with success for soil moisture measurements [33].

85.3 Technology for Wireless Systems

85.3.1 Operational Issues

The selection of the operation frequency is not an easy issue because different factors distinguish between them. In the first place, the dimensions of the antennas are imposed by the frequency. For an acceptable efficiency, the antenna size must be on the order of one-fourth of the wavelength, λ (m), which is given by $\lambda = c/f$ with $c = 3 \times 10^8$ m·s^{-1} being the speed of light in vacuum and f (Hz) the frequency of operation. Decreasing the dimension of an antenna implies the use of high frequencies. This issue can be solved by modulation. The most used modulations in wireless instruments (especially in laboratory environments, e.g., closed and relatively free of common band interference) are *amplitude-shift keying* (ASK), *phase-shift keying* (PSK), and *frequency-shift keying* (FSK). Both ASK and PSK need the same bandwidth given by BW = $2R_b$ (Hz), where R_b (bps) is the bit rate per second. The bandwidth required by FSK modulation is slightly higher and is BW = $2R_b + |f_1 - f_2|$, where $|f_1 - f_2|$ is the frequency shift between the two carriers, f_1 (Hz) and f_2 (Hz). The bit error probability (BEP) of each modulation can be found in Table 85.2, and the plots of their respective values as a function of the ratio E_b/N_0 are shown in Figure 85.7. The ratio E_b/N_0

TABLE 85.2 BEP for the Modulations ASK, PSK, and FSK

Modulation	BEP	
ASK with coherent detection	$Q\left(\sqrt{\frac{E_b}{N_0}}\right)$	$Q\left(\sqrt{\frac{S}{N}}\right)$
ASK with noncoherent detection	$\frac{1}{2}e^{-(1/2)\times(E_b/N_0)}, \left(\frac{E_b}{N_0}\right) > \frac{1}{4}$	$\frac{1}{2}e^{-(S/N)}, \left(\frac{S}{N}\right) > \frac{1}{8}$
PSK	$Q\left(\sqrt{2\times\frac{E_b}{N_0}}\right)$	$Q\left(\sqrt{2}\times\sqrt{\frac{S}{N}}\right)$
FSK	$Q\left(\sqrt{\frac{E_b}{N_0}}\right)$	$Q\left(\sqrt{1+\frac{\lvert f_1 - f_2\rvert}{4r_b}}\times\sqrt{\frac{S}{N}}\right)$

Source: Couch II, L.W., *Digital and Analog Communication Systems*, 5th edn., Prentice Hall, Englewood Cliffs, NJ, 1996.

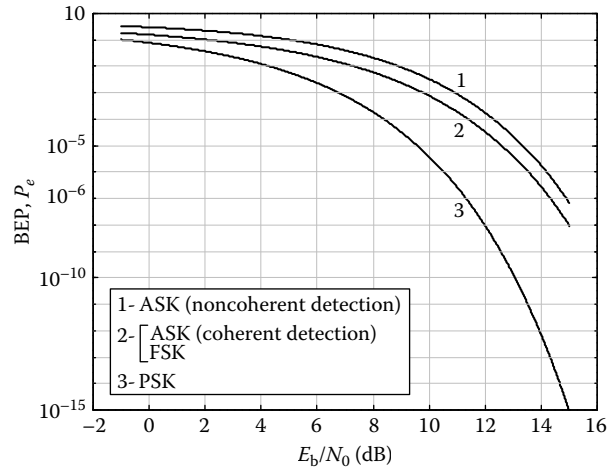


FIGURE 85.7 BEP versus the E_b/N_0 (dB) ratio for the modulations ASK, PSK, and FSK.

is the energy per bit, E_b (J), divided by the spectral density of the noise (additive white Gaussian noise), N_0 (W), which is given by [34]:

$$\frac{E_b}{N_0} = \frac{S/r_b}{N/\mathrm{BW}} = \left(\frac{S}{N}\right) \times \frac{\mathrm{BW}}{r_b} \tag{85.1}$$

where $N = N_0\mathrm{BW}$ (W) is the filtered noise at the output of a band-pass filter BW (Hz).

The selection of the frequency taking only in account the bit rate and antenna size isn't enough because as it is general knowledge, the antenna is perhaps the most critical subsystem in wireless communications. This imposes an antenna small enough for integration with the transmitter but not so small as to compromise this same miniaturization. The size reduction can be a problem because the antenna must be designed for transferring the highest possible power to the receiver. In this context, the size shortening of antennas can introduce additional problems of impedance matching [35] that must be solved. Wafer-level packaging (WLP) techniques can be used for overcoming these problems [36].

The communication range is another issue to take in account when selecting the frequency. This is related to the attenuation of RF signals, whose free-space value increases with transmitter/receiver distance, d (m), and frequency, f (Hz) [37]:

$$L_f(d,f) = -20\log_{10}(d) - 20\log_{10}(f) + K_f(\mathrm{dB}) \tag{85.2}$$

with $K_f = -20\log_{10}[c/(4\pi)]$ (dB). This means that for a simultaneously given transmitted power, P_t (dB), and receiver sensitivity, S_r (dB), the frequency is limited by the range, d_{max} (m):

$$f \le 10^{\frac{(P_t - S_r) - 20\log_{10}(4\pi d_{max})}{20}} (\mathrm{Hz}) \tag{85.3}$$

The free-space model is the most optimistic approach for calculating the link budget because the additional losses due to the surrounding environment (lossy propagation mediums, buildings, terrain conditions, vehicles, persons, shadowing, systemic implementations, among other factors) is not taken in account. Therefore, a general loss model must be used [37]:

$$l(d) = \alpha.c^{-n} + \chi, \quad n \ge 2 \text{ and } \alpha < 1 \tag{85.4}$$

Alternatively, $L(d) = -10n\log_{10}(d) + 10\log_{10}(\alpha) + \chi_{\mathrm{dB}} = A\log_{10}(d) + B + \chi_{\mathrm{dB}}$ (dB). The factor $A = -10n$ is very important to analyze because it justifies why the distance dependence loss is higher than that observed in free space. The signal fluctuations (also known as fading) don't contribute statically to the loss but dynamically. This dynamic behavior can impose severe restrictions when designing a wireless link because a super dimensioned RF receiver must be provided to overcome the temporary losses of signal power. The distance-dependent loss model is normally enough to predict the link budget especially for short distances (typically under 20 m) and closed spaces (laboratories, hospitals, residences, trains, among others). Very good references can be found in [38–40] to deal with the fading.

Figure 85.8 shows the available frequency bands for the different technologies used in wireless communications. Suitable frequencies for possible use in wireless instruments are those belonging to the so-called ISM band (industrial, scientific, and medical), due to its unregulated usage. These frequencies can be freely used without being subject to standardization but keeping the emission power below the maximum levels imposed by regulations. This usage flexibility leads to widespread new applications as will be discussed further.

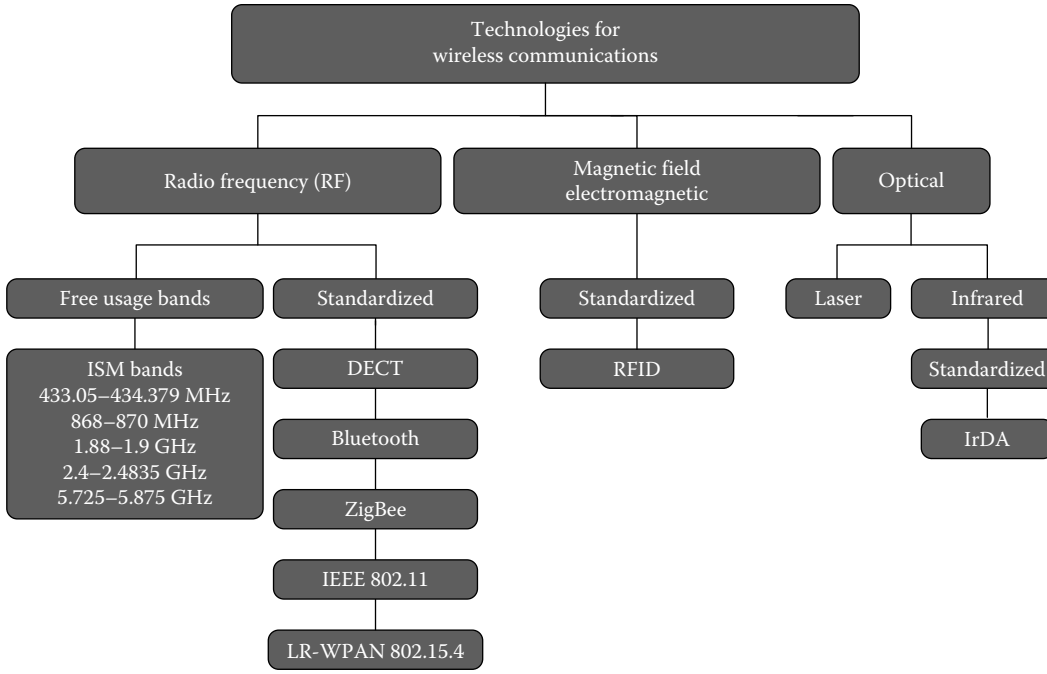


FIGURE 85.8 Currently available frequencies for wireless applications.

85.3.2 RF Interfaces and Examples

A wireless instrument communicates with the external world by RF. Thus, a wireless interface must be provided for allowing RF communications. Figure 85.9 shows a generic schematic block of a wireless microsystem performing functions of a stand-alone wireless instrument. These microsystems are composed of transducers and other electronic components for control and signal processing, by memory and by an RF interface (the RF transceiver) for connecting to an associated antenna. The dimensions of the RF transceiver must be comparable with other elements integrated in the microsystem (e.g., the transducers and remaining electronics). The miniaturization of electronics and the spreading of fabrication processes for integrating heterogeneous technologies (e.g., CMOS, SiGe, III/V technologies, MEMS, among others) will result in the mass production of wireless microsystems at low prices. All these issues combined with the flexibility to select which and the number of transducers for integrating together with the RF transceiver and remaining electronics allow engineers to design a wide number of devices for a wide number of applications. This last goal can be easily achieved with multi-chip module (MCM) techniques applied to a limited number of components (which can be of different technologies). In conclusion, the technology is also a major point to allow the fabrication of

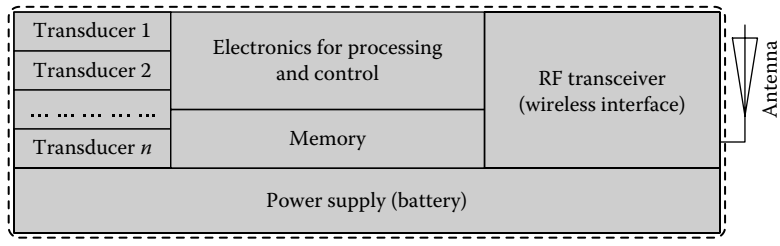


FIGURE 85.9 Generic microsystem architecture connected to an associated antenna.

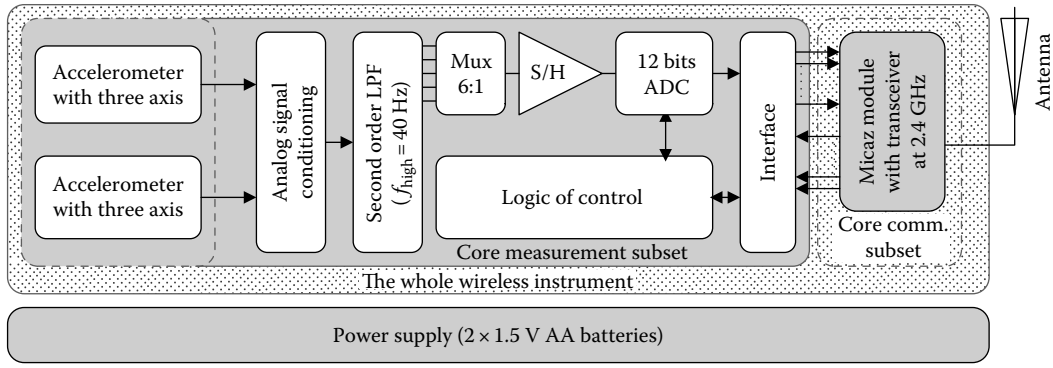


FIGURE 85.10 Block diagram of a wireless instrument at 2.4 GHz ISM band for monitoring the body movements of individuals.

wireless microsystems for use in wireless instruments. In this section, a few examples for each of the ISM bands in Figure 85.8 are presented for a better view of wireless instrument's potential.

85.3.2.1 Monitoring Body Movement

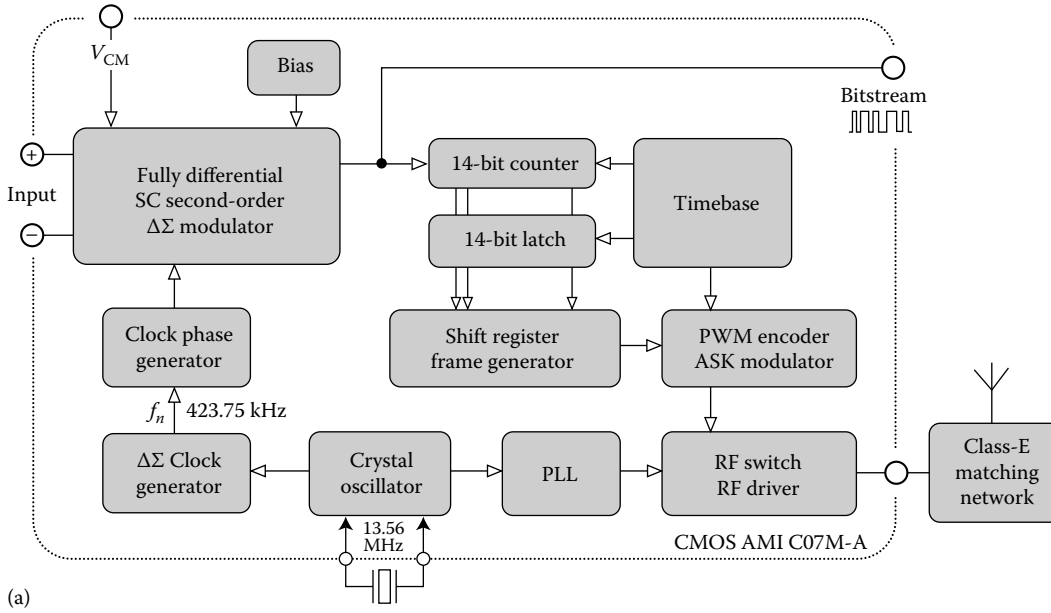
Figure 85.10 shows the block diagram of a wireless instrument for monitoring the body movements of individuals during hydrotherapy [41]. This wireless instrument is modular and is composed of a module with two MEMS accelerometers with three degrees of freedom (with three axes), a module with low-level multiplexing for signal processing and analog-to-digital conversion, and a third-party RF module (with a RF transceiver at 2.4 GHz and control electronics). The analog electronics in the core measurement subset is controlled by the core communications subset. The core communications subset is a Micaz RF module at 2.4 GHz (fabricated by the Crossbow company) for communicating with external devices and for controlling and managing the data acquisition process [42]. The accelerometers module was designed for measuring the movements of the individuals by obtaining information about the instantaneous roll, yaw, and pitch. This wireless instrument was designed for low-power and high-throughput communications using a specific MAC protocol for achieving such goals [43].

85.3.2.2 Connecting to Transducers

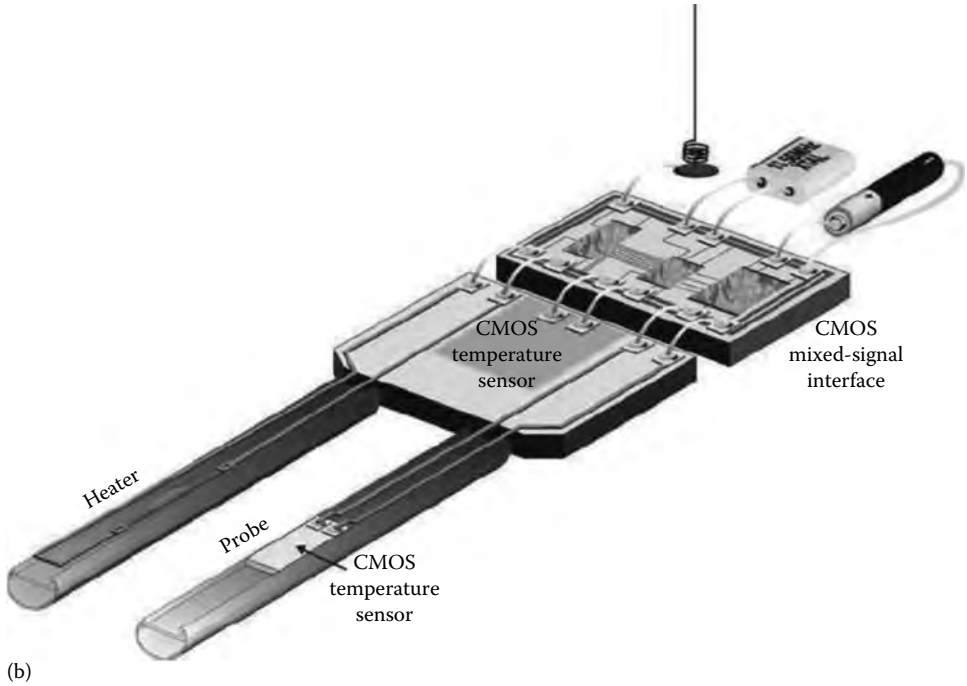
Figure 85.11 illustrates the block diagram of a sensor RF interface for operation at 433 MHz [32]. This RF interface has a differential (instrumentation) analog input for interference reduction purposes and allows the connection of other types of transducers. The analog-to-digital conversion is done by $\Sigma\Delta$ modulator for coding the input analog signals, whose result is a bitstream for encapsulation in a frame for RF transmission. As stated by Morais et al. [32], this interface was especially designed for connecting into the soil moisture transducer developed by Valente et al. [33]. The wireless instrument illustrated in Figure 85.11b was developed for measuring the soil moisture of greenhouses and uses the latter RF interface to send the acquired data to an external storing and/or analyzing unit.

85.3.2.3 Implantable Systems

Figure 85.12a shows the block diagram of a receiver for operation in the 433 MHz ISM band that was developed for use in implantable microsystems [31]. The selected architecture explores the superregeneration phenomena to achieve a high sensitivity. This receiver can be supplied with a voltage of only 3 V for demodulating signals with powers in the range (−100, −40) dB. The combination of modulation and coding scheme is OOK modulation combined with a variation of the Manchester code (e.g., a biphase code). The AMIS 0.7 μm CMOS process was selected for targeting the requirement to fabricate a low-cost receiver. Figure 85.12b shows a photograph of the first prototype (shaded area), which was integrated in a die with an area of 5×5 mm^2. An advantage of this receiver is being fully compatible with commercial



(a)



(b)

FIGURE 85.11 (a) The block diagram of a sensor interface for operation in the 433 MHz ISM band. (Reproduced from *J. Sens. Actuat. A*, 115, R. Morais, A. Valente, C. Couto, and J. H. Correia, A wireless RF CMOS mixed-signal interface for soil moisture measurements, 376–384, Copyright (2004), with permission from Elsevier.) and (b) a wireless instrument composed by the latter RF interface mounted in a soil moisture transducer for utilization in greenhouse environments. (Reproduced from *J. Sens. Actuat. A*, 115, A. Valente, R. Morais, C. Couto, and J. H. Correia, Modeling, simulation and testing of a silicon soil moisture sensor based on the dual-probe heat-pulse method, 434–439, Copyright (2004), with permission from Elsevier.)

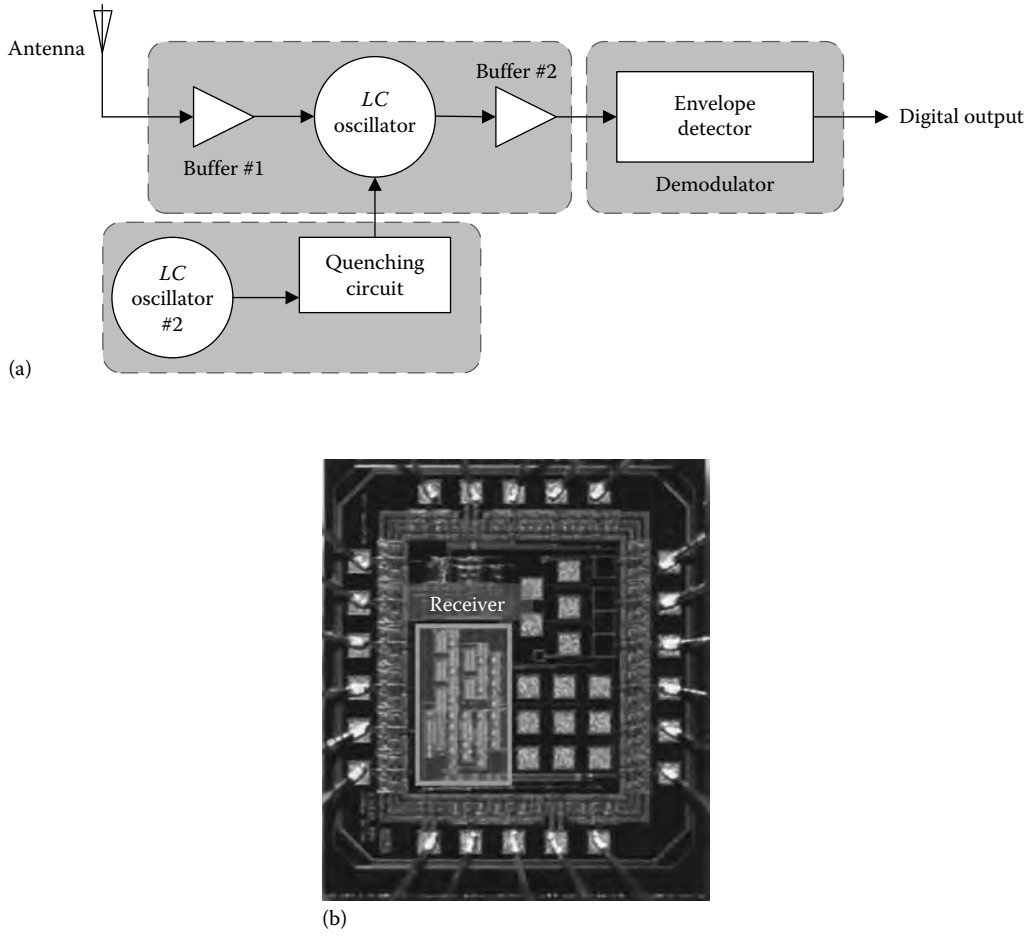
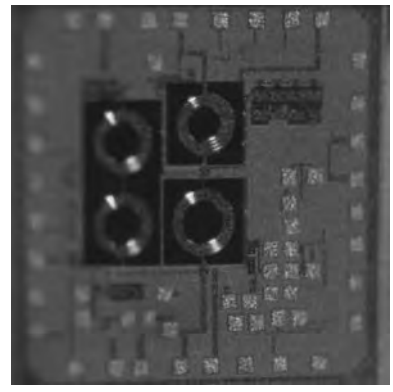


FIGURE 85.12 (a) The block diagram of the superregenerative receiver at 433 MHz and (b) a die photograph containing the first prototype of the superregenerative receiver (the rectangular area under the word Receiver). (Reproduced from *Microelectron. J.*, 42, J. P. Carmo, J. C. Ribeiro, P. M. Mendes, and J. H. Correia, Super regenerative receiver at 433 MHz, 681–687, Copyright (2011), with permission from Elsevier.)

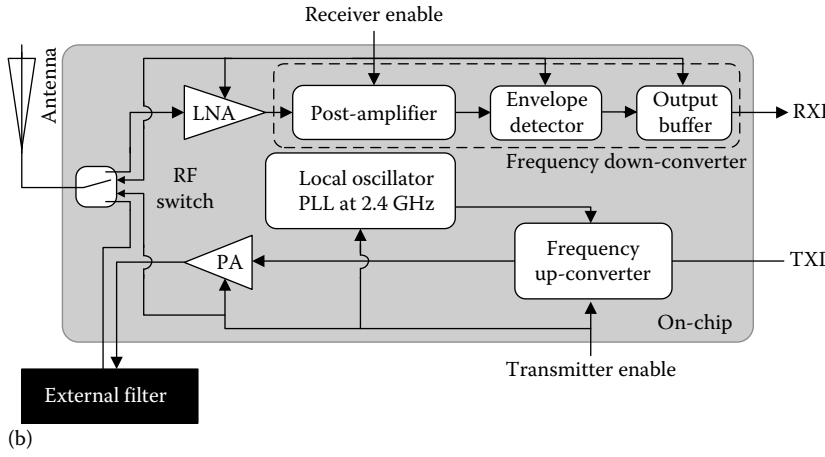
transmitters and the transmitter fabricated by Morais et al. [32], for the same coding scheme (the variation of the Manchester code) in the transmitter.

85.3.3 Power Consumption

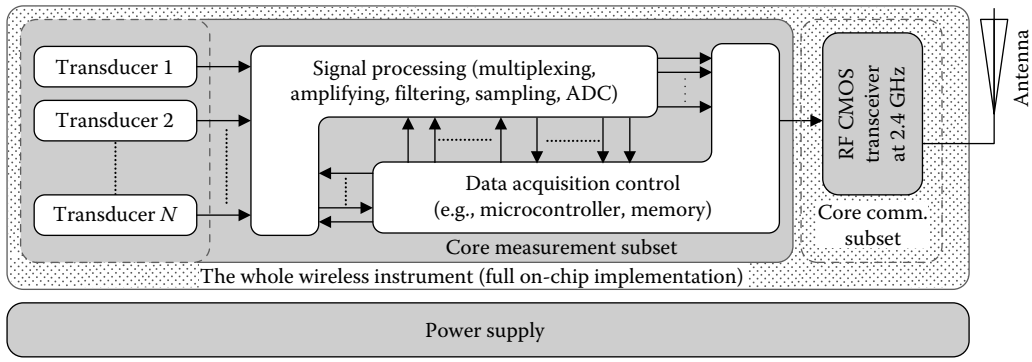
The power consumption of a wireless instrument limits its working time, especially when functioning with batteries. In this context, the selection (or even further, the design) of RF transceivers can't neglect this issue because this is the block with major impact for the total power consumption, when compared with the whole electronics in the instrument [44]. Furthermore and despite the spreading of microelectronics fabrication processes with the potential to achieve smaller power consumption, the RF transceiver is irremediably the subsystem of higher power consumption [45,46]. This demands the integrated definition of architectures and methods of control, as well as provides means to predict the power consumption of the RF system. Figure 85.13a shows an RF CMOS transceiver at 2.4 GHz that allows the implementation of control actions for optimizing the power consumption [23,47]. This RF CMOS transceiver was fabricated in a standard 0.18 μm CMOS process for achieving low power consumption with



(a)



(b)



(c)

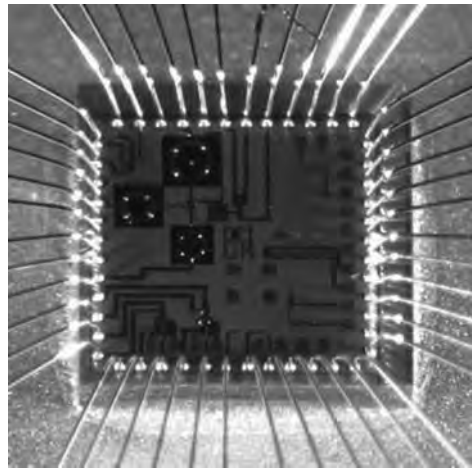
FIGURE 85.13 (a) Photograph and (b) block schematic of a RF CMOS transceiver at 2.4 GHz especially designed for stand-alone wireless instruments in biomedical applications. (c) A schematic illustrating the integration concept of the RF CMOS transceiver, transducers, and electronics in the same microsystem.

a low-voltage supply. As illustrated in Figure 85.13b, the design of this RF CMOS transceiver predicted the use of a control signal to either select the transmitter or the receiver in order to allow its integration with electronics to perform custom control, Figure 85.13c.

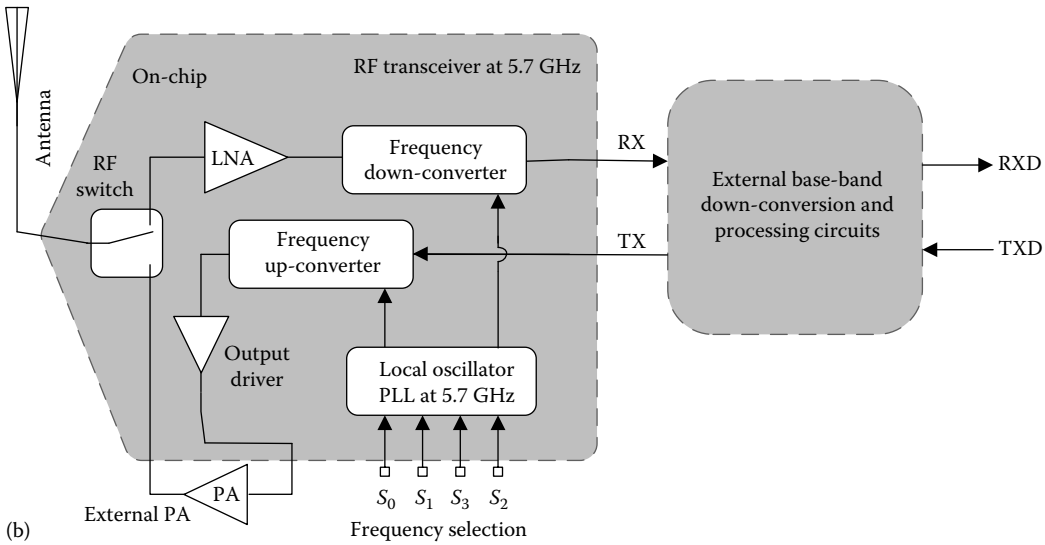
It is possible to explore the band located between the 5.7 and 5.89 GHz for implementing wireless instruments [48]. This band permits the fabrication of antennas, whose small dimensions allows their

integration with the electronics using WLP techniques [36]. The integration of antennas and electronics in the same microsystem results in smaller impedance mismatching problems. Moreover, the antenna and electronics co-integration systematizes the fabrication process and at the same time results in microsystems with a small cost per unit. The work presented in [49] takes all of this into account to provide a low-power/low-voltage wireless interface at 5.7 GHz with dry electrodes for implementing wireless instruments as parts of cognitive networks. Figure 85.14a shows a photograph of the wireless interface measuring of 1.5 × 1.5 mm. The schematic illustrated in the Figure 85.14b shows the block diagram of the RF part at 5.7 GHz. The digital signals $\{S_0, S_1, S_2, S_3\}$ select the target frequency in the range $f_{out} = f_{ref} \times (400 + 2S) = f_{ref} \times [400 + 2(S_0 + 2S_1 + 4S_2 + 8S_3)]$, whose range is located between 5.42 and 5.83 GHz for a reference frequency f_{ref} = 13.56 MHz.

An integrated low-cost solution for wireless instruments based on a microdevice fabricated with low power consumption using a 0.18 μm CMOS process is in [50]. This microdevice is composed of an RF transceiver, a reduced instruction set computer (RISC) microcontroller, a random access memory (RAM),



(a)



(b)

FIGURE 85.14 For the frequency of 5.7 GHz: (a) photograph and (b) schematic block of the RF part of a wireless interface at 5.7 GHz.

a power-supply management circuit, analog electronics of signal conditioning and analog-to-digital conversion (ADC), and circuits for providing communication based in SPI and I²C buses. The control electronics was developed for implementing a specific communication protocol for use with multiple wireless instruments and low power consumption, e.g., the WiseMAC protocol. According to the authors, this protocol working together with their RF transceiver achieves power consumptions thirty times smaller than those obtained with the IEEE 802.15.4 (which defines OSI layers 1 and 2 functions). Furthermore, the operation frequency can be selected from 433 to 868 MHz, as well as either with the OOK modulation or the FSK modulation. According to [50], their RF transceiver has a receiver power of 2.5 mW and a transmit power of 39 mW.

85.3.4 Networks of Wireless Instruments

A wireless sensors network can be considered a network of wireless instruments whose sensor nodes are the wireless instruments itself. The stand-alone operation without the need of a human operator for doing maintenance and/or for replacing the batteries (that provide the supply of power) are the main differences of these kind of wireless network compared to conventional networks. In this sequence and as illustrated in Figure 85.15, a wireless sensors network can be considered a distributed sensor network constituted by a high density of nodes. It is expected for each node to run simple protocols and provide low data rates in order to keep power consumption below reasonable levels. These issues are of major interest especially for nodes powered by batteries and without energy-harvesting capabilities because their useful life will be limited by the remaining charge.

In a wireless sensors network, each node acquires the physical data from the sensors and transmits by RF these same data toward a base station for storing and further analysis. The adjacent sensorial nodes can be used for storing the information and relaying into the base station, when the establishment of an RF link between a given sensor node and the base station is not possible. The information of a hidden node (with relation to the base station) is not lost because it can be relayed across the distributed infrastructure that forms the network of wireless sensors. However, their distributed nature and the nonexistence of a central controller imply the development of complex algorithms to deal with the multiplicity of nodes. This is especially true, when the topology of the network changes due to the malfunction of a node and/or when a new node is placed in the network or even when the existent nodes occupy new physical positions [51]. Moreover and contrary to what happens on wired networks, this type of network brings new problems: propagation aspects of RF signals and power-supply issues. The shared nature of the propagation medium is another problem because this makes the signals vulnerable to interference and multiple paths (fading), especially for mobile nodes and when a wide number of surrounding obstacles are present [52]. The shared medium also introduces problems related to the security and confidentiality of the data.

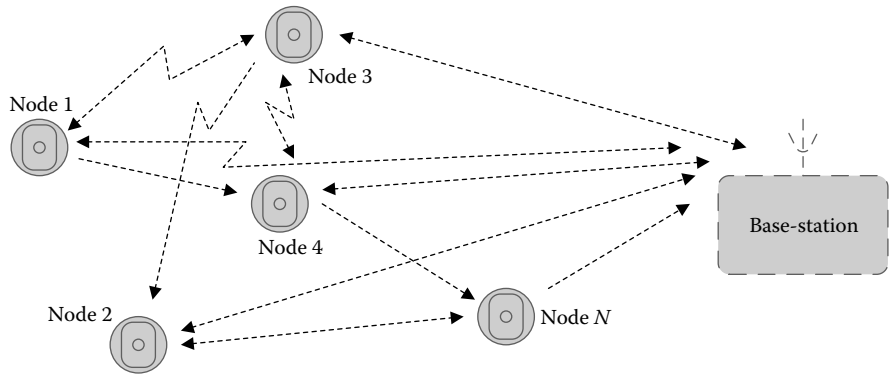


FIGURE 85.15 A schematic illustrating the wireless sensors network concept.

The increased application potential of wireless sensors networks in several fields of the human society (such as industrial, biomedical, transportation, domestic, and energy fields, among others) resulted in the need of their standardization as well as in their wide acceptance. Historically, the first wireless networks were mere technologic extensions of IEEE 802 local networks. Basically, the target of the local wireless networks was the interconnection of computers (as it was a common wired network). With time, other wireless networks appeared such as Bluetooth for connecting computers to their peripherals and IEEE 802.15.4 for wireless sensors networks. The IEEE 802.11 and the Bluetooth protocols are very difficult, complex, and, thus, with the potential to require devices with high power consumption. These are the reasons that made these protocols not suitable for wireless sensors networks but only for point-to-point connections. In this context, the need of protocols with low power consumption and simple procedures resulted in the establishment of the IEEE 802.15 workgroup. The joint actions developed by this workgroup resulted in the proposal of three different classes of wireless operation. The focus of the first class of operation targeted bit rates, whereas the second targeted power consumption, while the third class was more concerned with the quality of service (QoS). The need of protocols for low-power devices resulted in the proposition of the IEEE 802.15.4, as a basic set of rules for application in wireless sensors networks. The IEEE 802.15.4 protocol was developed for low-complexity applications and distances up to 10 m, allowing bit rates up to 250 kbps. Furthermore, the IEEE 802.15.4 protocol was proposed for a wide range of uses, ranging from consumer electronics, industrial and domestic automation, personal healthcare, and interconnection of computer peripherals. The IEEE 802.15.4 protocol defines two lower OSI layer functions. Figure 85.16 shows two versions of the physical layer of the IEEE 802.15.4 protocol [53].

The first version of the physical layer uses 868 MHz in Europe or 915 MHz in the United States. The European version permits only one RF channel for transmitting the maximum bit rate of 20 kbps, whereas the U.S. version allows the use of 10 simultaneous channels spaced by 2 MHz and a maximum bit rate of 40 kbps per channel. The second version of physical layers uses the 2.4 GHz band and supports the use of 16 simultaneous channels spaced by 5 MHz and a maximum bit rate of 250 kbps per channel. Table 85.3 shows that the IEEE 802.15.4 protocol uses spread spectrum techniques for increasing the resilience against a variety of factors that include interference from other radio stations and fading resulting from a multiplicity of radiowave paths. The spread spectrum techniques also ease the clock synchronization task in the receiver. These modulations belong to the constant amplitude modulations group and are very complex to implement because analog products in four quadrants are required. Fortunately, the RF *transceiver* CC2420 from the Chipcon company [54] is commercially available.

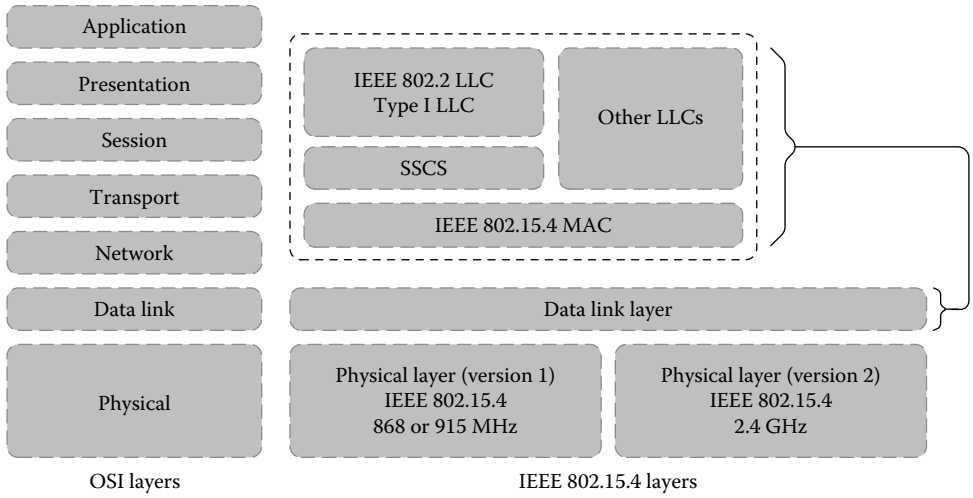


FIGURE 85.16 The OSI layer functions can be found in the IEEE 802.15.4 protocol.

TABLE 85.3 The Main Characteristics of the Physical Layer of the IEEE 802.15.4 Protocol

Band	Frequencies	Bit Rate (kbps)	Symbol Rate (ksymbols/s)	Spreading Parameters	
				Modulation	Chip Rate
868 MHz	868–868.6 MHz	20	20	BPSK	300 kchips/s
915 MHz	902–928 MHz	40	40	BPSK	600 kchips/s
2.4 GHz	2.4–2.4835 GHz	250	62.5	O-QPSK	2 Mchips/s

This RF transceiver contains a core ready for implementing all the IEEE 802.15.4 functions (naturally, the second version of the physical layer) and consumes only 19.7 mW when operating in the receiving mode as well as 17.4 mW when operating in the transmitting mode.

The IEEE 802.15.4 also defines data link layer standardized protocols (e.g., MAC and LLC). In this context, the LLC sublayer of the IEEE 802.15.4 uses the same type I LLC frame formats and procedures specified by the standard IEEE 802.2. The main difference between those used by local area networks and those used by wireless sensors networks is the adopted MAC sublayer. The MAC sublayer adopted by the IEEE 802.15.4 (e.g., the IEEE 802.15.4 MAC) is closer in hardware than the ordinary MACs adopted for local area networks. The service-specific convergence sublayer (SSCS) allows the adoption of other proprietary LLCs as an alternative to this one defined by the IEEE 802.2 (e.g., the type I LLC). The purpose of this model is to allow the IEEE 802.15.4 MAC to implement medium access mechanisms not defined in the IEEE 802.2 [53]. The structure of the MAC sublayer frames is flexible enough to allow the deployment of networks with a wide range of topologies and applications. Typically, an IEEE 802.15.4 MAC frame contains the following fields: a control field to indicate its type, a sequence field to indicate the number of frames for transmission, two fields with receiver and sender addresses information, a field with the information itself (designated as payload), and a field for data integrity check (e.g., the CRC for transmission errors verification).

Figure 85.17 shows that the ZigBee protocol is an extension of the IEEE 802.15.4 protocol. ZigBee uses the IEEE 802.15.4 protocol to implement the physical and data link layer functions. Furthermore, ZigBee supports a wider range of high-level functionalities (not present in the IEEE 802.15.4, which is closer to the hardware) as used in cryptography, management policies in environments with multiple users, as well as error control [53]. CC2430 is another integrated circuit fabricated by the Chipcon company [55] that eases the task of implementing ZigBee networks. CC2430 contains internally an RF transceiver and an additional core for implementing the ZigBee high-level functions. The RF part of CC2430 consumes 21 mW in the receiving mode and 30 mW in the transmitting mode. The CC2430 is at this date the first

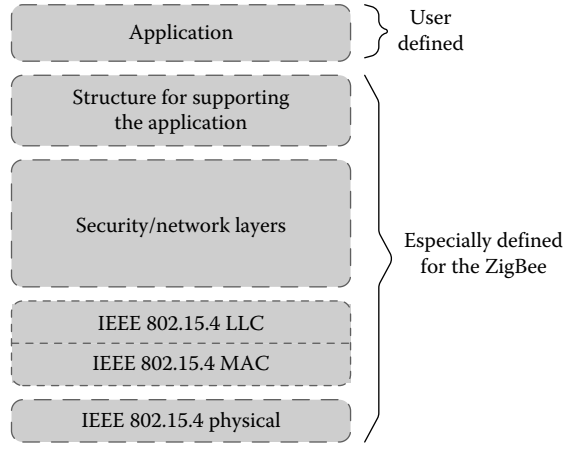


FIGURE 85.17 Adopted layers from the OSI model for use by ZigBee.

integrated circuit to offer a full ZigBee solution for the market. This integrated circuit has flash memory up to 128, 8 kbytes of RAM, an SPI interface, and a few pins for connecting analog and digital signals.

The ZigBee protocol was defined in response to the increased demand for wireless sensors by industry and the need for new personal applications. Thus, ZigBee allows fast prototyping of wireless sensors networks. A wide number of solutions are possible for wireless sensors networks. A few companies (that include Crossbow [42], Dust Networks [56], and Sensicast Systems [57]) offer products such as radios (motes) and sensor interfaces. The motes are battery-powered devices that run specific software. In addition to running the software networking stack, each mote is easily customized and programmed, since it runs open-source operating systems that provide low-level event and task management. Mote processor/radio module families working at 2.4 GHz ISM band that support IEEE 802.15.4 and ZigBee are available from the Crossbow company.

However, despite the inherent ease of the solutions based on motes, they can be very expensive when full custom network prototypes are required. The wireless sensors network solution found in [18] uses microcontrollers PIC from the Microchip company to meet a wide range of small volume applications with a low cost and in a ready to use fashion. Their solution uses a PIC microcontroller to provide the basic services of communication and control. Thanks to the serial connection of ADC chains, this solution is scalable in the sense that is possible to expand the number of attached transducers. The main drawback is that the maximum sampling frequency is limited by the number of transducers: the maximum sampling frequency per transducer is limited to f_S/N (Hz), where f_S (Hz) is the maximum sampling frequency when only one transducer is present and N is the number of transducers. To finish, the reprogramming of the microcontroller increases the functionalities to yield new services of nodes.

85.4 Biomedical Applications

85.4.1 COTS and Customized Applications

Biomedical applications have a high potential for using wireless instruments. An example that confirms this statement is the wireless monitoring systems of human body information as a growing field. A body area network comprises smart sensors able to communicate wirelessly to a base station.

An example of these applications is the wireless electroencephalogram (EEG), which is expected to provide a breakthrough in monitoring, diagnostics, and treatment of patients with neural diseases. Wireless EEG modules composed of the neural electrodes, processing electronics, and a RF transceiver with an associated antenna will be an important breakthrough in EEG diagnostics. Two approaches can be used for implementing wireless EEG systems: commercial off-the-shelf (COTS) and customized solutions. A COTS solution uses discrete integrated circuits and passive components for making wireless instruments, whereas a customized solution is designed from the scratch and further integrated on a single microdevice in order to optimize the size and power consumption and allow a power supply with small batteries (e.g., class AA, coin-sized batteries). The system proposed by Dias et al. [58] is an example of a COTS system for acquiring EEG signals and transmission by RF. This wireless EEG system uses a Micaz module [42] at 2.4 GHz for RF transmission and for controlling and converting the physical data. This system uses two 1.5 V class AA batteries for power supply and achieves maximum bit rates of 120 kbps. Other features of this system include 4 μV resolution, 15 mW power consumption, five single-ended channels, and size of 5.7 × 4.8 × 2.0 cm. Figure 85.18 shows the acquisition part of this wireless EEG system. A reference voltage must be added to the acquired signals before the analog-to-digital conversion because negative potentials are not provided by the power supply (the batteries can provide the following electrical potentials: ground, V_{dd}, and $V_{dd}/2$). Further explanations of this analog circuit (especially for the necessity of neutral and signal ground, SGND, electrodes) can be obtained in [59].

Customized solutions require the development of dedicated microelectronic systems or at least dedicated application-specific integrated circuits (ASICs). The wireless EEG system proposed by Yazicioglu

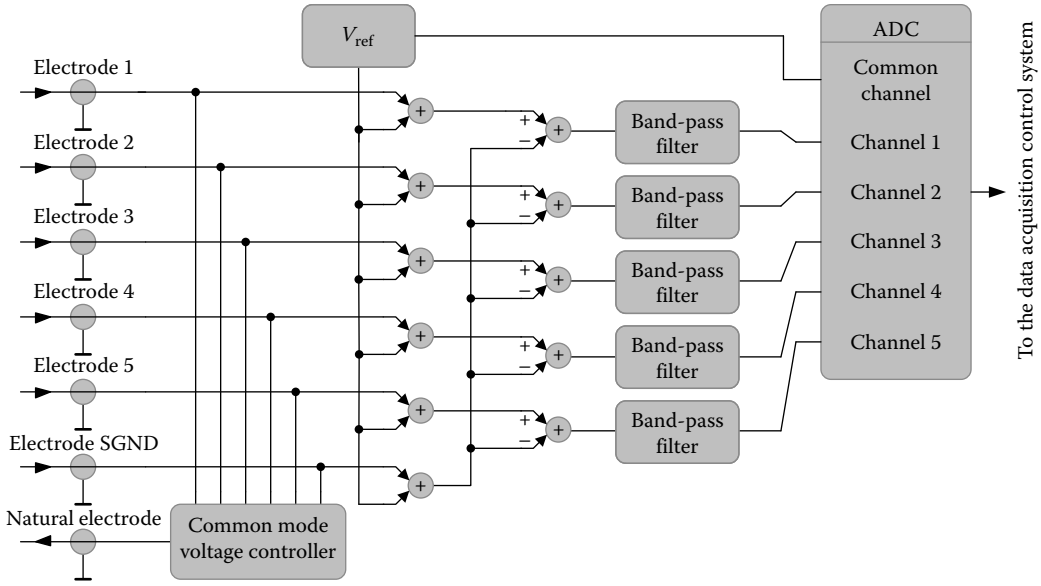


FIGURE 85.18 The analog signal-processing part found in. (From Dias, N.S. et al., A wireless system for biopotential acquisition: An approach for non-invasive brain-computer interface, *Proceedings of IEEE International Symposium on Industrial Electronics—ISIE 2007*, Vigo, Spain, pp. 2709–2712, June 4–7, 2007.)

et al. [60] pushes further the concept of wireless EEG, by using the heat of human body for powering the whole wireless instrument itself. Their wireless EEG system uses a thermoelectric generator to convert the temperature differences between the environment (the coldest side, at temperature T_c [K]) and the forehead of the subject (the hottest side, at temperature T_h [K]) [60]. The output voltage depends on the Seebeck effect [61] temperature difference $\Delta T = T_h - T_c$ (K). This wireless EEG system can acquire signals from eight EEG channels, whose inputs are differential (instrumentation) for noise and interference reduction. Each channel uses a new concept developed by them and known as AC-coupled chopper-stabilized instrumentation amplification (ACCIA) for achieving high common-mode rejection ratio (CMRR) and at the same time eliminating the flicker noise of the transistors as well as for filtering the differential DC voltage generated between two EEG electrodes [60]. A dedicated ASIC was developed for achieving a complete readout front end for the eight EEG channels, and thanks to it, the complete wireless EEG module mounted with the RF front end and with a backup lithium battery occupies a volume of 1 cm^3.

Another example of customized solution is the Sensium TZ1030 sensor interface illustrated in Figure 85.19a [62]. The Sensium TZ1030 sensor interface was developed by Toumaz Technology Limited Company for operation in the following bands: 863–870 MHz in Europe and 902–928 MHz in the United States. Internally, the TZ1030 is composed of analog and digital electronics for interfacing and calibrating the sensors. The sensors are external to the TZ1030 and can attach directly to it. An RF transceiver, an 8051 compatible microcontroller, RAM, and flash memories are also provided and make the TZ1030 a compact solution for an easy placement on the subject's body. The control software contains procedures for local processing of the information, in order to reduce the amount of information and the total transmission times. These features make the TZ1030 a low power consumption solution. Together with an appropriate external sensor, the TZ1030 is ready for acquiring electrocardiograms (ECGs), temperature, glucose levels, and oxygen levels in the blood. Figure 85.19b illustrates a possible architecture for rapid development of wireless body area networks (BANs), offering the possibility.

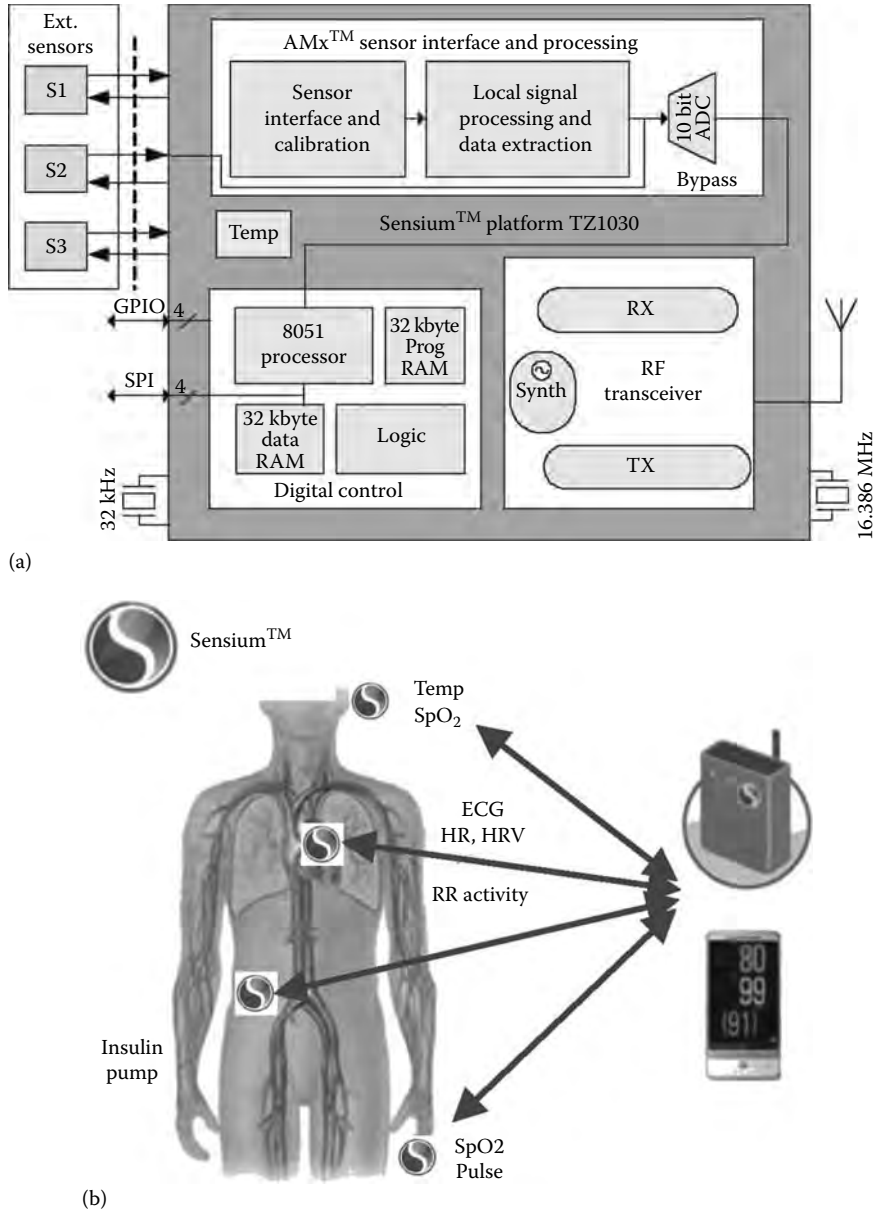


FIGURE 85.19 (a) Sensium TZ1030 sensor interface for transmission at 863–870 MHz, and (b) system architecture for biomedical applications using the TZ1030. (From Sensium TZ1030, Ultra low power smart sensor interface and transceiver platform, Toumaz Technology Limited. Online http://www.toumaz.com, accessed on February 27, 2012.)

85.4.2 New Concepts for Wireless Instruments

New techniques for implementing wireless instruments can be found in the literature. These techniques are extremely innovative due to the breakthrough introduced in the way the measurements are done. The work proposed by Karacolak et al. [63] takes in account the variation of the electric parameters for continuously measuring glucose (the electric parameters vary with the sugar concentration).

Alternatively, the research group of Chow et al. [64] explores an uncommon (but still very innovative) methodology that makes use of cardiovascular stents to receive RF signals inside the human body.

In this work, the stents are used as radiating structures for transmitting the measurements through the tissues of the human body.

Finally, the work proposed by Rodrigues et al. [65] uses a MEMS antenna with a U-shaped cantilever structure. This cantilever is sensitive to the magnetic field component of electromagnetic waves and will oscillate. A piezoelectric material layer of polyvinylidene fluoride (PVDF) is used to convert the magnetic field into a voltage useful enough to be acquired by the reading circuit. The major innovation of this technique allows the integration of antennas with implantable devices by way of WLP techniques for achieving the fabrication of small-sized devices. Their antenna occupies an area of only 1.5×1.5 mm [65].

References

1. D. Buchla and W. McLachlan (eds.), *Applied Electronic Instrumentation and Measurement*, 1st edn., Prentice Hall, Englewood Cliffs, NJ, 1991.
2. W. D. Cooper and A. D. Helfrick, *Electronic Instrumentation and Measurement Techniques*, 3rd edn., Prentice Hall, Englewood Cliffs, NJ, 1985.
3. J. H. Correia, G. de Graaf, M. Bartek, and R. F. Wolffenbuttel, A single-chip CMOS optical microspectrometer with light-to-frequency converter and bus interface, *IEEE Journal Solid-State Circuits*, 37(10), 1344–1347, October 2002.
4. G. R. Tsai and M. C. Lin, FPGA-based reconfigurable measurement instruments with functionality defined by user, *EURASIP Journal on Applied Signal Processing*, Article ID 84340, 1–14, January 2006.
5. I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, Wireless sensor networks: A survey, *Computer Networks*, 38(4), 393–422, March 2002.
6. V. Raghunathan, C. L. Pereira, M. B. Srivastava, and R. K. Gupta, Energy-aware wireless systems with adaptive power—Fidelity tradeoffs, *IEEE Transactions on Very Large Scale Integrated (VLSI) Systems,* 13(2), 211–225, February 2005.
7. W. Wu, M. A. T. Sanduleanu, X. Li, and J. R. Long, 17 GHz RF front-ends for low-power wireless sensor networks, *IEEE Journal of Solid State Circuits*, 43(9), 1909–1919, September 2008.
8. H. Eren, *Wireless Sensors and Instruments: Networks, Design and Applications*, CRC Press, Boca Raton, FL, 2006.
9. J. H. Correia, G. de Graaf, M. Bartek, and R. F. Wolffenbuttel, A CMOS optical microspectrometer with light-to-frequency converter, bus interface and stray-light compensation, *IEEE Transactions on Instrumentation and Measurement*, 50(6), 1530–1537, December 2001.
10. F. J. Naivar, CAMAC to GPIB interface, *IEEE Transactions on Nuclear Science*, 25(1), 515–519, February 1978.
11. L. Korba, S. Elgazzar, and T. Welch, Active infrared sensors for mobile robots, *IEEE Transactions on Instrumentation and Measurement*, 43(2), 283–287, April 1994.
12. R. Mukaro and X. F. Carelse, A microcontroller—Based data acquisition system for solar radiation and environmental monitoring, *IEEE Transactions on Instrumentation and Measurement*, 48(6), 1232–1238, December 1998.
13. D. R. Muñoz, D. M. Pérez, J. S. Moreno, S. C. Berga, and E. C. Montero, Design and experimental verification of a smart sensor to measure the energy and power consumption in a one-phase AC line, *Measurement: Elsevier Science Direct*, 42(3), 412–419, April 2009.
14. A. Depari, A. Flammini, D. Marioli, and A. Taroni, USB sensor network for industrial applications, *IEEE Transactions on Instrumentation and Measurement*, 57(7), 1344–1349, July 2008.
15. G. Bucci, E. Fiorucci, C. Landi, and G. Ocera, Architecture of a digital wireless data communication network for distributed sensor applications, *Measurement: Elsevier Science Direct*, 35(1), 33–45, January 2004.
16. A. Wheeler, Commercial applications of wireless sensor networks using ZigBee, *IEEE Communications Magazine*, 45(4), 70–77, April 2007.

17. L. Ferrigno, V. Paciello, and A. Pietrosanto, Performance characterization of a wireless instrumentation bus, *IEEE Transactions on Instrumentation and Measurement*, 59(12), 3253–3261, December 2010.
18. J. P. Carmo, P. M. Mendes, C. Couto, and J. H. Correia, A low-cost wireless sensor network for industrial applications, *Proceedings of Wireless Telecommunications Symposium 2009*, Praha, Czech Republic, Session D-2, pp. 1–4, 22–24, April 2009.
19. A. Bakker and J. H. Huijsing, Micropower CMOS temperature sensor with digital output, *IEEE Journal of Solid-State Circuits*, 31(7), 933–937, July 1996.
20. M. Motz, D. Draxelmayr, T. Werth, and B. Forster, A chopped hall sensor with small jitter and programmable «true power-on» function, *IEEE Journal of Solid-State Circuits*, 40(7), 1533–1540, July 2005.
21. J. Chae, H. Kulah, and K. Najafi, A monolithic three-axis micro-g micromachined silicon capacitive accelerometer, *IEEE Journal of Microelectromechanical Systems*, 14(2), 235–244, April 2005.
22. A. Arnaud and C. Galup-Montoro, Fully integrated signal conditioning of an accelerometer for implantable pacemakers, *Analog Integrated Circuits and Signal Processing*, 49(3), 313–321, 2006.
23. W. Stallings, *Data and Computer Communications*, Prentice Hall, Englewood Cliffs, NJ, 2003.
24. B. Pattan, *Robust Modulation Methods and Smart Antennas in Wireless Communications*, Prentice Hall, Englewood Cliffs, NJ, 1999.
25. H. R. Silva, L. A. Rocha, J. A. Afonso, P. C. Morim, P. M. Oliveira, and J. H. Correia, Wireless hydrotherapy smart-suit network for posture monitoring, *Proceedings of IEEE International Symposium on Industrial Electronics—ISIE* 2007, Vigo, Spain, pp. 2713–2717, June 2007.
26. N. S. Dias, J. P. Carmo, P. M. Mendes, and J. H. Correia, Wireless instrumentation system based on dry electrodes for acquiring EEG signals, *Medical Engineering & Physics*, 34(7), 1–10, 2012.
27. Taiyo Yuden Functional Modules, Taiyo Yuden Co. Online http://www.yuden.co.jp/ut/product/category/module/, accessed on February 27, 2012.
28. Linx RF modules, Linx Technologies Inc. Online http://www.linxtechnologies.com/, accessed on February 27, 2012.
29. Radiomtrix Wireless Data Transmission, Radiometrix Ltd. Online http://www.radiometrix.com/, accessed on February 27, 2012.
30. Wireless Solutions for a Connected World, Low Power Radio Solutions (LPRS Ltd). Online http://www.lprs.co.uk/, accessed on February 27, 2012.
31. J. P. Carmo, J. C. Ribeiro, P. M. Mendes, and J. H. Correia, Super regenerative receiver at 433 MHz, *Microelectronics Journal*, 42(5), 681–687, May 2011.
32. R. Morais, A. Valente, C. Couto, and J. H. Correia, A wireless RF CMOS mixed-signal interface for soil moisture measurements, *Journal Sensors and Actuators A*, 115, 376–384, September 2004.
33. A. Valente, R. Morais, C. Couto, and J. H. Correia, Modeling, simulation and testing of a silicon soil moisture sensor based on the dual-probe heat-pulse method, *Journal Sensors and Actuators A*, 115, 434–439, September 2004.
34. L. W. Couch II, *Digital and Analog Communication Systems*, 5th edn., Prentice Hall, Englewood Cliffs, NJ, 1996.
35. M. D. Weiss, J. L. Smith, and J. Bach, RF coupling in a 433-MHz biotelemetry system for an artificial hip, *IEEE Antennas and Wireless Propagation Letters*, 8, 916–919, 2009.
36. P. M. Mendes, J. H. Correia, M. Bartek, and J. Burghartz, Analysis of chip—Size antennas on lossy substrates for short-range wireless microsystems, *Proceedings SAFE 2002*, Veldhoven, the Netherlands, pp. 51–54, November 27–28, 2002.
37. J. D. Parsons, *The Mobile Radio Propagation Channel*, Pentech Press, London, U.K., 1992.
38. M. Pätzold, *Mobile Fading Channels*, Wiley–Blackwell, Chichester, U.K., 2002.
39. M. Blaunstein and J. B. Andersen, *Multipath Phenomena in Cellular Networks*, Artech House Publishers, Norwood, MA, 2002.

40. H. L. Bertoni, *Radio Propagation for Modern Wireless Systems*, Prentice Hall, Englewood Cliffs, NJ, 2000.
41. L. A. Rocha, J. A. Afonso, P. M. Mendes, and J. H. Correia, *A Body Sensor Network for E-Textiles Integration*, *Proceedings of Eurosensors XX*, Gothenburg, Sweden, pp. 1–4, September 2006.
42. Crossbow. (2009). Wireless measurement systems, Crossbow Inc. Online http://www.xbow.com, accessed on February 2012.
43. J. A. Afonso, L. A. Rocha, H. R. Silva, and J. H. Correia, MAC protocol for low-power real-time wireless sensing and actuation, *Proceedings of IEEE International Conference on Electronics, Circuits and Systems—ICECS 2006*, Nice, France, pp. 1248–1251, December 2006.
44. J. A. Gutierrez, M. Naeve, E. Callaway, M. Bourgeois, V. Mitter, and B. Heile, IEEE 802.15.4: Developing standards for low-power low-cost wireless personal area networks, *IEEE Network*, 5(15), 12–19, September/October 2001.
45. C. Enz, N. Scolari, and U. Yodprasit, Ultra low-power radio design for wireless sensor networks, *Proceedings of the IEEE International Workshop on Radio Frequency Integration Technology: Integrated Circuits for Wideband Communication and Wireless Sensor Networks*, Singapore, pp. 1–17, December 2005.
46. C. C. Enz, A. El-Hoiydi, J. D. Decotignie, and V. Peiris, WiseNET: An ultralow—Power wireless sensor network solution, *IEEE Computer*, 378, 62–70, August 2004.
47. J. P. Carmo and J. H. Correia, Low-power/low-voltage RF microsystems for wireless sensors networks, *Microelectronics Journal*, 40(12), 1746–1754, December 2009.
48. E. H. Callaway Jr., The physical layer. In: *Wireless Sensor Networks: Architectures and Protocols,* Chapter 3, pp. 41–59, CRC Press, Boca Raton, FL, 2004.
49. N. S. Dias, J. P. Carmo, P. M. Mendes, and J. H. Correis, A low-power/low-voltage CMOS wireless interface at 5.7 GHz with dry electrodes for cognitive networks, *IEEE Sensors Journal*, 11(3), 755–762, March 2011.
50. A. El-Hoiydi, C. Arm, R. Caseiro, S. Cserveny, J. D. Decotignie, C. Enz, F. Giroud et al., The ultra low-power WiseNET system, *Proceedings Design, Automation and Test in Europe, DATE'06*, Munich, Germany, pp. 1–5, March 6–10, 2006.
51. J. A. Afonso, H. D. Silva, P. Macedo, and L. A. Rocha, An enhanced reservation-based MAC protocol for IEEE 802.15.4 networks, *Sensors*, 11(4), 3852–3873, April 2011.
52. W. Y. Lee, *Wireless and Cellular Communications*, 2nd edn., McGraw-Hill, New York, 1998.
53. E. Callaway, P. Gorday, L. Hester, J. A. Gutierrez, M. Naeve, B. Heile, and V. Bahl, Home networking with IEEE 802.15.4: A developing standard for low-rate wireless personal area networks, *IEEE Communications Magazine*, 40(8), 2–9, August 2002.
54. Smart RF CC2420, 2.4 GHz IEEE 802.15.4/ZigBee-ready RF transceiver, Texas Instruments Incorporated. Online http://www.ti.com/, accessed on February 27, 2012.
55. Smart RF CC2430, A true system-on-chip solution for 2.4 GHz IEEE 802.15.4/ZigBee, Texas Instruments Incorporated. Online http://www.ti.com/, accessed on February 27, 2012.
56. Dust, Dust Networks Inc. Online http://www.dust-inc.com/, accessed on February 27, 2012.
57. Sensicast, Sensicast Systems. Online http://www.sensicast.com/, accessed on February 27, 2012.
58. N. S. Dias, J. F. Ferreira, C. P. Figueiredo, and J. H. Correia, A wireless system for biopotential acquisition: An approach for non-invasive brain-computer interface, *Proceedings of IEEE International Symposium on Industrial Electronics—ISIE 2007*, Vigo, Spain, pp. 2709–2712, June 4–7, 2007.
59. A. C. M. van Rijn, A. Peper, and C. A. Grimbergen, High-quality recording of bioelectric events, Part 1: Interference reduction, theory and practice, *Medical & Biological Engineering & Computing*, 28(5), 389–397, September 1990.
60. R. F. Yazicioglu, T. Torfs, P. Merken, J. Penders, V. Leonov, R. Puers, B. Gyselinckx, and C. V. Hoof, Ultra low-power biopotential interfaces and their applications in wearable and implantable systems, *Microelectronics Journal*, 40(9), 1313–1321, September 2009.

61. J. P. Carmo, L. M. Goncalves, and J. H. Correia, Thermoelectric microconverter for energy harvesting systems, *IEEE Transactions on Industrial Electronics*, 57(3), 861–867, March 2010.
62. Sensium TZ1030, Ultra low power smart sensor interface and transceiver platform, Toumaz Technology Limited. Online http://www.toumaz.com, accessed on February 27, 2012.
63. T. Karacolak, A. Z. Hood, and E. Topsakal, Design of a dual band implantable antenna and development of skin mimicking gels for continuous glucose monitoring, *IEEE Transactions on Microwave Theory and Techniques*, 54(4), 1001–1008, April 2008.
64. E. Y. Chow, Y. Ouyang, B. Beier, W. J. Chappell, and P. P. Irazoqui, Evaluation of cardiovascular stents as antennas for implantable wireless applications, *IEEE Transactions on Microwave Theory and Techniques*, 57(10), 2523–2532, October 2009.
65. F. J. O. Rodrigues, J. H. Correia, and P. M. Mendes, Modeling of a neural electrode with MEMS magnetic sensor for telemetry at low frequencies, *Proceedings MicroMechanics Europe, MME 2009*, Toulouse, France, pp. D19/1–D194, September 20–22, 2009.

Partial List of Manufacturers

BWI Eagle, Inc., 105 Bonnie Dr., Butler, PA 16002, Phone: 888-366-7850 (toll free), Fax: 724-283-5939, bwieagle.rtrk.com/?scid=1808652&rl_alt=h…

Cognex Corp., 1 Vision Dr., Natick, MA 01760-2059, Phone: 508-650-3100, 800-677-2646 (toll free), Fax: 508-650-3344, www.cognex.com

Cooper Instruments, 400 Belle Air Lane, P.O. Box 3048, Warrenton, VA 20188, Phone: 540-349-4746, 800-344-3921 (toll free), Fax: 540-347-4755, www.cooperinstruments.com

Micro-Epsilon America, 8120 Brownleigh Dr., Raleigh, NC 27617, Phone: 919-787-9707, Fax: 919-787-9706, www.micro-epsilon.us

Microstrain, Inc., 459 Hurricane Lane, Suite 102, Williston, VT 05495, Phone: 802-862-6629, 800-449-3878 (toll free), Fax: 802-863-4093, www.microstrain.com

Siemens Corp, 300 New Jersey Avenue, N.W., Suite 1000, Washington, DC 20004-2611, Phone: 202-434-4800, 800-SIEMENS (toll free), Fax: 202-347-4015, www.usa.siemens.com/entry/en/



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

86

Wireless Sensor Node Hardware

Michael Healy
University of Limerick

Thomas Newe
University of Limerick

Elfed Lewis
University of Limerick

86.1 Introduction **86**-1
86.2 Development History of Sensor Node Hardware **86**-1
86.3 Sensor Node Components **86**-5
Computing Subsystem • Communication Subsystem • Power Systems • Sensing Subsystems
86.4 Future Directions **86**-10
86.5 Summary **86**-12
References **86**-12
Partial List of Vendors and Suppliers **86**-15

86.1 Introduction

The fundamental building blocks of wireless sensor networks (WSNs) are the sensor nodes themselves. For the last 10–12 years, a number of research institutions and companies have been designing and producing nodes. However, the capabilities of the resulting nodes vary widely, but only within certain bounds. All the nodes have a number of features in common, i.e., a low-power microcontroller, a low-power, short-range, radio, a power supply, and either onboard sensors or the capability to attach sensors. However, it is the variance within these bounds that complicates application development.

The resources of conventional computing systems, such as personal computers and mobile phones, increase almost exponentially, and cost is reduced as technology improves. This has not happened with sensor nodes and, indeed, is not expected to happen going forward. WSN deployments are likely to become more widespread, thus allowing economies of scale to come into play at the manufacturing level, driving costs down. Today most sensor nodes consist of a number of components that work together, but there is a move toward system-on-a-chip (SoC) devices that have, at the least, the communication, processing, and sensing subsystems all on a single piece of silicon. This consolidation of component parts will also help to drive down costs.

Table 86.1 lists nodes that have been developed from 1998 to the present day. This list is extensive but perhaps not comprehensive as a large number of universities and research institutions would have created their own nodes from scratch, tailored to their own particular research requirements, and the details of these nodes would not be available.

86.2 Development History of Sensor Node Hardware

Sensor node hardware has always been a very important factor in WSN research and development. The earliest known wireless sensor nodes, from the late 1960s and early 1970s, were the units used in operation Igloo White [1]. These units were used to detect troop movements along the Ho Chi Minh trail,

TABLE 86.1 Selection of Sensor Nodes

Platform	CPU	Clock (MHz)	RAM/Flash/ EEPROM	Radio Transceiver	BW (kbps)	Freq. (MHz)	OS	Year
WeC	Atmel AT90LS8535	4	512/8K/32K	RFM TR1000	10	916.5	TinyOS	1998
Rene 1	Atmel AT90LS8535	4	512/8K/32K	RFM TR1000	10	916.5	TinyOS	1998
AWAIRS 1	Intel StrongARM SA1100	59–206	1M/4M	Conexant RDSSS9M	100	900	MicroC/ OS	1999
μAMPS	Intel StrongARM SA1100	59–206	1M/4M	National LMX3162	1000	2400	μOS	1999
Rene 2	Atmel ATmega 163	8	1K/16K/32K	RFM TR1000	10	916.5	TinyOS	2000
Dot	Atmel ATmega 163	8	1K/16K/32K	RFM TR1000	10	916.5	TinyOS	2000
Mica	Atmel ATmega 128L	4	4K/128K/512K	RFM TR1000	40	916.5	TinyOS	2001
BT node	Atmel ATmega 128L	8	4K/128K/4K	ZV4002 BT/ CC1000	1000	2400	TinyOS	2001
SpotON	Dragonball EZ	16	2M/2M	RFM TR1000	10	916.5		2001
Smart-its	PIC 18F252	8	3K/48K/64K	Radiometrix	64	433	Smart-its	2001
Mica2	Atmel ATmega 128L	8	4K/128K/512K	Chipcon CC1000	38.4	900	TinyOS	2002
Mica2Dot	Atmel ATmega 128L	4	4K/128K/512K	Chipcon CC1000	38.4	900	TinyOS	2002
iBadge	Atmel ATmega 128L	8	4K/128K	Ericsson ROK101007 BT	1000	2400	Palos	2002
CENS Medusa MK2	Atmel ATmega 128L/Atmel AT91FR4081	4/40	4K/32K 136K/1M	RFM TR1000	10	916	Palos	2002
iMote	Zeevo ZV4002 (ARM)	12–48	64K/512K	Zeevo BT	720	2400	TinyOS	2003
U3	PIC 18F452	0.031–8	1K/32K/256	CDC-TR-02B	100	315	Pavenet	2003
Spec	8-bit AVR-like RISC	4–8	3K	FSK Transmitter	100		TinyOS	2003
RFRAIN	Chipcon CC1010 (8051)	3–24	2K/32K	Chipcon CC1010	76.8	0.3–1000	RFRAIN Libraries	2003
Nymph	Atmel ATmega 128L	4	4K/128K/512K	Chipcon CC1000	38.4	900	Mantis	2003
Telos	TI MSP430F149	8	2K/60K/512K	Chipcon CC2420	250	2400	TinyOS	2004
MicaZ	Atmel ATmega 128L	8	4K/128K	Chipcon CC2420	250	2400	TinyOS	2004
CIT sensor node	PIC 16F877	20	368/8K	Nordic nRF903	76.8	868	TinyOS	2004
BSN node	TI MSP430F149	8	2K/60K/512K	Chipcon CC2420	250	2400	TinyOS	2004
MITes	nRF24E1 (8051)	16	4K/512	Nordic nRF24E1	1000	2400	Custom	2004
AquisGrain	Atmel ATmega 128L	4	4K/128K/512K	Chipcon CC2420	250	2400	Custom	2004
RISE	Chipcon CC1010 EM (8051)	3–24	2K/32K	Chipcon CC1010 EM	76.8	0.3–1000	TinyOS	2004
Particle2/29	PIC 18F6720	20	4K/128K/512K	RFM TR1001	125	868.35	Smart-its	2004
Pluto	TI MSP430F149	8	4K/60K/512K	Chipcon CC2420	250	2400	TinyOS	2004
DSYS25	Atmel ATmega 128L	4	4K/128K	Nordic nRF2401	1000	2400	TinyOS	2004

TABLE 86.1 (continued) Selection of Sensor Nodes

Platform	CPU	Clock (MHz)	RAM/Flash/ EEPROM	Radio Transceiver	BW (kbps)	Freq. (MHz)	OS	Year
EnOcean TCM120	PIC 18F452	10	1.5K/32K/256	Infineon TDA 5200	120	868	TinyOS	2005
EyesIFXv2	TI MSP430F1611	8	10K/48K	Infineon TDA 5250	64	868	TinyOS	2005
iMote2	Intel PXA 271	13–104	256K/32M	Chipcon CC2420	250	2400	TinyOS	2005
uPart0140 ilmt	rfPIC 16F675	4	64/1K	rfPIC 16F675	19.2	868	Smart-its	2005
TelosB/Tmote Sky	TI MSP430F1611	4/8	10K/48K/1M	Chipcon CC2420	250	2400	TinyOS	2005
Ember RF module	Atmel ATmega 128L	8	4K/128K	Ember 250	250	2400	EmberNet	2005
XYZ sensor node	OKI ML67Q500x (ARM/ THUMB)	1.8–57.6	4K/256K/512K	Chipcon CC2420	250	2400	SOS	2005
Ant	TI MSP430F1232	8	256/8K	Nordic nRF24AP1	1000	2400	Ant	2005
ProSpeckz II	Cypress CY8C2764	12	256/16K	Chipcon CC2420	250	2400	Speckle net	2005
Fleck	Atmel ATmega 128L	8	4K/2G/512K	Nordic nRF903	76.8	902–928	TinyOS	2005
Sun SPOT	Atmel AT91RM9200	180	512K/4M	Chipcon CC2420	250	2400	Squawk VM (Java)	2005
Tinynode 584 [16]	TI MSP430F1611	8	10K/560K	Semtech XE1205	153	433/ 568/ 915	TinyOS	2005
MSB430 [17]	TI MSP430F1612	8	5K/55K	TI CC1020	153	402–470/ 804/960	Contiki	2005
ECO [18]	nRF24E1 (8051)	16	4K/512/32K	Nordic nRF24E1	1000	2400	Custom	2006
Shimmer [19]	TI MSP430F1611	4/8	10K/2G	WML-C46A BT/ CC2420	250	2400	TinyOS	2006
Firefly [20]	ATmega 32L	8	2K/32K/2G	Chipcon CC2420	250	2400	NanoRK	2006
IRIS [21]	Atmel ATmega 1281	8	8K/640K/4K	Atmel AT86RF230	250	2400	TinyOS	2007
Mulle [22]	Renesas M16C	10	31K/384K	WML-C46A BT	100	2400	TinyOS Contiki	2008
Epic [23]	TI MSP430F1611	4	10K/2M	TI CC2420	250	2400	TinyOS	2008
Ubi-Mote2 [24]	TI MSP430F2618	16	8K/116K	TI CC2520	250	2400	TI ZStack	2009
EZ430-RF2500 [25]	TI MSP430F2274	16	1K/32K	TI CC2500	500	2400	Custom	2009
Zolertia Z1 [26]	TI MSP430F2617	16	8K/2M	TI CC2420	250	2400	TinyOS Contiki	2010
Shimmer2R [27]	TI MSP430F1611	4/8	10K/2G	RN-42 BT/ CC2420	250	2400	TinyOS	2010
G-Node [28]	TI MSP430F2418	16	8K/1M/116K	TI CC1101	500	868	TinyOS	2010
Waspmote [29]	Atmel ATmega 1281	8	8K/2G/4K	Xbee/eUnistone 31308/2 BT/ Hilo GPRS	250	2400/868/ 900/850/ 1800/1900	Custom	2010

Source: Expanded and updated from Healy, M. et al., Wireless sensor node hardware: A review, *The Seventh IEEE Conference on Sensors (IEEE SENSORS 2008)*, Lecce, Italy, October 26–29, pp. 621–624, 2008.



FIGURE 86.1 Smart dust sensor node that includes optical communication, sensors, and power. Approximately the size of a grain of rice.

from North Vietnam through Laos and into South Vietnam. However, the cost of this operation was astronomical, and the project was not continued after the Vietnam War.

The next documented example of wireless sensor node hardware comes from approximately 15 years later, in the mid-1980s, and was part of the DARPA-funded DSN [2,3] project. The wireless node was so large that it needed to be mounted on a truck for mobility purposes. The substantial power requirements were met by a dedicated generator, which was mounted at the rear of the vehicle in an acoustically quieted enclosure.

Progress on wireless sensor node hardware was equally slow for the next 10–15 years. The primary reasons for this were due to the cost and complexity of constructing the nodes from fully custom hardware and software. Examples include the low-power, wireless, integrated, microsensor (LWIM) node from the University of California, Los Angeles (UCLA) [5], constructed circa, 1996, and the smart dust node shown in Figure 86.1 [6]. While these devices were significantly smaller than the previous examples, they were both fully custom affairs, from the silicon level upward.

The revolution in wireless sensor node hardware, which can be seen to have kick-started all current WSN research, was to make use of commercial-off-the-shelf (COTS) components to construct these nodes.

The Hidra node from Rockwell, shown in Figure 86.2, was arguably the first commercially available sensor node, released in 1998. This was 3.5 in. × 3.5 in. × 3 in. in size, but primarily because of the StrongARM 1100 processor running at 133 MHz, either a large external power supply was required to operate it or, if run off battery power, the operating time was measured in minutes [7].

A more significant progression from 1998 was the development of the WeC node, the first entry in Table 86.1. This was developed as part of the proof-of-concept leg of the Smart Dust project [8] at the University of California, Berkeley. This node is significant because it is the arguably the first *low-power* wireless sensor node and is essentially the ancestor of most of the current sensor nodes available today. Approximately 30 units of the WeC mote were constructed, before being superseded by the Rene mote in 1999, of which there are 850+ built. The Rene mote in turn was further improved upon, to give the Dot mote in 2000 and the Mica mote in 2001 [6,9] (Figure 86.3). The next iterations, the Mica2 and Mica2Dot motes, from 2002 onward, were commercialized by Crossbow Technology, Inc. [10]. The Mica2 and its direct descendants, the MicaZ and Iris mote, are still in active use today.



FIGURE 86.2 Rockwell WINS and Hidra nodes.

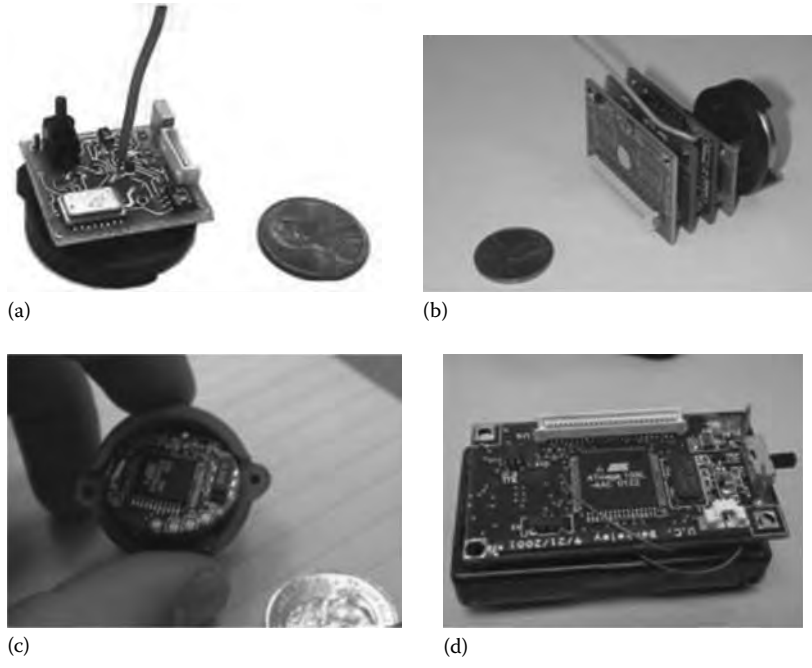


FIGURE 86.3 Early Berkeley motes. (a) WeC, (b) Rene, (c) Dot, (d) Mica.

86.3 Sensor Node Components

Despite the variances, almost all sensor nodes are composed of four basic subsystems: a computing subsystem, a communication subsystem, a power subsystem, and a sensing subsystem.

86.3.1 Computing Subsystem

The computing subsystem is responsible for controlling the components of the sensor node and performs any required computations. It consists of two subunits: the processor unit and the storage unit. An important aspect of processors in sensor nodes is the different operational modes, usually active, idle,

and sleep modes as a minimum. This is important so as to preserve power as much as possible without impeding the operation of the processor when it is required.

As can be seen from Table 86.1, the processors used in sensor nodes have, to date, ranged from ultralow-power 8 bit microcontrollers, which can be clocked at speeds as low as 31 kHz, to powerful 32 bit ARM processors that can be clocked at greater than 200 MHz. The price that needs to be paid for using faster and more powerful processors is increased power consumption and purchasing cost. The trend in new generations of sensor nodes is to use low-power processors that focus on very low energy consumption while inactive. As for most WSN applications, the processor is in sleep mode for over 99% of the time. The one major exception to this is the Sun SPOT mote, which uses a powerful processor; the reason for this is to enable the mote to be easily programmed using Java.

An anomaly that can be gleaned from Table 86.1 is that wireless sensor node hardware seems to be very slow to update the processor being used. By far, the most popular processor used in WSN hardware today is the MSP430F1611 from Texas Instruments [11], followed by the Atmel ATmega 128L [12]. While well suited to WSN use, these processors were already well established in the mid-2000s when they were first employed for this purpose. More efficient, faster, cheaper, and lower-power versions of both have been available from their own family of devices for a number of years, but the primary reason they have not been used in this context is due to the lack of open-source compiler support. The WSN operating systems described in the next chapter depend on ports of the GNU C compiler (GCC) [13] tool chains for their build systems, msp430-gcc [14] and avr-gcc [15] to be specific. In recent years, these chip family-specific tool chains have not been updated to support the product line extensions in a timely manner, making it difficult to include the latest and greatest devices on sensor motes. However, this is starting to change, both because of a resurgence of interest from developers who are willing to work on these tool chains, and also, importantly, the manufacturers themselves are starting to recognize the value of these tool chains to their business and as a result are supporting their development.

The storage unit of a node usually consists of both flash memory, containing the program code for the node, and random access memory (RAM), which stores sensed information and any data needed for computations. Some motes also have nonvolatile storage for off-line data capture for later retrieval. Additionally a number of motes have built-in microSD/miniSD/SD card interfaces, or attachable daughter cards to add this functionality, as can be seen in Figure 86.4, allowing up to 2 GB or more of data storage.

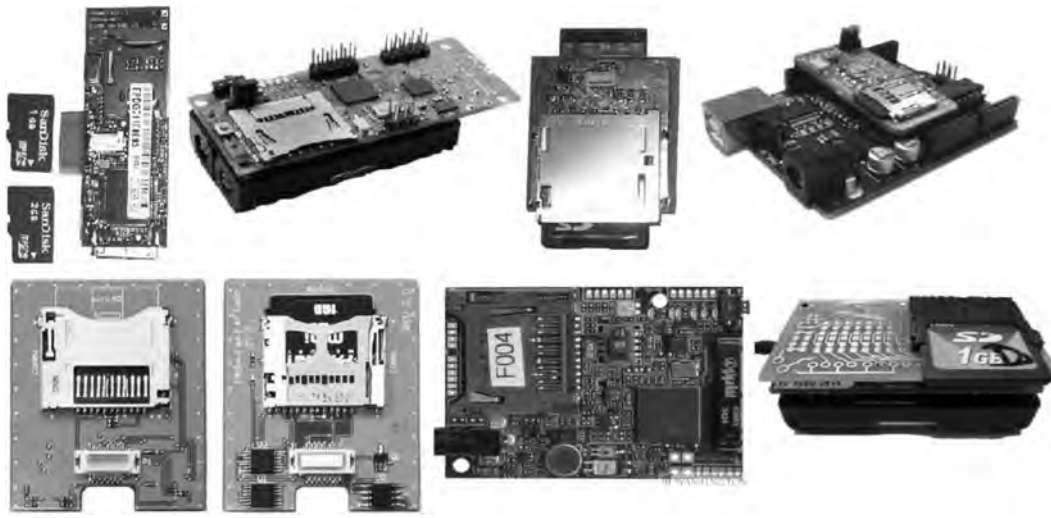


FIGURE 86.4 Selection of motes with SD card support. Top row (L–R): Shimmer, Firefly, Carnivore [30], and SquidBee with expansion board [31]. Bottom row (L–R): expansion board for Sun SPOT (2 miniSD slots, one on top and bottom) [32], expansion board for iMote2 [33], and Luster (MicaZ expansion board) [34].

As a result, there can be anything up to an order of magnitude or more of a disparity between the storage capabilities of different sensor nodes, the choice being made on the basis of the required and desirable storage capacity versus extra cost.

The Carnivore mote shown in Figure 86.4 is tailored to specific application use for wildlife monitoring and is not generally available to developers. It is based on the MSP430 processor and the CC2420 communications platform. The SquidBee mote has generally been preceded by the Waspmote.

86.3.2 Communication Subsystem

The communication subsystem is required to enable the sensor nodes to communicate with each other and with a base station. Generally the communication subsystem is a short-range radio, but the use of infrared communication, ultrasound, and inductive fields has also been explored [35].

While cheap and suitable for implementation with low power consumption, the major drawback with infrared communication is the need for an unobstructed line of sight between the communicating devices. Ultrasound is usually ruled out due to the fact that the network coordinator requires high energy and the form factor of the equipment is also a major problem (i.e., miniaturization). Despite being used extensively for radio-frequency identification (RFID) applications, inductive field communication is also not suitable due to the high-energy requirements of the network coordinator and also because of its very low range. Radio-frequency (RF) communication is ideal for sensor nodes because it is not limited by line of sight and current technology allows implementation of low-power radio transceivers with data rates and ranges scalable according to the application. The RF spectrum is a scarce resource and is regulated by most governments, but there are unlicensed bands, known as the industrial, scientific, and medical (ISM) bands, inside which it is free to operate once the device conforms to the rules that control the band.

Table 86.1 shows a wide variety of radios have been used on sensor nodes. Many current motes use a radio chip that conforms to the IEEE 802.15.4 standard [36], and some nodes use Bluetooth [37] as an alternative. The argument for using Bluetooth is that it allows easy interoperability with a range of existing devices such as mobile phones and laptop computers without the need for additional hardware. The price that needs to be paid for this interoperability is increased energy consumption. Older generations of nodes tended to use radios that conformed to proprietary standards or to no standards at all. Some newer-generation motes are also going in this direction, e.g., those that use the Texas Instruments CC2500 chip [38]. The advantage these nonconforming radios offer is that, because they do not necessarily need to follow ageing standards, they can allow higher throughput or lower power consumption, or both, than the standards-compliant radios. Of course the downside is the lack of interoperability with radios and devices from other manufacturers. Finally, the standards-compliant radios tend to be packet-level devices, whereas the noncompliant radios tend to allow manipulation at the bit level. Bit-level devices allow for more flexibility, but the cost that needs to be paid is that they are usually more difficult to use. However, as the WSN operating systems try to abstract the radio operation so that the application developer does not need to worry about it, this is not a major area for concern. On the other hand, assuming 802.15.4, Bluetooth and other applicable standards are well designed, and there is no reason to date to think otherwise; in the majority of applications, it makes no sense to go to the trouble of creating a communication protocol from scratch, i.e., to reinvent the wheel, especially as the application developer is unlikely to be as experienced or as expert in the field as the committees that developed the standards in the first place, potentially resulting in, to continue the metaphor, a suboptimal wheel design.

An important point to note is that the ISM bands were originally chosen to allow electromagnetic interference (EMI) emissions of products to bleed into without causing problems for other devices. As this is still the case, any device, even ones that use no RF communication, is a potential source of interference. Also humans resonate at 900–1000 MHz, and as the ISM bands are multiples of our fundamental frequency, crowds of people, or even a single person in an inconvenient location, can have a major impact

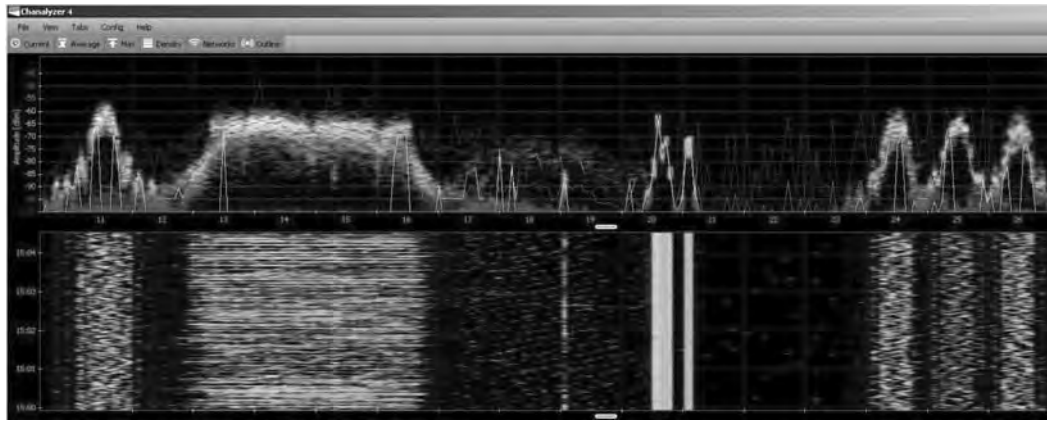


FIGURE 86.5 Density graph and waterfall trace of the 2.4 GHz band.

on the effectiveness of the communications [39]. However, WSNs still tend to operate inside these bands in order to keep costs down. Other problems that should be considered is that the unlicensed bands change from country to country, and there are lots of other devices that also operate in these bands, e.g., wireless local area networks (WLAN), wireless keyboards, home automation systems, wireless surveillance cameras, and digital enhanced cordless telecommunication (DECT) phones, providing additional interference.

Figure 86.5 shows a snapshot of the 2.4 GHz band from a typical industrial environment captured using a Wi-Spy spectrum analyzer [40]. This is the ISM band that the majority of current wireless sensor nodes operate in, a fact that is obvious from looking at Table 86.1. The *x*-axis of both the density graph and the waterfall trace shows the sixteen 802.15.4 channels. When this trace was taken, there was only one WLAN in range, spanning channels 13–16. As a comparison, the activity on channels 11, 24, 25, and 26 is almost entirely from wireless sensor nodes (Shimmer2R devices) transmitting 98 byte 802.15.4 packets at 10 Hz. Eight time-multiplexed nodes were operating in each of channels 11, 24, and 26, and 5 nodes in channel 25, with all 29 nodes being within a meter of the Wi-Spy receiver and all transmitting at their highest power setting. The Wi-Fi router was approximately 20 m away, on the other side of a partition wall. The relatively high-intensity, and very constant, transmissions in channels 20 and 21 were from an unknown source.

This trace shows that even in such close proximity, the 802.15.4 radio can barely match the more distant Wi-Fi radio in intensity level. Also, the amount of data being sent per channel by the 802.15.4 networks would be considered close to the upper bounds of what is realistically achievable even in uncongested channels while maintaining packet reception rates of greater than 90%.

Despite this, the channel bandwidth consumed by the Wi-Fi network is still significantly higher, as can be seen in Table 86.2.

TABLE 86.2 Channel Duty Cycle of 2.4 GHz Band Trace Shown in Figure 86.5

Channel	Duty Cycle (%)	Channel	Duty Cycle (%)
11	21.0	19	1.6
12	0.4	20	20.3
13	29.9	21	13.3
14	38.2	22	0.1
15	37.9	23	0.1
16	21.0	24	15.9
17	0.8	25	10.8
18	2.3	26	15.1

The consequence of this is that when the same 802.15.4 networks were run under these conditions in the channels occupied by the Wi-Fi network, or the unknown transmitters in channels 20 and 21, the packet reception rate dropped from greater than 98% to less than 50% in all cases and less than 10% in some of the channels. This goes to show how great an impact background noise can have on the communication in WSNs and why the RF used is a very important deployment decision.

86.3.3 Power Systems

The power subsystem usually consists of a battery that supplies power to the sensor node.

For many applications of WSNs, the required lifetime of the sensor nodes may be weeks, months, or even years. Battery recharging or replacement is unlikely to be feasible, especially in large-scale deployments with thousands of widely dispersed nodes, or for nodes placed in hazardous environments. For these reasons the developer must ensure every aspect of the network, such as communication algorithms, localization algorithms, and sensing devices, must be as efficient as possible in their power usage.

A power generator may also be included to recharge the battery on site. Photovoltaic, motion/vibration, and thermoelectric energy conversion are all possible sources of power, depending on the location of the node [41].

For a significant subset of WSN applications, the power supply is not a major limiting factor, a category many current deployments fit into. WSNs enable very remote, very-large-scale deployments but are also useful in more mundane situations. For example, for industrial monitoring, while a suitable wired networking infrastructure, or WLAN infrastructure, might not be present in a factory and excessively costly to install, very often a power infrastructure is already present, i.e., the main power supply. Another example is that for many health applications, the sensor nodes are easily accessible, allowing recharging or replacement of batteries at regular intervals. This is a fact that is largely ignored by the research community, primarily as limited energy capacity results in the more difficult and more interesting research problems. However, as most commercially viable sensor networks fall into this category, at least for the moment, suboptimal solutions for the major issues that affect these networks are the only ones currently available, especially along the lines of data aggregation, querying, etc.

That said, in most cases the available power is still the driving factor behind many decisions made about WSN node deployments, including whether or not to include often optional functionality like security, which is why any optional functionality solution needs to be as energy efficient as possible.

86.3.4 Sensing Subsystems

Sensor transducers translate physical phenomena to electrical signals. Therefore, the sensing subsystem of the node is its link to the outside world. The output of the sensors may be digital or analog signals. If the output is analog, the node must also include an analog-to-digital converter (ADC) in order to allow the processor to read the data. Some nodes have sensors built in, such as temperature, humidity, and light sensors, on the Tmote Sky [42] and a three-axis accelerometer on the Shimmer [19], but many do not, instead providing suitable ports and connectors to allow a variety of sensors to be attached for more versatility.

Attaching a sensor with a digital output to a mote is generally a simple exercise, but analog sensors offer some complications. If a node does not have a built-in ADC, one must be included, and without careful design, external ADCs can often send an excessive number of interrupts to the microcontroller unit (MCU), seriously impacting on the other functions of the node. The two most commonly used processors on sensor nodes to date, the Atmel ATmega 128L and the Texas Instruments MSP430, both contain integrated ADCs. However, the resolution and quality of these ADCs can be insufficient for some applications, with the ATmega 128L supplying a 10 bit ADC [12] and the MSP430 offering a 12 bit ADC [11]. Additionally direct memory access (DMA) by the ADC subsystem can be important

for low-jitter, fast-timing accurate sampling, a feature found on the MSP430 processers, but not on the Atmel ATmega devices.

An important feature of sensors that are attached to sensor nodes is the ability to turn on, stabilize, and be read quickly and also power down in a timely manner. When a sensor is switched on, it is consuming power, and as already mentioned, preserving battery life on the nodes is often a prime concern.

A large variety of add-on boards are commercially available for current WSN nodes, capable of measuring a wide variety of phenomena. The following list is a selection of the sensors available from two wireless sensor node vendors, Shimmer Research [43] and Libelium [44], but it is representative of the range available:

- Pressure, acceleration, rotation, vibration, magnetic field, strain, electroencephalography (EEG), electromyography (EMG), electrocardiography (ECG), galvanic skin response (GSR), location (GPS), gas (CO, CO_2, NO_2, etc.), presence (passive infrared), noise, moisture, ultrasound, luminosity, and radiation

This is by no means an exhaustive list as almost any possible sensor could conceivably be connected to a wireless sensor node with the correct interface hardware being available.

86.4 Future Directions

The last few years of the twentieth century and the first five or so years of the twenty-first century saw significant changes in the field of wireless sensor node hardware. However, since then, advancements have been much slower and much more incremental in nature.

For example, the Tinynode 584 from Shockfish [16] was first released in 2005 and is still being sold today in exactly the same form in 2012. The incremental nature of the changes can be seen in the Shimmer family of motes, initially designed at Intel [45], but now developed and sold by Shimmer Research [27]. The current version, the Shimmer2R is the third generation over a 5 year period, but it still uses most of the same parts and has identical functionality and performance as the original Shimmer mote. The entirety of the changes over that period was as follows:

- The accelerometer sensor was changed, but this was only updated because the original chip is at end of life (EOL).
- The Bluetooth radio module employed was changed for the same reason.
- The original Shimmer could not be switched off, but this was a common requirement to preserve battery life when not in use, as the rechargeable battery is soldered directly to the main board and so cannot be easily removed, so power on/off functionally was added.
- Additionally a passive tilt/vibration sensor was added to allow the shimmer to enter the lowest power state possible when idle.
- Power control of the peripherals on the Shimmer was improved, i.e., the ability to power off the Bluetooth radio and the microSD card when not in use.
- Finally, a patented bypass system was added to allow data to be read off the microSD card quickly without having to remove it from the Shimmer device [46].

This trend of slow, incremental changes of existing offerings is likely to continue for the foreseeable future, with the primary aim of making the devices cheaper and more energy efficient, but not necessarily giving them more processing power or more memory capacity.

One of the areas that advancing technology is making an impact with wireless sensor node hardware, and poised to make a much bigger impact, is the components of the power subsystem. This progression is being driven by the power requirements of modern mobile phones, especially the demands of smart phones where higher-capacity batteries are continually being shrunk to fit into smaller packages. Again taking the Shimmer node as an example, when first produced, it was shipped with a rechargeable 280 mAh battery but is now supplied with a 450 mAh battery in the same form factor [47]. Additionally

there have been a number of recent advancements in the field of energy harvesting that should help to extend battery life for WSNs [48–50].

Another trend that is starting to become evident is the emergence of more focused wireless sensor node devices. All the devices listed in Table 86.1 are generic prototyping platforms that allow researchers the flexibility to test new ideas and expand into new application domains. However, this flexibility is expensive, with features and components being present on a mote that are not necessary for a given application. As a result, devices designed specifically, and only, for a particular application domain are starting to become available. For example, the medical domain has offerings from Toumaz [51] and Zepher [52]. The latest example from the Berkeley family of sensor motes, the Epic [23], embodies this specificity by designing each of the sensor node subsystems as stand-alone interchangeable modules. Given that the Epic mote has been released as an open-source hardware (OSHW) platform, anybody can come along and take the required, well-tested, parts and build a custom mote for a particular purpose, including only the required components.

In the general case, there is a need for easy-to-use WSNs as a whole, but in some cases at least, this requirement extends down to the hardware itself. As a result, there are a few recently released hardware platforms that are marketed as easy to set up and use, but this ease of use comes at the cost of the devices being more "locked down," i.e., while the device might be easy to use within certain bounds, it cannot be used outside these bounds even by advanced users. Examples include the WSN hardware from National Instruments, which integrates with their LabVIEW software environment [53]; the very specific targeted wireless sensors from Monnit [54]; and the disposable, self-adhesive wireless sensors from GreenGoose [55]. These types of wireless sensor nodes are going to become more popular as the technology matures and the exact requirements of the different application domains become better understood.

A different approach to making wireless sensor nodes easier to use is to make them programmable using the Java language [56]. The reasoning behind this advancement is that there are many more programmers in the world capable of, and comfortable, using this language than the C [57] or supersets of C, the languages traditionally used to develop applications on wireless sensor nodes. It is also thought that Java would make this application development easier all round, both for experienced software developers and beginners, thus enabling more people to enter the field. This venture was initially driven by Sun Microsystems with their Sun SPOT mote [58] and later by Sentilla Corporation [59] with their JCreate mote. However, despite a lot of excitement in this area between 2005 and 2008, there has been very little news or developments in the area since then. In fact, Sentilla no longer seem to be in the hardware business, instead focusing entirely on application "solutions" [60]. Whether or not Java, or any other higher-level programming language, can be used to program embedded, mote class devices in an efficient and comprehensive manner is still an open question.

Complete SoC wireless sensor nodes have been around for a number of years, such as the original equipment manufacturer (OEM) edition of the Iris node from Memsic [61] and Dust Networks mote-on-chip range [62]. Recently some large processor manufacturers have started producing products that, while not complete SoC wireless sensor nodes, have at least got the computing and communication subsystems on a single chip. An example of this is the CC2531 from Texas Instruments, which includes an 802.15.4 radio transceiver and an 8051 MCU on a single chip [63]. Texas Instruments already have a number of other similar products available, and other manufacturers are likely to follow suit to keep pace with the current movement of adding wireless functionality to any device/product possible.

These same microcontroller manufacturers who also produce low-power, short-range radios are starting to produce node-class devices, both as development kits and also for the hobbyist market. To date these solutions have been much cheaper than any of the dedicated wireless sensor node vendors, albeit not as fully featured or as flexible. These offerings include the EZ430-RFxxxx series from Texas Instruments [25] and the RzRaven from Atmel [64]. This trend is likely to continue as these large companies are recognizing the business opportunities in the "hobbyist" market. Additionally these devices are being targeted toward the education market, which has always been an important segment for these businesses.

Current WSNs sense their environment and report back to the operator. It is envisioned that in the future, WSNs will also be able to influence this environment by having actuators attached to the nodes.

These actuators can then either be controlled by the operator or alternatively the WSN will determine the correct course of action itself using distributed processing, operating as a sort of "hive mind" [65,66].

Additionally, security is going to become much more important in WSNs, and the hardware employed for these networks must be able to support such functionality [67]. As WSNs become more ubiquitous, the collected data will become more valuable to attackers.

86.5 Summary

This chapter gave an overview of the field of WSN hardware. The development kits for sensor nodes available today cost anywhere between 20 U.S. dollars [68] and 900 U.S. dollars [69] and upward. The price of most current sensor motes used for research purposes fall somewhere between these two extremes [70–72]. The vision is for these motes to be truly disposable, with an eventual price point of 1–2 dollars each and a lifetime of at least a year, but this, of course, is totally application dependent. By this reckoning, most commercial sensor nodes will need to drop in price by 50–100 times or more before WSNs become truly ubiquitous. This might seem a massive requirement, but a number of factors need to be remembered:

- Firstly, current commercial products are aimed at being prototyping or development platforms, with costly components and features that are likely to be unnecessary for many deployments.
- As technology improves, causing prices to drop, WSN deployments are likely to become more widespread, thus allowing economies of scale to come into play at the manufacturing level, driving costs down further.

Finally, today most sensor nodes consist of a number of components that work together, but there is a move toward SoC devices that have, at the very least, the communication, processing, and sensing subsystems all on a single piece of silicon. This consolidation of component parts will also help to drive down costs.

References

1. J. T. Correll, Igloo White, *Air Force Magazine*, Vol. 87, November 2004.
2. R. Lacoss and R. Walton, Strawman, Design of a DSN (Distributed Sensors Networks) to Detect and Track Low Flying Aircraft. *Proceedings of the Distributed Sensor Nets Workshop*, Pittsburgh, PA, Department of Computer Science, Carnegie Mellon University, AD-P003791, pp. 41–52, 1978.
3. Distributed sensor networks, MIT Lincoln Laboratory, Lexington, MA, Report number ESD-TR-88-175, 1986.
4. M. Healy, T. Newe, and E. Lewis, Wireless sensor node hardware: A review, *The Seventh IEEE Conference on Sensors (IEEE SENSORS 2008)*, Lecce, Italy, October 26–29, 2008, pp. 621–624. ISBN: 978-1-4244-2581-5, ISSN: 1930-0395, 2008.
5. K. Bult, A. Burstein, D. Chang, M. Dong, M. Fielding, E. Kruglick, J. Ho et al., Low power systems for wireless microsensors, *Proceedings of the International Symposium on Low Power Electronics and Design*, Monterey, CA, 1996, pp. 17–21.
6. B. Warneke, M. Scott, and B. Leibowitz, Ultra-low energy circuits for distributed sensor networks (smart dust), *The EECS Research Summary for 2003*, Berkeley, CA, University of California, 2003.
7. M. Srivastava, Wireless sensor and actuator networks, *Tutorial at 11th Annual International Conference on Mobile Computing and Networking (MobiCom '05)*, Cologne, Germany, 2005.
8. J. M. Kahn, R. H. Katz, and K. S. J. Pister, Mobile networking for smart dust, *Presented at ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom 99)*, Seattle, WA, August 17–19, 1999.
9. J. L. Hill and D. E. Culler, MICA: A wireless platform for deeply embedded networks, *IEEE Micro*, 22, 12–24, 2002.
10. Products, Crossbow Technology, Inc., http://www.xbow.com (accessed on May 1, 2012).

11. Texas Instruments, MSP430F15x, MSP430F16x, MSP430F161x mixed signal microcontroller (Rev. G), http://www.ti.com/lit/ds/symlink/msp430f1611.pdf (accessed on May 1, 2012).
12. Atmel Corporation, 8-bit Atmel microcontroller with 128 Kbytes in-system programmable flash, ATmega128/ATmega128L Datasheet, http://www.atmel.com/Images/doc2467.pdf (accessed on June 5, 2012).
13. Free Software Foundation, Inc., GCC, the GNU compiler collection, http://gcc.gnu.org (accessed on June 5, 2012).
14. Geeknet, Inc., The GCC toolchain for the Texas Instruments MSP430 MCUs, http://mspgcc.sourceforge.net/ (accessed on June 5, 2012).
15. AVR Freaks, AVR GNU toolchain, www.avrfreaks.net/AVRGCC/ (accessed on June 5, 2012).
16. Shockfish SA, Tinynode 584, http://www.tinynode.com/?q=product/tinynode584/tn-584–868 (accessed on June 5, 2012).
17. Freie Universtät Berlin, ScatterWeb: Modular sensor board 430, http://cst.mi.fu-berlin.de/projects/ScatterWeb/mod_MSB-430.html (accessed on June 5, 2012).
18. C. Park, P. H. Chou, Y. Bai, R. Matthews, and A. Hibbs, An ultra-wearable, wireless, low power ECG monitoring system, *Proceedings of the IEEE Biomedical Circuits and Systems Conference (BIOCAS)*, London, U.K., 2006.
19. A. Burns, B. R. Greene, M. J. McGrath, T. J. O'Shea, B. Kuris, S. M. Ayer, F. Stroiescu, and V. Cionca, SHIMMER—A wireless sensor platform for non-invasive biomedical research, *IEEE Sensors Journal*, 10, 1527–1534, 2010.
20. A. Rowe, R. Mangharam, and R. Rajkumar, FireFly: A time synchronized real-time sensor networking platform, in S.-L. Wu and Y.-C. Tseng (Eds.), *Wireless Ad Hoc Networking: Personal-Area, Local-Area, and the Sensory-Area Networks*, Boca Raton, FL: CRC Press, 2006.
21. Memsic Corporation, Iris datasheet, http://www.memsic.com/support/documentation/wireless-sensor-networks/category/7-datasheets.html?download=135%3Airis (accessed on June 5, 2012).
22. J. Eliasson, Low-power design methodologies for embedded internet systems, PhD thesis, EISLAB, Department of Computer Science and Electrical Engineering, Luleå University of Technology, Luleå, Sweden, 2008.
23. P. Dutta, J. Taneja, J. Jeong, X. Jiang, and D. Culler, A building block approach to sensornet systems, in *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems (SenSys '08)*, Raleigh, NC, pp. 267–280, 2008.
24. Centre for Development of Advanced Computing, Ubiquitous computing, http://www.cdacb.in/html/products_details.aspx (accessed on June 5, 2012).
25. Texas Instruments, MSP430 wireless development tool: EZ430-RF2500, http://www.ti.com/tool/EZ430-RF2500 (accessed on June 5, 2012).
26. Zolertia, Z1 WSN module, http://www.zolertia.com/ti (accessed on June 5, 2012).
27. Realtime Technologies Ltd., Shimmer wireless sensor unit/platform, http://www.shimmer-research.com/p/products/sensor-units-and-modules/shimmer-wireless-sensor-unitplatform (accessed on June 5, 2012).
28. SOWNet Technologies, Introducing the G-Node, http://www.sownet.nl/index.php/products/gnode (accessed on June 5, 2012).
29. Libelium Comunicaciones Distribuidas S.L., Waspmote—Wireless sensor networks 802.15.4 Zigbee mote, http://www.libelium.com/products/waspmote (accessed on June 5, 2012).
30. M. Rutishauser, V. V. Petkov, J. Boice, K. Obraczka, P. Mantey, T. Williams, and C. Wilmers, CARNIVORE: A disruption-tolerant system for studying wildlife, *Proceedings of the 19th International Conference on Computer Communications and Networks (ICCCN '10)*, Zurich, Switzerland, 2010.
31. Wireless Sensor Networks Research Group, Adding a MicroSD flash module to a SquidBee, http://sensor-networks.org/index.php?page=0827727742 (accessed on June 5, 2012).
32. R. Meike, New Sun SPOT hardware open source, http://blogs.oracle.com/roger/entry/new_sun_spot_hardware_open (accessed on June 5, 2012).

33. University of Washington—Ubicomp Research, Sensor boards, http://ubi.cs.washington.edu/wiki/index.php/Sensor_Boards (accessed on June 5, 2012).
34. L. Selavo, A. Wood, Q. Cao, T. Sookoor, H. Liu, A. Srinivasan, Y. Wu et al., LUSTER: Wireless sensor network for environmental research, *Presented at the 5th International ACM Conference on Embedded Networked Sensor System (SenSys'07)*, Sydney, Australia, 2007.
35. S. Roundy, D. Steingart, L. Frechette, P. Wright, and J. Rabaey, Power sources for wireless sensor networks, *Proceedings of First European Workshop on Wireless Sensor Networks (EWSN'04)*, Berlin, Germany, 2004.
36. IEEE, *802.15.4: Wireless Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs)*, New York: IEEE Standards Association, 2003.
37. The Bluetooth SIG Membership Website, Bluetooth Special Interest Group, http://www.bluetooth.org (accessed on June 5, 2012).
38. Texas Instruments, CC2500, http://www.ti.com/product/cc2500 (accessed on June 5, 2012).
39. K. Lorincz, B. Chen, G. W. Challen, A. R. Chowdhury, S. Patel, P. Bonato, and M. Welsh, Mercury: A wearable sensor network platform for high-fidelity motion analysis, *Presented at the 7th ACM Conference on Embedded Networked Sensor Systems (SenSys'09)*, Berkeley, CA, 2009.
40. Metageek LLC, Wi-Spy, http://www.metageek.net/products/wi-spy/ (accessed on June 5, 2012).
41. S. Roundy, P. K. Wright, and J.M. Rabaey, *Energy Scavenging for Wireless Sensor Networks: With Special Focus on Vibrations*, 1st edn., New York: Springer, 2003.
42. Moteiv Corporation, Tmote Sky, http://www.moteiv.com/products/tmotesky.php (accessed on June 5, 2012).
43. Shimmer Research, Shimmer sensors and modules, http://www.shimmer-research.com/p/sensor-and-modules (accessed on June 5, 2012).
44. Libelium Comunicaciones Distribuidas S.L., Waspmote sensor boards, http://www.libelium.com/products/waspmote/sensors (accessed on June 5, 2012).
45. Products, Topics, Intel Corporation, http://www.intel.com/technology/itj/2009/v13i3/ITJ9.3.7-Research-Platform.htm?wapkw=shimmer (accessed on June 5, 2012).
46. Shimmer Research, Shimmer revisions, http://www.shimmer-research.com/products-2/shimmer-revisions (accessed on June 11, 2012).
47. Shimmer Research, Shimmer Research increase battery capacity by 60% on the Shimmer low power wireless sensing platform, http://www.shimmer-research.com/blog/shimmer-research-increase-battery-capacity-by-60-on-the-shimmer-low-power-wireless-sensing-platform/1805 (accessed on June 11, 2012).
48. M. D. Seeman, S. R. Sanders, and J. M. Rabaey, An ultra-low-power power management IC for energy-scavenged Wireless Sensor Nodes, *Proceedings of the IEEE Power Electronics Specialists Conference (PESC '08)*, Rhodes, Greece, pp. 925–931, 2008.
49. T. Krupenkin and J. A. Taylor, Reverse electrowetting as a new approach to high-power energy harvesting, *Nature Communications*, 2, 23, 2011.
50. K. Bourzac, Power-scavenging batteries, in MIT Technology Review, http://www.technologyreview.com/energy/38434/?mod=chfeatured (accessed on June 11, 2012).
51. Toumaz Healthcare, Toumaz Microsystems, Toumaz UK Ltd., http://www.toumaz.com (accessed on June 11, 2012).
52. Products, Zepher Technology, http://www.zephyr-technology.com/ (accessed on June 11, 2012).
53. National Instruments Corporation, NI wireless sensor networks, http://www.ni.com/wsn/ (accessed on June 11, 2012).
54. Monnit Corporation, Monnit. The leader in low cost wireless sensors, http://www.monnit.com (accessed on June 11, 2012).
55. Green Goose Incorporated, GreenGoose—Play real live, http://www.greengoose.com (accessed on June 11, 2012).
56. J. Gosling, B. Joy, G. Steele, and G. Bracha, *The Java Language Specification*, Santa Clara, CA: Addison-Wesley, 2005.

57. B. W. Kernighan and D. M. Ritchie, *The C Programming Language*, Upper Saddle River, NJ: Prentice Hall, 1988.
58. Sun Microsystems, Inc., Project Sun SPOT: Sun small programmable object technology, http://sunspotworld.com/ (accessed on June 11, 2012).
59. Reuters, World's smallest java-based computers from Sentilla to Debut at JavaOne, http://www.reuters.com/article/2008/05/01/idUS102368+01-May-2008+BW20080501 (accessed on June 11, 2012).
60. Sentilla Corporation, Data center analytics software—Sentilla, http://www.sentilla.com (accessed on June 11, 2012).
61. Memsic Corporation, Iris OEM edition datasheet, http://www.memsic.com/support/documentation/wireless-sensor-networks/category/7-datasheets.html?download=136%3Airis-oem-edition (accessed on June 11, 2012).
62. Dust Networks Incorporated, Motes-on-chip product documentation, http://www.dustnetworks.com/products/list/?field_product_type_value_many_to_one=Mote-on-Chip (accessed on June 11, 2012).
63. Texas Instruments, CC2531: System-on-Chip solution for IEEE 802.15.4 and ZigBee applications, http://www.ti.com/product/cc2531 (accessed on June 11, 2012).
64. Atmel Corporation, RZRAVEN, http://www.atmel.com/tools/RZRAVEN.aspx (accessed on June 11, 2012).
65. K. S. J. Pister, From smart dust to smart plants: The evolution of wireless sensor networking, *Keynote Speaker International Society of Automation (ISA) Exposition*, Houston, TX, 2008.
66. M. Welsh, The next decade of sensor networking, *Keynote Speaker European Conference on Wireless Sensor Networks (EWSN '10)*, Coimbra, Portugal, 2010.
67. G. Press and C. Kane, Smart energy, *ON Magazine*, Issue 1, 2010.
68. Texas Instruments Incorporated TI eStore MSP430 wireless development tool target board, http://www.ti-estore.com/Merchant2/merchant.mvc?Screen=PROD&Product_Code=EZ430-RF2500T (accessed on June 11, 2012).
69. Imote2 development kit from MEMSIC.CH, http://www.memsic.cn/index.php?option=com_phocadownload&view=category&download=168%3Aimote2.builder-q-a&id=7%3Adatasheets&Itemid=86&lang=zh (accessed on June 11, 2012).
70. Shimmer Research, Shimmer, http://www.shimmer-research.com/products_detail.php?cmd=detail&id=SH-SHIM-KIT-001 (accessed on June 11, 2012).
71. TelosB based mote, http://www.advanticsys.com/shop/asxm1000-p-24.html (accessed on June 11, 2012).
72. Eistec AB, Mulle Pricing and availability, http://www.eistec.se/pricing.php (accessed on June 11, 2012).

Partial List of Vendors and Suppliers

Decton, 1 Research Forest Dr., The Woodlands, TX 77381, Phone: 713-559-9333, Fax: 281-292-2860, www.detcon.com

EMRA Controls, Inc., 1502 Pike St. N.W., Suite 4B, Auburn, WA 98001, Phone: 253-351-0292, 800-351-0293 (toll free), Fax: 253-804-2678, www.emracontrols.com

Neal Systems, Inc., 122 Terry Dr., Newtown, PA 18940, Phone: 215-968-7577, Fax: 215-968-6480, www.nealsystems.com/newsite/

PICS INC, 81 Lancaster Ave., Suite 202, Malvern, PA 19355, Phone: 610-644-7701, Fax: 610-644-9297, www.picscontrols.com

Schmitt Industries, Inc., 2765 N.W. Nicolai St., Portland, OR 97210, Phone: 888-425-8020 (toll free), Fax: 503-223-1258, schmittind.rtrk.com/?scid=1407361&rl_alt…

Steven Engineering, 230 Ryan Way, South San Francisco, CA 94080, Phone: 888-790-0022 (toll free), Fax: 800-314-0716, 800-314-0716 (toll free), stevenengineering.rtrk.com/?scid=1975164…

Wireless Cables, Inc., 2880 Zanker Rd., Suite 203, San Jose, CA 95134, Phone: 408-850-1884, Fax: 408-228-0687, www.aircable.net



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

87

Mobile Instrumentation with Wireless Design and Implementation

87.1 Introduction **87**-2
87.2 Wireless Hub Design Considerations **87**-4
"PC"-Based Hub Design Approach: PCVAN • Form Factor • Network Support • Processor, Memory, and I/O Considerations • Power Supply Selection
87.3 Implementing I/O for Vehicle and Mobile Assets **87**-5
Data Throughput Considerations • Analog versus Digital I/O Considerations • Data Acquisition Using Asynchronous Network Sources • Disadvantages of Using PCVAN-Based Hub Design Approach
87.4 Embedded Vehicle Area Network Hub Architectural Overview **87**-7
Using Microcontrollers • Always Watching! • Additional Design Considerations for EVAN Wireless Vehicle Networks • Peripherals • Divide and Conquer
87.5 EVAN Power Supply Design Considerations **87**-10
86.6 Physical and Electric Footprint Advantages **87**-11
Bibliography **87**-11

Frederick Fortson
Solidica, Inc.

Kenneth Johnson
Solidica, Inc.

Mobile instrumentation has been in use for decades ranging from telemetric navigation systems and aircraft instruments to space shuttles and satellite systems. But today, combined with the communication technologies and microcomputers, it is finding wider proliferation in many consumer products and industrial applications. For instance, the GPS systems, the sensors on board the mobile telephones, coupled together with application software make it possible to employ mobile instrumentation in many areas such as remote health monitoring, personal safety, and many other areas. In this chapter, as an example, the principles and the technology behind the mobile instrumentation of vehicular systems will be explained. Fully automated vehicles are being developed for applications in mining, industrial and personal transportation systems, underground and underwater explorations, and in military. In here, although one example is given in detail, the general principles will apply to many other and diverse ranges of applications. Some examples are also provided in chapters on radiolocation and radio navigation and in the chapter on wireless instruments.

87.1 Introduction

Wireless instruments consist of five main components—sensors and signal conditioners, programmable digital hardware, memory and storage, input/output and communication components, and others, for example, interface with computers, displays, keypads, and power supplies. A typical wireless instrument is illustrated in Figure 87.1. Instruments differ from each other by the way that they transmit, handle, and display information.

Recent developments in technology related to wireless instrumentation and sensor networking have significantly impacted the cost, capabilities, and efficacy of using wireless networks to implement mobile data acquisition, measurement, and control systems. Design engineers architecting wireless sensor and control systems are taking advantage of advances in low power operation, new chip-level system on a chip (SoC) device, and increasingly sophisticated wireless modules that benefit from higher levels of integration. Applications for mobile and remote platforms may now incorporate battery-powered sensor nodes to augment data collection in systems for which mains power may be unavailable and wiring for communications networks is unacceptable.

Previous limitations placed on battery-powered sensor networks involving wireless communications range, power consumption, network topologies, and physical footprint available for mounting have been effectively addressed using newly available devices and protocols. Advances in the state of the art are also mirrored in the embedded processing space, with new microcontroller and microprocessor offerings that also consume less power while offering greater processing speed and improved on-chip peripherals including memory, I/O, and mixed signal capabilities.

Leading practitioners in this field have successfully leveraged recent advances in wireless sensors and enabling technologies to develop novel Vehicle Area Network (VAN) applications that augment and replace existing wired sensor and control networks for mobile asset monitoring in the commercial and military space. These applications take advantage of wireless sensor networks to provide functionality that would not otherwise be possible or practical using traditional wired networks or directly wired sensors. Such applications include distributed tire pressure monitoring system (TPMS), planetary gear temperature sensors for armored vehicles in military applications, aerial platform sensors for "bucket trucks" used in utility service applications, remote digital level sensors to measure and report precise angles of operation for deployment of boom-mounted augers used on "digger derrick" trucks, and precise measurement of drill head position to calculate boring hole depth in mining applications. An example of a commercially available wireless vehicle sensor node that is able to run on battery power for up to 10 years and communicate vibration, temperature, and up to four additional analog inputs via 802.15.4 wireless is shown in Figure 87.2.

Central to mobile asset sensor network efficiency is a sensor or telematics hub, which serves as a bridge to connect wired and wireless data streams and to provide a "central" location for data storage and processing. The wireless sensor hub may also serve as a waypoint prior to moving data off platform to an enterprise application or server. The transport of sensor data off the mobile asset may be performed wirelessly, utilize direct wired connection or what is euphemistically referred to as "sneaker net": using

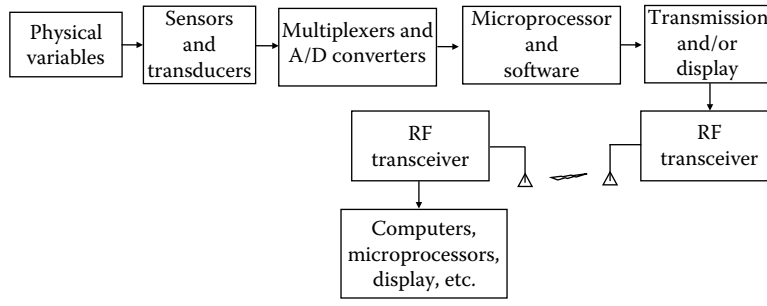


FIGURE 87.1 Components of a wireless instrument.



FIGURE 87.2 Ruggedized wireless wheel vibration and temperature sensor. (Courtesy of Solidica, Inc., Ann Arbor, MI.)

a USB stick or flash memory card as transport mechanism. Telematics and sensor hubs can typically be implemented as either an embedded device in which all processing and I/O are self-contained or by utilizing a ruggedized laptop or other computing resource running a high-level operating systems such as Linux or Microsoft Windows. Both implementations have advantages and disadvantages including cost, size, power consumption, and systems administration overhead. Additionally, the requirements (and constraints) for on-platform sensor networks are typically vastly different than for enterprise networks; a wireless sensor or telematics hub must be able to accommodate both. An overview of a complete system-wide implementation is shown in Figure 87.3.

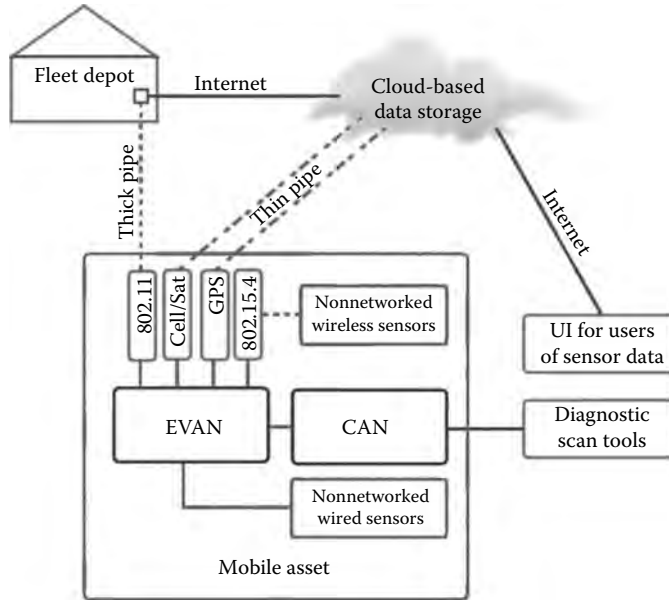


FIGURE 87.3 System-wide EVAN-based architecture.

This chapter provides an overview of the design methodologies for both the wireless sensor hub and the wireless sensor network itself. Advantages and disadvantages of various approaches to constructing wireless sensor networks using dedicated, embedded devices as well as those utilizing generic computing platforms will be discussed. Several real-world examples of VAN implementations will be considered with special attention given to what may be accomplished in commercial and military applications that consist of a multimode telematics hub and a connected sensor network. Considerations and competing requirements such as data security versus computing resources will be examined within the context of establishing wireless sensor networks for various applications and their enterprise requirements.

87.2 Wireless Hub Design Considerations

87.2.1 "PC"-Based Hub Design Approach: PCVAN

Common designs for wireless hubs fall into two broad categories: embedded designs based on microcontrollers or small microprocessors and "PC"-based designs that leverage the x86 architecture and widely available operating systems and expansion hardware. We will refer to this approach as PCVAN.

Designing a PCVAN wireless sensor network hub around an architecture originally used to implement desktop personal computers and industrial control systems has the advantage of being able to incorporate a diverse set of hardware, software, and operating systems available from a wide range of suppliers and OEMs. Development tools are also particularly easy to come by: the same compilers, network protocol stacks, debugging tools, etc., as used for designing PC applications may be leveraged to develop a wireless sensor hub for almost any application.

The development cycle itself for a PCVAN solution may be also significantly shortened using this approach, as applications may be developed, modeled, and test units fielded with readily available, inexpensive hardware designed for operation in a desktop or lab environment, and later refined to use more specialized (and expensive) hardware able to withstand extremes of temperature, shock, vibration, and environmental hazards such as saltwater intrusion.

87.2.2 Form Factor

Choices of platform form factor also exist that must be considered during the design process. Single-board computers (SBCs) that natively run Microsoft Windows, Linux, and real-time operating systems (RTOS) are available in small form factors, rated for industrial temperature and vibration and housed in NEMA-rated enclosures. Although business, residential, and light industrial applications may not require a "hardened" approach to design, and an inexpensive laptop computer is often all that is required, the harsh environment present in mobile and remote asset operation requires that all aspects of the hub be able to withstand daily abuse. Ruggedized laptops, such as the Panasonic Toughbook™, have been successfully employed in commercial, military, fire, police, and other similar telematics applications, although they come with a high initial capital costs.

87.2.3 Network Support

The use of "PC"-related hardware and software also has advantages insofar as wireless network support is concerned. Support for 802.11x (Wi-Fi) as well as Bluetooth wireless networks is commonly provided directly on the system motherboard, with additional sensor network hardware such as 802.15.4 easily added through the use of third-party adapters connected through a USB port or installed internally on an appropriate I/O expansion port.

Any connection to wired networks that may be present on a vehicle or other mobile or marine asset will typically require a protocol converter interface as direct support for mobile and industrial networks such as controller area network (CAN) may not be natively supported on the processor hardware.

For example, support for the J1939 CAN network for large vehicles and trucks might be provided through the use of an internal expansion card, a CF or SD form factor expansion card, or an external protocol converter "box" connected via an RS-232 or USB serial cable. It is important to note that software device drivers available from the manufacturer of the network interface may not be available for a given operating system or may require that additional software protocol stacks be purchased or written in order to support a given vehicle network.

For example, although a generic CAN interface card may be supported by the host hardware and operating system environment, additional drivers are required to support the specific CAN network present on a vehicle or remote asset. SAE-J1979, SAE-J1939, DeviceNet, and CANopen each require a unique protocol stack and may communicate over different network speeds to talk to a supported device on the network, even though the physical CAN network interface hardware is identical for each.

87.2.4 Processor, Memory, and I/O Considerations

The choice of processors for use of an "x86"-based PCVAN architecture typically centers on low cost, lower power microprocessors such as the relatively new Intel Atom family, as well as the well-established advanced micro devices' Geode processor family. Both support Microsoft Windows and Linux operating systems and are available in pre-integrated SBC's and motherboards in a variety of form factors. Memory and mass storage may be sized appropriate to the application and directly connected to the SBC or motherboard through OEM standard interfaces such as SATA and therefore not require special adapter cards for installation.

87.2.5 Power Supply Selection

Most "x86" motherboards, laptops, and SBCs utilize power supply bricks that expect 19 V DC for operation. For systems able to operate from mains power, an external AC adapter provides operating current to the unit. For mobile and portable applications, however, 12 or 24 V DC is more common and will require an appropriately sized DC to DC power converter.

87.3 Implementing I/O for Vehicle and Mobile Assets

Although analog and digital signal I/O is typically not part of the default architecture for SBC's and motherboards, adding connectivity features for PCVAN-based systems is supported by an entire industry dedicated to this task. Connecting thermocouples for measuring temperature, for example, is easily accomplished in much the same was as adding wireless sensors: Attach the needed I/O using externally available serial or network ports or use internal I/O expansion connectors supplied by the OEM to achieve any desired complement of analog, digital, or mixed signal expansion. One important point to note is that such expansion typically defaults to the voltage range most commonly associated with laboratory gear, providing direct connection to sensor signals in a ±5 or 10 V range. Although perfectly acceptable for tabletop or laboratory installations, additional signal conditioning modules will need to be installed between an A/D converter, for example, and a nominal vehicle voltage of 13.8–28 V, depending on the platform for the installation. Note also that most vehicle sensors and actuators operate in "single-ended" mode and may not require much in the way of signal conditioning to "knock down" system voltages for reading by the sensor network hub.

For example, a simple resistor divider network can perform the task to make a 13.8 V alternator output signal readable by an analog-to-digital converter designed to accept up to 5 V. This approach, however, introduces noise on the signal (resistors = noise!) but is generally acceptable in automotive applications, where tolerances are less stringent due to the inherently noisy nature of a vehicle electric subsystem. Using an A/D converter card with buffering provided by an instrumentation amplifier will provide a far less noisy signal and increase the input impedance to several MΩ at the same time.

This may be especially important in situations where it is desirable not to "load" an electric signal present on a vehicle subsystem. Excess loading by devices that present too low input impedance may adversely affect the system under test and measurement, resulting in poor readings and possible system malfunction.

87.3.1 Data Throughput Considerations

Another area of concern for measurement of vehicle systems is related to the sample rate, accuracy, and resolution of the analog data acquisition system being used. Contrary to the "more is better" principle typically applied when measuring and testing electric subsystems, the noise floor present in automotive systems renders as irrelevant extremely fast, high-accuracy data conversion. For example, using a 16 bit A/D converter to measure a source that itself may only be accurate to 10 or 12 bits is a waste of effort and may indeed only increase the perceived noise in the measurement itself, making the task of post-processing data more difficult. Likewise, sample rates must be specified in a manner consistent with the signal under measurement: there is no point in measuring data from a thermocouple attached to a hydraulic subsystem at 100 kHz if the fluid temperature varies slowly, as is usually the case. Samples for this kind of measurement need only be taken at once per minute in most cases. Reducing the data sample rate to values that are reasonable for the system under test generates data sets that are smaller and more manageable and reflect a better understanding of the problem to be solved.

87.3.2 Analog versus Digital I/O Considerations

Sometimes it is required to detect state change in a vehicle electric system, for example, by reading the voltage present on a vehicle switch, dial, or other actuator to determine if a subsystem is energized or otherwise activated. It is common practice to attach an A/D input to a vehicle cab dome light to infer if a driver or passenger side door is open or closed. This type of measurement presents two difficulties: The first difficulty is related to "using up" an A/D input on what is essentially a "digital" (on or off) measurement, and the second difficulty is dealing with this type of "open-switched" signal, which varies from an electric "open" circuit when not used, to a grounded, "logic-zero" state for low-side drive systems or a "logic-one" state at 13.8 V, for systems that switch system electric power to the device under measurement. Thus, inputs being read by an A/D system may only have a defined state when energized and be left floating in an ambiguous logic state when not in use.

Both issues are readily addressed by using a dedicated "digital" input (if available on an I/O adapter) instead of an analog input and by implementing a pull-up or pull-down resistor circuit on the signal line under test. In the previous example of detecting a "door-open" condition, an inexpensive 4.7 kΩ resistor may be connected between the signal line and the ground on the "high side" of the cab dome light: Logic 0 (ground) will be read when the circuit is un-switched an open, and logic 1 (13.8 V) will be present when the door is open and the dome light is switched on.

Of course, the use of any external signal I/O would only be required for signals that are not represented on a vehicle network bus as described earlier in this chapter. The vehicle and the engine OEM have already taken great pains to place their most important state and operational data on the vehicle network, in effect turning the vehicle into its own piece of dedicated test equipment. The manufacturer will also have previously determined and implemented appropriate data sampling rates, accuracy, and resolution for each sensor and subsystem on the vehicle. Taking advantage of this resource reduces overall system complexity and cost while decreasing the development effort and time required for implementation.

87.3.3 Data Acquisition Using Asynchronous Network Sources

Another important difference between wireless vehicular test, measurement, and control applications from more traditional laboratory and industrial applications is the concept of asynchronous data

present on the vehicle network and the more commonly employed synchronous sampling methodologies employed elsewhere. For example, many engineers familiar with traditional analog data acquisition equipment will be familiar with the concept of "aggregate data sample rate" specifications for test equipment. A ten-port analog data acquisition module with an aggregate sample rate capability of 100 kHz would be able to be programmed to concurrently sample data from all ten ports at 10 kHz each, or 5 inputs at 20 kHz, or one input at the full device bandwidth of 100 kHz. This common paradigm of dividing total sampling capacity by the number of inputs may leave an engineer that is not familiar with the asynchronous nature of vehicle network traffic confused by the very nature of how data is presented on a network: Some items may be reported with regularity, such as a temperature being reported once per second. Other items, such as speed, O_2 sensor, throttle position, and fault codes, may be present at varying rates on the network, and values will have to be interpolated or simply thrown away if the engineer tries to fit asynchronous data collection and reporting into methodologies that are designed around synchronous data collection and analysis.

Another area of concern occurs when "request/response" techniques are misused to increase data reporting rates on a vehicle network. If a vehicle network is asked to report vehicle speed many times greater than would otherwise be typical to decrease "granularity" in speed data reporting, other systems might not have network bandwidth left to operate. A common problem encountered when an engineer miscalculates network utilization is such a manner as to overwhelm system messages that may render a vehicle inoperable by dint of preventing system messages from being correctly sourced and delivered: dashboard and lower-priority messages fail first, with unpredictable behavior evident on the affected subsystems. As bus utilization grows further, the vehicle may enter its default "limp-home" mode or fail to operate altogether.

87.3.4 Disadvantages of Using PCVAN-Based Hub Design Approach

Although there are considerable advantages of PCVAN-based designs for implementing a vehicle area wireless test measurement and control network, there are inherent disadvantages of this approach that include high power consumption, physical footprint, large memory footprint, large total code size required for running applications, difficulty ruggedizing and enclosing such systems, large overhead in systems administration, time-stamp "skew," poor real-time performance for data event processing, lack of adaptability/scalability, and expense. Additionally, although using a PC-based architecture is often associated with a large available pool of software developers already familiar with the platform, the developers often struggle with unique challenges posed by automotive and remote application development that is more typical of the challenges faced by those with experience programming for embedded systems.

87.4 Embedded Vehicle Area Network Hub Architectural Overview

A rapidly emerging approach to remote mobile sensor applications that addresses many of the challenges embodied in the PCVAN approach is the Embedded Vehicle Area Network (EVAN) design philosophy. These new breed of low power devices offer a stand-alone, integrated device that incorporates multiple-wired and wireless sensor networks including 802.11b/g, 802.15.4, 2G and 3G cellular, SAE J1979 and SAE J1939 Vehicle CAN, SAE J1708 Vehicle Diagnostic Network, USB, and Ethernet. Also integrated into these new platforms are a variety of analog and digital data acquisition and output capabilities, GPS, real-time clock with battery backup, wide-input voltage-range vehicle power supplies, and water-proof, rugged enclosure with automotive grade connectors. Additionally, module high-value-added sensor subsystems such as TPMS desired for retrofit applications can be provided internally through an expansion connector that mates with the TPMS OEM adapter card. An example of a commercially available EVAN is shown in Figure 87.4.



FIGURE 87.4 Commercially available EVAN device with J1979, J1708, J1939, GPS, cellular, 802.15.4, and 802.11b/g capability. (Courtesy of Solidica, Inc., Ann Arbor, MI.)

87.4.1 Using Microcontrollers

In an EVAN, a dedicated 32-bit microcontroller provides the computational bandwidth required to process and transport multiple data streams to and from various networks simultaneously. The use of an integrated microcontroller offers significant advantages over using a microprocessor as described in the previous examples. Most modern microcontrollers, such as the Freescale ColdFire and ARM core variants, offer deterministic interrupt latencies for true real-time interrupt handling, thereby permitting highly efficient data acquisition, manipulation, and storage algorithms that are not possible when using an "x86"-style processor that was designed originally for desktop computing requirements. Modern microcontrollers also include significant complements of on-chip memory, including single-cycle access RAM, FLASH, and EEPROM (or equivalents) that permit faster interrupt handling and highly efficient data manipulation and buffering on a single die. This also permits the system to run at higher "MIPS per watt" ratios, reducing system power consumption and heat generation. A typical RISC-based microcontroller, using only on-chip memory and peripherals, may consume several orders of magnitude less power to perform essentially the same task set as a larger and faster microprocessor-based approach that would otherwise require significant off-chip memory and I/O resources.

87.4.2 Always Watching!

Significantly, this design approach permits an embedded wireless hub to operate in an "always-on" mode, where the system may operate even when the vehicle is turned off or the remote asset has limited power generation capabilities. Typical configurations will involve having the unit "sleep" when a vehicle engine is off and waking up periodically to check system status and report significant events that would otherwise go unnoticed on PCVAN devices that can operate only when the engine is running. PCVAN solutions that feature low power shutdown or sleep modes are also constrained by what kind of events may be used to "wake up" the device. A hardwired ignition input may be required, as opposed to being able to wake up on a vehicle network or wireless network message. Additionally, current draw for a PCVAN device is too prohibitive for frequent "wake-up" cycles, further limiting their effectiveness. Most newly emerging EVAN systems consume less power when operating than PCVAN approaches consume while "sleeping."

Vehicle electric systems are notoriously difficult to diagnose when there are intermittent failures; the value of an always-on EVAN system is magnified when it is also used to continuously monitor the vehicle during engine starting and shutdown, which places the most demand on the vehicle battery, alternator, and associated electronics. An EVAN system will always able to track electric system behavior,

whereas a PCVAN approach will miss the most taxing part of vehicle operation while it is booting an operating system or recovering from standby.

Data, configuration information, and run-time software may be buffered locally on SD card–based flash memory cards to provide features such as store and forward of vehicle network and sensor data for transmission off platform via Internet and corporate intranet connections. Platform firmware upgrades may also be remotely uploaded in situ while the unit is in service. Another advantage of the code efficiency and use of "lightweight"-embedded RTOS use in the EVAN approach is that large off-line storage and rotating physical media are not required. Booting is usually direct from on-chip flash memory, and system size, complexity, expense, and power are thereby significantly reduced.

The power and flexibility inherent in the EVAN architecture allows the platform to be used in a wide variety of fleet telematics and remotes sensing applications and use cases, from straightforward asset monitoring and remote diagnostics of long-haul trucking fleets to highly specialized applications such as electric utility areal platform work trucks that require wireless monitoring of hydraulic fluid pressures and temperatures in addition to vehicle network sensor awareness.

87.4.3 Additional Design Considerations for EVAN Wireless Vehicle Networks

When designing a wireless sensor network using an EVAN approach for use in a mobile asset, it is often important to take advantage of any features that may be present in an existing network that is "organic" to the platform, such as an existing wired vehicle network. Modern commercial, consumer, and military vehicles incorporate one or more vehicle networks to establish sensor and control communication between electronic body, engine, and transmission modules in place of the direct point to point wiring that dominated the electric systems of legacy vehicles. The use of such networks serves to save space, eliminate costly and failure prone wiring harnesses, improve vehicle safety, and provide operating features that would not be otherwise possible. Servicing a modern vehicle will typically involve the connection of a "scan tool" to one or more on-board diagnostic (OBD) ports to read and reset trouble codes, reprogram or update the firmware of an attached control module, directly query an attached sensor, or read system voltage and current values of interest.

Early designs (OBD 1) would incorporate the majority of functionality in the engine control unit (ECU), where test equipment would be connected to an access point located in the engine compartment for service purposes. Modern vehicles have a more distributed approach to data processing and control, wherein functionality is distributed throughout the vehicle among intelligent nodes connected over a vehicle network such as the SAE-J1939 and SAE-J1979 CAN. This approach lends itself to the combined use of small, simple "body modules" for such functions as door lock and window control and more sophisticated modules for engine and transmission control. Multiple vehicle networks may be also employed where a secondary network is dedicated for high-speed real-time control messages between the ECU and a transmission module, for example. Such systems typically have a diagnostic and debug connector for network access in the passenger compartment of cars and light-duty trucks (J1979 OBD 2) or in the vehicle cab (J1939) for heavy truck, bus, and military vehicle applications. A simple schematic of a modern automotive network configuration is shown in Figure 87.5.

The presence of a vehicle network also presents an opportunity for designers of vehicle sensor and telematics equipment to utilize data available on the network to supplement the data to be used in the wireless network as many data items of interest may be collected without requiring the installation of additional sensors to the vehicle. For example, fuel level, battery voltage, vehicle speed, engine speed, and oil temperature and pressure are all available to be read directly on the vehicle network and may be used in conjunction with other sensor data for whatever use case is desired. Although originally designed to be used for service applications and emissions testing purposes, the use of a vehicle network for other applications such as data acquisition and control has become commonplace. Using EVAN architecture increases the efficiency and reliability of processing and managing this heterogeneous data flow.

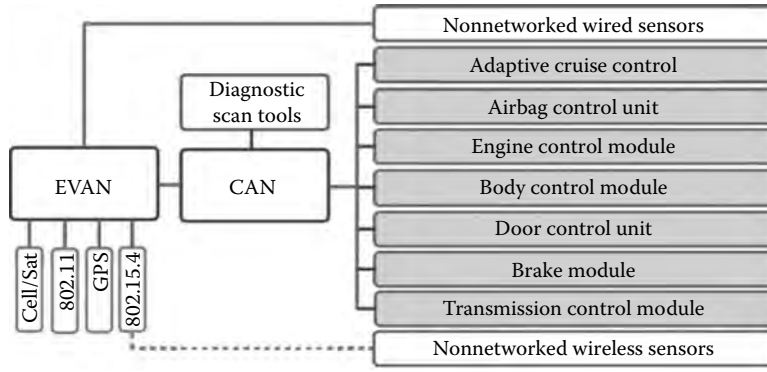


FIGURE 87.5 Modern network schematic with inclusion of EVAN hub.

87.4.4 Peripherals

The microcontroller-based EVAN devices are able to take advantage of on-chip peripherals such as Ethernet, USB, UARTs for legacy serial devices, timers, DMA controllers, and built-in analog and digital I/O that would otherwise need to be added to a PC-based VAN through the use of additional hardware peripherals otherwise not present on the microprocessor itself. This allows the EVAN to be designed using far fewer components, further reducing cost, size, and power consumption and increasing reliability.

87.4.5 Divide and Conquer

Recent advances in microprocessor integration have also revolutionized the off-chip peripheral options as well. For example, in order to implement 802.11x Wi-Fi for a PCVAN, a digital radio subsystem will typically be added to the motherboard, with the Wi-Fi protocol stack and TCP/IP stack left to be implemented on the "x86" microprocessor in addition to all of the other tasks it is expected to perform. An embedded VAN approach, however, may use a highly integrated Wi-Fi module, available from multiple OEMs, which completely encapsulates the Wi-Fi and TCP/IP stacks on its own microcontroller, thereby freeing up the host micro to concentrate on its core mission.

Such highly integrated, microcontroller-based modules are readily available to cover a wide variety of embedded applications including cellular machine to machine modems, GPS, Bluetooth, TPMS, and 802.15.4 support for ZigBee and numerous other wireless network protocols. Such modules leverage the advantages described earlier by encapsulating their functionalities and offering direct, "no glue logic" connection to a host processors serial UART or SPI peripheral complement.

To preserve the advantages in low power operation offered by the highly integrated host microcontroller, most wireless peripherals offer their own internal low power sleep and shutdown modes, enabling an even lower power budget for the system as a whole.

87.5 EVAN Power Supply Design Considerations

To further reduce current consumption, increase reliability, reduce size, and permit remote battery-powered operation in addition to using vehicle power, a large assortment of new, highly efficient power supply modules are available to leverage an EVAN design. Because the total power consumption of an EVAN device is able to be orders of magnitude less than a PCVAN approach, alternate designs may be used to provide power to the system. Wide-input voltage-range switching supply modules with greater than 90% efficiency are now available from many manufacturers. At less than a square

centimeter in area, the modern switching regulator is far smaller than the power supply brick required to operate a PC or laptop. An integrated switching supply module is typically built directly on the EVAN board, as opposed to residing a dedicated off card module. Because the total current needed to be supplied is relatively low, these devices often have an input voltage range of 6–36 VDC, which enables direct connection to 12 or 24 V vehicle electric systems with no external DC to DC converter required.

Another power supply "trick" that may be employed to further reduce current draw and permit always-on operation is the use of a small low-dropout linear regulator to supply power to the host microprocessor when in sleep mode. This permits the host to physically shut down the main switching regulator when not needed, further reducing current consumption.

For supplying power to "hungry" onboard peripherals, such as a Wi-Fi module, the use of a dedicated current switch or switching regulator may also be employed to electrically remove the device from the system when not needed. A well-designed EVAN will thereby only have peripherals in a powered on state when actually required for operation, further reducing the total system power budget. This is in stark contrast to the design choices presented for a PCVAN approach, where many (or all) of the system peripherals will be powered up at the same time.

86.6 Physical and Electric Footprint Advantages

The EVAN approach offers an increase in efficiency and reduced footprint and mounting requirements over the PCVAN approach, thereby allowing applications to exist that would otherwise be impractical to implement. For example, with an EVAN design, it is now possible to place wireless telematics devices in otherwise inaccessible locations, such as the bucket of an aerial platform rig for power line maintenance.

Because the nature of the electric power line service marketplace, which require that the bucket and support arm be free of electric wiring of any kind, a battery-powered EVAN may be placed on board the device to monitor position, vibration, and abuse and determine hours and modes of operation that would not be possible if a large, power-hungry PCVAN were to be used.

Numerous examples of EVAN installations are possible in a diverse range of enterprise applications for which use of a PCVAN is not suitable, such as battery-operated use in a refrigerated cargo container, unpowered trailer asset monitoring, pallets, remote pumping stations, weather and road condition monitoring, factory automation, shipping containers, and large fleet deployments where the expense and the maintenance overhead associated with a PC-based design approach are not workable.

Bibliography

Eren, H., *Electronic Portable Instruments-Design and Applications*, CRC Press, LLC, Boca Raton, FL, 2004.

Eren, H., *Wireless Sensors and Instrument Networks—Networks Design and Applications*, CRC Press, Boca Raton, FL, 2006.

Erl, T., *SOA Design Patterns*, Prentice Hall, Upper Saddle River, NJ, 2009.

Jamsa, J. and Luimula, M., Advanced car navigation: Future vehicle instrumentation for situation-aware services, *12th IEEE International Conference on Mobile Data Management (MDM 2011)*, Lullea, Sweden, pp. 7–10, 2011.

Johnson, K. and Fortson, F., Sense and respond, lessons learned in support of condition-based maintenance programs through data acquisition, analysis and secure transmission, *Ground Vehicle Systems Engineering and Technology Symposium Technical Paper Publication*, Vol. 2, pp. 531–538, August 2010.

Mariusz, M., Zabiegaa, B., and Namiesnic, J., Mobile systems (portable, handheld, transportable) for monitoring air pollution, *Critical Reviews in Analytical Chemistry*, 42(1), 2–15, Taylor & Francis Inc., 2012.

Pecht, M.G., *Prognostics and Health Management of Electronics*, Wiley, Hoboken, New York, 2008.

Zhong, H., Li, P., Hu, Y., Lv, Z., Yin, J., Yu, B., and Wu, J., Solution for the data collection in the field survey based on mobile and wireless GIS, *18th International Conference on Geoinformatics*, Beijing, China, IEEE, Piscataway, NJ, p. 5, 2010.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

88

Powering Autonomous Sensors

Manel Gasulla
Universitat Politècnica de Catalunya

Maria Teresa Penella
Urbiotica S.L.

Oscar Lopez-Lapeña
Universitat Politècnica de Catalunya

88.1 Autonomous Sensors....**88**-1
Primary Batteries • Energy Harvesting
88.2 Ambient Energy Sources and Transducers....**88**-3
Radiant Energy • Mechanical Energy • Thermal Energy • AC Magnetic Energy • Biochemical Energy
88.3 Energy Storage Units....**88**-6
Batteries • Supercapacitors • Hybrid Storage
References....**88**-9

Wireless sensor networks (WSN) have recently become feasible thanks to reductions in the power consumption of the sensor nodes and to the advent of specific communication standards and commercial transceivers. However, for WSN to thrive, several challenges must first be addressed, including power supply to the nodes.

Primary batteries provide a simple solution to power the sensor nodes. However, batteries have a limited energy, and their replacement can increase the maintenance costs and become unpractical or even unfeasible. Harvesting energy from the ambient is an alternative. This chapter presents both powering alternatives with a higher focus on the second one. First, autonomous sensors are described. Then, the main characteristics and differences between primary batteries and energy harvesters are presented. Finally, a focus is given to some specific parts of energy harvesters such as the ambient energy sources, the related energy transducers, and the energy storage devices.

88.1 Autonomous Sensors

Autonomous sensors are measurement devices in which cables are not used for transmitting data or for the power supply. They serve as the nodes of distributed data acquisition systems and WSNs and are used in numerous fields, including health care [1], aerospace [2], environmental monitoring [3], or Smart Cities [4]. To increase their autonomy, autonomous sensors seek to reduce their average power consumption by working in low-power modes whenever possible. They spend most of the time in *sleep* (standby) mode, and only *wake up* to perform specific actions—namely, measurement, processing, and transmission/reception of data [5].

* Most of the material of this chapter has extracted with kind permission from Springer Science+Business Media: *Powering Autonomous Sensors: An Integral Approach with Focus on Solar and RF Energy Harvesting*, 2011, Penella-López, M.T. and Gasulla-Forner, M.

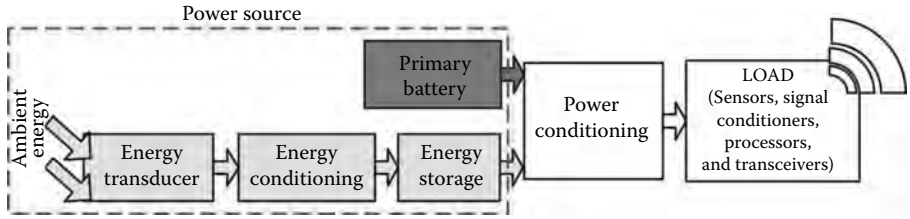


FIGURE 88.1 Block diagram of the power supply of an autonomous sensor. Either a primary battery (dark-gray box) or an energy harvester subsystem (light-gray blocks) can be used.

An autonomous sensor is usually composed of one or more sensors and their corresponding signal conditioners, a processor and a transceiver. Sensors convert a signal from a physical or chemical quantity to a corresponding signal in the electric domain. The sensors to be used depend on the desired application. A signal conditioner (or analog processor) matches a sensor output to the digital processor (usually, a low-cost microcontroller). Commercial transceivers are used for wireless communication. They transmit in the free-licensed ISM bands and can use a proprietary or standard (e.g., IEEE 802.15.4) protocol.

Power needs to be supplied to the autonomous sensor. Figure 88.1 shows a generic block diagram for the power supply. The left-hand rectangle with dashed line shows two options for the power source: either a primary battery (dark-gray box) or an energy harvester (light-gray blocks). Energy harvesting implies more complexity. An optional *power conditioning* stage matches the power source to the power supply requirements of the *load* (sensors, signal conditioners, processor, and transceiver).

88.1.1 Primary Batteries

Primary batteries provide a simple solution yet have several drawbacks, including size, limited energy storage, and poor indication of the remaining capacity. Whenever their replacement is unpractical or unfeasible, batteries with an adequate amount of energy (E_{bat}) for powering the autonomous sensor over its lifetime should be used. Thus, the required E_{bat} for a certain operation lifetime period, T_{op}, can be computed as

$$E_{bat} = \overline{P}_c \cdot T_{op} \tag{88.1}$$

whereby $\overline{P_c}$ is the average power demanded from the power source and includes the power demanded by the load plus the power losses of the different parts, such as that of the power conditioning stage or the battery itself. On the other hand, the peak-power capability of the batteries must be higher than the peak power demanded by the autonomous sensor. Small-size batteries, which are prone to be used with autonomous sensors, present rather high impedances and thus a limited peak-power capability.

88.1.2 Energy Harvesting

Energy harvesting enables autonomous sensors to run in nearly perpetual operation with little or no maintenance; however, it increases circuit complexity and is a fledgling technology that still poses certain challenges. Optical [6], mechanical [7], thermal [8] or even radio frequency (RF) [9] energies have been proposed. The energy harvester comprises a transducer, an energy conditioning stage, and an energy storage unit (Figure 88.1). The transducer converts ambient energy into electric energy with a given efficiency. Typical ambient energy sources operate with a certain degree of variability (e.g., night-to-day variability when harvesting solar energy); thus, an energy storage unit is needed to provide the demanded power flow to the load. An energy conditioning block is used to efficiently charge the storage unit from the output of the transducer. To maintain

sustainable operation of the autonomous sensor when harvesting energy, the average generated power $\left(\overline{P_g}\right)$ must comply with the following:

$$\overline{P}_g \geq \overline{P}_c \tag{88.2}$$

Here

- $\overline{P_c}$ includes the demanded power by the load plus the power losses of the energy and power conditioning blocks and of the storage unit
- $\overline{P_g}$ depends on the ambient conditions at the physical location of the node and on the efficiency of the transducer

The storage unit must supplement the power transferred to the load whenever $P_g < P_c$, whereby P_g and P_c correspond to the instantaneous generated and consumed power, respectively. Then, for any arbitrary long period of time, T, a long-term storage ($E_{storage}$) unit must fulfill the condition

$$E_{storage} \geq \max\left\{\int_T \left(P_c - P_g\right)\cdot dt\right\} \tag{88.3}$$

Both supercapacitors and secondary (rechargeable) batteries can be used as energy storage units and will be reviewed more extensively later on.

88.2 Ambient Energy Sources and Transducers

Thomas et al. [10] extensively described many energy sources for powering small-scale unmanned systems. Based on their initial classification, we have categorized harvested energy for powering autonomous sensors into five types: *radiant, mechanical, thermal, magnetic,* and *biochemical* (Figure 88.2).

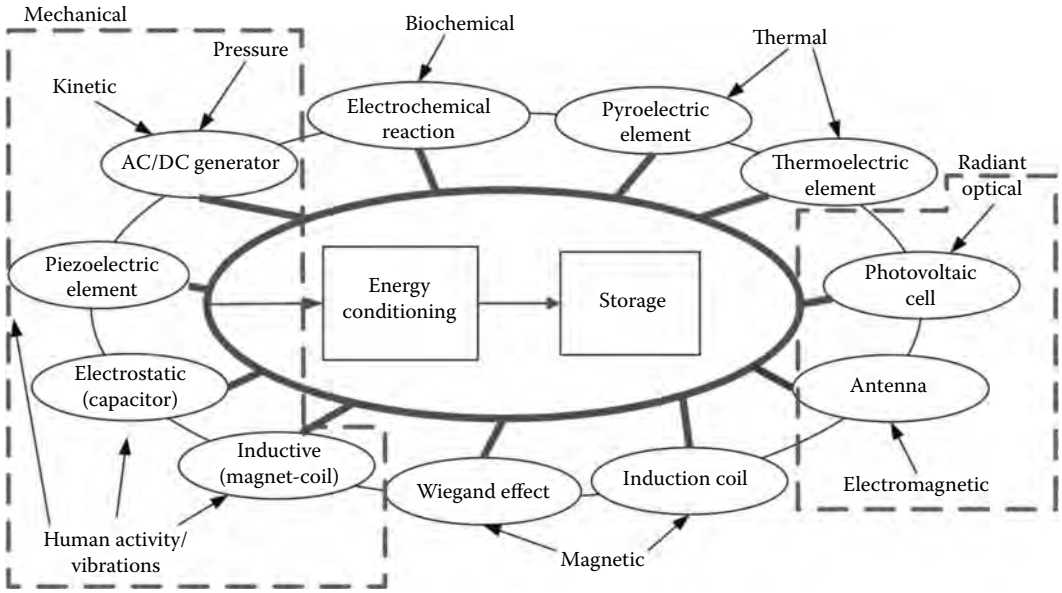


FIGURE 88.2 Energy sources and their respective transducers for powering autonomous sensors. (Adapted from Thomas, J.P. et al., *J. Power Sour.*, 159, 1494, 2006, with additional power sources added.)

Here, we consider energy already present in the environment, although in some cases energy can be expressly distributed, in a process called *remote powering*. Other analyses, more detailed descriptions, and typical power densities of energy sources can be found in the literature, for example, in [11] or [12]. The choice of the ambient energy source strongly depends on the ambient energy density at the autonomous sensor's location and on the load's power needs. On the other hand, energy transducers must be used to convert the ambient energy source into an electric output. Efficiency here is defined as the ratio of the transducer output power to its input power. In this section, we will briefly review different energy sources and associated transducers.

88.2.1 Radiant Energy

Radiant energy is emitted as electromagnetic waves, which includes radioactive, optical, and RF energy. Radioactive materials have extremely high energy densities, and their half-lives span from fractions of a second to millions of years. Radioactive properties can be found in uranium, nickel, or phosphorous isotopes among others. However, because these materials can be extremely dangerous, they must be confined to a closed volume. The energy from radioactive sources that emit alpha and beta particles can be converted by piezoelectric or betavoltaic elements, which determine their efficiency [13]. Commercial tritium batteries based on the betavoltaic effect are already available from City Labs.

Optical energy accounts to the portion of the electromagnetic spectrum spanning from infrared to ultraviolet light. Outdoors, optical energy chiefly comes from the Sun. Part of the energy radiated by the Sun is absorbed by the atmosphere; the rest arrives to the surface of the Earth as both direct and indirect radiations. On cloudy days, indirect radiation predominates. The available power density depends on several factors, including latitude, weather conditions, day of the year, and time of day. For example, at noon on a sunny day, the incident light on the surface of the Earth, in the Equator, can reach a power density of a little more of 1000 W/m^2. Indoor optical energy comes from outdoors (e.g., near windows) and from artificial illumination (e.g., incandescent and fluorescent bulbs). Optical power density, when the lights of a room are on, typically ranges from 100 to 1000 $\mu W/cm^2$. For autonomous sensors, optical energy is harvested via photovoltaic (solar) cells. A solar or photovoltaic cell is basically a semiconductor diode with a large p–n junction in the plane of the cell. A simple model includes a DC current source with a diode in parallel. Most current solar cells are silicon based and fall into one of three categories: monocrystalline, polycrystalline, and amorphous. Their efficiencies at full Sun range from less than 20% for monocrystalline cells down to 5% for commercial amorphous cells. Crystalline cells are better suited for outdoor use, whereas amorphous cells work better indoors. Organic (dye-sensitized solar cells, DSSC) and fourth-generation solar cells are becoming more popular and commercially available (e.g., Solaronix or Dyesol).

RF signals are deliberately radiated by broadcasting station antennas and cellular phone antennas. The RF energy transducer is an antenna, which can be modeled as a voltage AC source in series with output impedance. The generated signals are of low amplitude, so a rectifier and a boosting stage must be used. Except for locations near transmitters, the available ambient power density is in the nW/cm^2 range [14], which is too low to be useful. An alternative to power autonomous sensors is to deliberately transmit RF power on the industrial, scientific, and medical (ISM) frequency bands. Some commercial products are already available (Powercast).

88.2.2 Mechanical Energy

Mechanical energy harvesting [7,15,16] is based on kinetic energy, the sources of which include liquid or gas flow, vibrations, human activity, and pressure variations (e.g., acoustic noise and atmospheric pressure). Available power in flow energy increases cubically with an increase in the speed of the liquid or gas. Low-level vibrations occur in machinery, outdoor windows, and transport vehicles; they produce frequencies between 50 and 200 Hz and acceleration amplitudes between 1 and 10 m/s^2. Human activity can actively or passively generate kinetic energy. Active human power requires deliberate movement,

whereas passive human power exploits common daily activities (e.g., heel strike while walking); indeed, nearly all the energy produced in daily activities stems from when legs are bent and from heel strike.

Mechanical energy can be converted to electric energy by one of the following conversion principles: electrostatic, piezoelectric, and electromagnetic. Electrostatic converters are based on variable capacitors. One plate is fixed, and the other changes with the mechanical force, thereby changing the value of the capacitance. One attractive feature is that they are IC compatible. Reversible electric polarization in response to strain from mechanical stress is called the *piezoelectric effect*. Electromagnetic induction, based on Faraday's law, refers to generation of a voltage in a conductor (typically a coil) located within a magnetic flux, usually generated by a permanent magnet. The generated signals are AC in all three conversion principles and must be rectified in order to power autonomous sensors. The specific energy conditioning stage must be tailored to each transducer type. Based on the electromagnetic principle, some commercial products have appeared that harvest energy from water by using miniature hydroelectric generators (Splashpower), produce energy form vibrations (Perpetuum), or even get enough energy when pressing buttons or switches in order to control lights in buildings (EnOcean). Other products are based on the piezoelectric principle (e.g., Midé, Arveni).

88.2.3 Thermal Energy

Thermal energy can be converted to electric energy by thermoelectric or pyroelectric transducers, among others. The former rely on spatial variations in temperature, whereas the latter rely on temporal variations in temperature. Thermoelectric generators, which exploit the Seebeck effect, comprise several pairs of alternating p- and n-type semiconductor blocks arranged thermally in parallel and connected electrically in series. They can be modeled as a DC voltage source in series with a resistance. Thermoelectric generators are typically built from Bi_2Te_3 or poly-SiGe, and their thermal resistance should be matched to that of the thermal source in order to maximize the harvested power. Optimizing the extracted power requires additional elements—namely, a heat sink for efficient dissipation of heat into the environment and structures, known as *thermal shunts*, which direct the heat that passes between the hot and the cold plates into the thermocouple legs [17]. There are several companies commercializing thermoelectric generators (e.g., EnOcean, Micropelt, Thermolife, Tellurex).

88.2.4 AC Magnetic Energy

AC magnetic energy [10] is available near AC current–carrying conductors (e.g., power grids), magnets, and electric rotating machinery. Electromagnetic induction in a coil can be used to convert AC magnetic energy into electric energy. The transducer (coil) can be modeled as an AC voltage source with inductive series impedance. Wireless nonradiative power transfer through strongly coupled resonant objects (also known as *magnetic resonance*) has been reported in [18]. Coil-to-coil efficiencies of up to 40% over a distance of 2 m were achieved by using coils of 60 cm of diameter. A related company (WiTricity) has been raised. The concept is similar to that already used in low-frequency RFID systems in which energy is transferred via magnetic coupling from a reader in order to power nearby tags. Inductive powering has also been applied in biomedical implants [19]. The same principle is also found in some commercial products (eCoupled), and recently the Wireless Power Consortium has been created in order to set international standards for interoperable wireless charging of mobile devices.

88.2.5 Biochemical Energy

Biochemical energy can be extracted by using fuel cells. In a typical cell, fuel is continuously fed to the anode compartment, and an oxidant (e.g., oxygen from air) is continuously fed to the cathode compartment; electrochemical reactions occur at the electrodes, producing an electric current. For biochemical energy, the fuel is biological material, such as enzymes, microbes, glucose, or marine sediment.

88.3 Energy Storage Units

Both secondary (rechargeable) batteries and supercapacitors can be used as energy storage units in energy harvesters for autonomous sensors. Both of them are electrochemical devices, that is, they consist of two electrodes and an electrolyte. However, in a supercapacitor, in contrast with a battery, no electrochemical reaction takes place [20]. In this section, secondary batteries will be first described. Then, supercapacitors will be presented and compared to secondary batteries. Simple circuit models of batteries and of supercapacitors will be shown. Finally, the combined use of batteries and supercapacitors (*hybrid storage*) and its utility for autonomous sensors will also be addressed.

88.3.1 Batteries

Batteries extract electric power from a chemical reaction. They comprise one or more basic electrochemical units known as *cells*, which are connected in series or parallel, to obtain the desired voltage and capacity. Each cell contains a negative electrode (an *anode*), a positive electrode (a *cathode*), and an ionic conductor (an *electrolyte*). The anode and cathode are physically isolated; the electrolyte provides the medium for charge transfer (via ions) inside the cell, between the anode and the cathode (Figure 88.3a). In practical cells, the anode and the cathode are mechanically separated by a partition material that is permeable to the electrolyte [20]. A battery can be readily modeled with a zero-order electric model similar to that shown in Figure 88.3b, whereby V_B models the open-circuit voltage and R_B models the equivalent series resistance (ESR). The value of V_B decreases with decreasing internal stored charge. The value of R_B depends on the chemistry, battery design, and temperature.

Power density mainly depends on the internal impedance of the battery, which is tightly related to the current collector area and the electrolyte conductivity. Some batteries are specially designed to promote power density through reduced internal impedance by maximizing the current collector area. However, batteries designed to improve power density also present higher self-discharge and lower energy density. Small batteries (e.g., button type) generally exhibit high internal impedance. In addition, battery impedance increases with an increasing *depth of discharge* (DoD) and *number of charging/discharging cycles*.

The rate of chemical reactions increases with increasing temperature. This principle has beneficial consequences (e.g., increased power density because the resistance decreases) as well as detrimental ones (e.g., higher leakage). The lower limit for a battery's operating temperature range is partially determined by the electrolyte's freezing point.

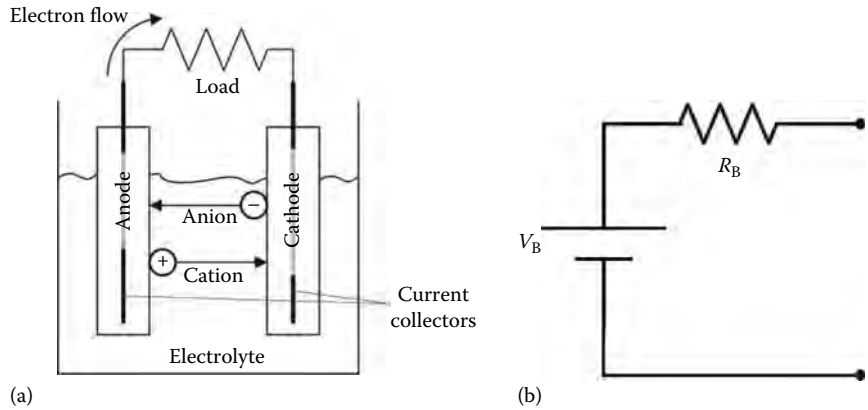


FIGURE 88.3 (a) Electrochemical cell at discharge and (b) a zero-order electric model.

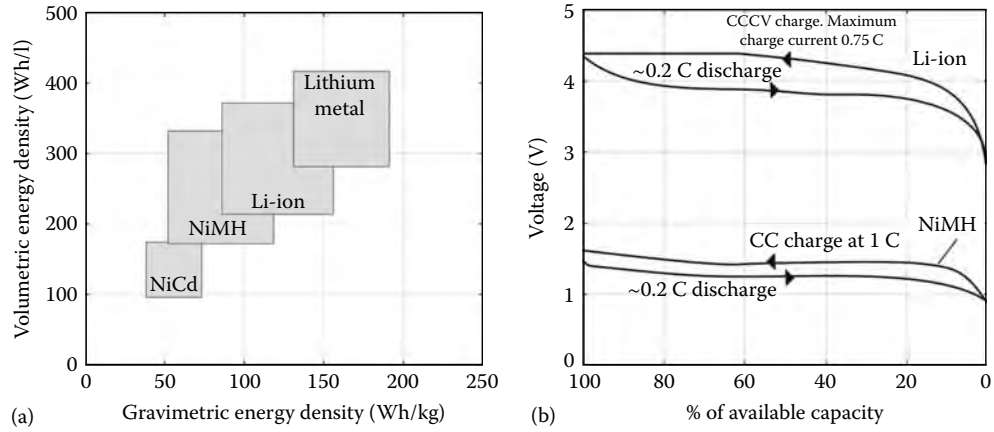


FIGURE 88.4 (a) Volumetric energy density plotted against gravimetric energy density for common secondary battery types. (Adapted from Linden, D. and Reddy, T.B., *Handbook of Batteries*, 3rd edn., McGraw-Hill, New York, 2001.) (b) Charge (at 1 C and 0.75 C) and discharge (at 0.2 C) curves for Li-ion and NiMH batteries.

The most widely used secondary batteries for portable and space-constrained devices are

- Lithium based (Li-ion and Li polymer)
- Nickel metal hydride (NiMH)

Figure 88.4a shows a plot of volumetric versus gravimetric energy density for typical secondary batteries. Metallic-lithium-based batteries offer better energy densities but present more safety risks. Batteries using lithium alloys in the negative electrode (e.g., button LiAl/MnO_2 batteries) imply less risks but at the expense of lower energy density and poor cycle life at deep DoD. These batteries are sold in coin configuration and are intended for low rate discharges. Lithium-ion batteries use lithium intercalation compounds as positive and negative materials. Over the past decade, they have become the most popular lithium-based batteries, as they are relatively safe and offer reasonably high energy densities. Lithium polymer batteries are Li-ion batteries that employ polymer-based electrolytes, which enable slimmer, and even flexible, designs. Figure 88.4b shows the charge and discharge curves for Li-ion and NiMH batteries. As observed in the plots, during charge, the battery voltage is higher than during discharge, which can be easily explained by the effect of the internal impedance. Lithium-ion batteries present higher voltage ranges than do NiMH batteries. Table 88.1 shows the principal characteristics of three of

TABLE 88.1 Principal Characteristics of the Three Main Types of Secondary Batteries

Battery Type	Li Ion	Li Polymer	NiMH
Internal impedance Z (mΩ)	30 to thousands	40 to thousands	50 to thousands
Overvoltage tolerance	Explosion risk	Damage risk	Temperature increase
Temperature range (°C)			
Charge	0–40	0–45	0–45[a]
Discharge	−20 to 60	−20 to 60	−20 to 60[a]
Self-discharge (% per month)	2	2.5	15–25
Cycle life (cycles)	500–1000	300–500	300–500
Float or calendar life (years)	>5	>5	4–6

Sources: Linden, D. and Reddy, T.B., *Handbook of Batteries*, 3rd edn., McGraw-Hill, New York, 2001; Woodbank Communications, Electropaedia, Battery and Energy Technologies, Available on: http://www.mpoweruk.com/ (accessed on 2nd August, 2012).

Note: Parameters are specified at 25 °C, otherwise specified.

[a] These are the permissible temperature ranges; however, the cell should operate as closely to room temperature as possible.

the most common types of secondary batteries. As can be seen, NiMH batteries present a higher self-discharge. Furthermore, the number of cycles is higher for Li-ion batteries. The cycle life is defined as the number of cycles at which the battery capacity decreases to 80% of its original capacity. The number of cycles typically increases with a decreasing DoD.

When harvesting energy from the environment, the charging current is usually quite low, thereby minimizing the need for complex charging and protection circuits. Simple over or under voltage protection circuits (e.g., a comparator plus a disconnection switch) can be considered.

Thin-film batteries are solid-state lithium batteries specially designed for integration onto silicon wafers, as they are very thin (micrometers). Both Li-ion and metallic-lithium thin-film batteries have been developed. They can sustain high temperatures (250 °C) for short periods of time and contain no liquid elements; therefore, they can support solder reflow [21]. Cycle lives of several thousands of cycles at room temperature and hundreds of cycles at 150 °C have been demonstrated [23]. These batteries exhibit energy densities of 100–1000 μWh/cm^2, depending on the cathode [24], and power densities (under continuous discharge) of up to 10 mW/cm^2 [23]. Therefore, their internal impedance is quite high compared to traditional Li-ion cells, which they otherwise resemble closely. There are companies that work on and commercialize these types of batteries, for example, Excellatron, Cymbet, and Infinite Power Solutions.

88.3.2 Supercapacitors

Supercapacitors feature activated carbon electrodes that have very large surface areas and that may be separated by distances as short as in the molecular range. These characteristics enable higher capacities per unit area than that of regular capacitors. The electrodes are impregnated with an electrolyte that enables charge transfer via ions. There are two types of electrolytes: *organic* and *aqueous*. Organic electrolytes provide higher voltage per cell and higher energy densities, whereas aqueous electrolytes lead to smaller cell-to-cell voltage differences.

Figure 88.5a shows a simple electric model for a supercapacitor, which is also used for regular capacitors, whereby C is the nominal value of the capacitor, R_C models the ESR, and R_{leak} models the supercapacitor leakage current. There are more complex supercapacitor models, such as one which includes several *RC* branches in parallel. Figure 88.5b shows an example with three *RC* branches and a parallel R_{leak} that models the supercapacitor leakage. The main branch (R_fC_f) is equivalent to CR_C in Figure 88.5a. The other *RC* branches model longer time constants (from seconds or minutes to days). Consequently, fully charging or discharging a supercapacitor may take several hours or days. However, C_f is much larger than the model's remaining capacitors, and consequently, so is its corresponding available energy. The leakage current is the current that passes through R_{leak} once the capacitor has been fully charged or discharged (i.e., the capacitors of all the branches) to some fixed voltage. Other complex models of supercapacitors can be found in [25] and [26].

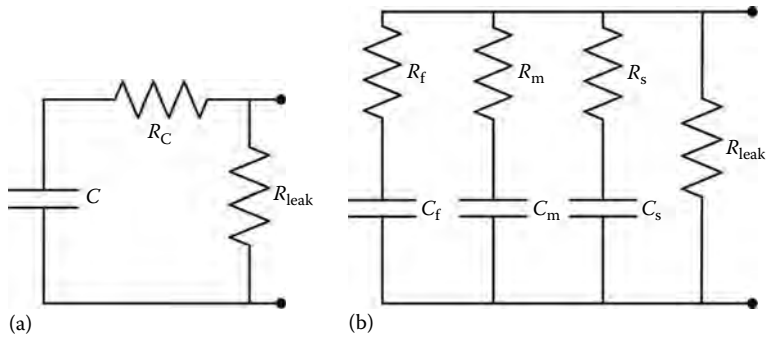


FIGURE 88.5 (a) Simple supercapacitor circuit model and (b) a more complex supercapacitor circuit model.

Supercapacitors designed for backup applications not only have a relatively high R_C (large voltage drop for a step current load) but also have a high R_{leak} (low leakage). Conversely, supercapacitors with very low R_C (tens to hundreds of milliohms) have a relative low R_{leak} (high leakage). Leakage current increases with increasing temperature (similarly to batteries) and with increasing voltage. Several companies offer supercapacitors, for example, Cap-XX, Maxwell, and AVX.

Compared to secondary batteries, supercapacitors feature energy densities of one to two orders of magnitude lower. However, their power densities can be much higher (and conversely, their internal impedances can be much lower) and offer longer lifetimes (in terms of the number of charging/discharging cycles). Additionally, their DoD can be easily obtained by measuring their OCV. However, the voltage decreases linearly with the DoD, which poses more challenges to the preceding energy conditioning stage and to the ensuing power conditioning stage. Moreover, supercapacitors with good power performance can be quite expensive. The advantages of secondary batteries and supercapacitors can be gained through their tandem use in *hybrid-storage units*, as will be reported as follows.

88.3.3 Hybrid Storage

Due to the internal impedance of the battery (Figure 88.3b) and the pulsed current profile of autonomous sensors, a voltage drop across the battery terminals appears at the activation time of the autonomous sensor. This phenomenon, which is more pronounced for high-impedance batteries, can hinder the full extraction of the available energy from the battery. Additionally, some power (and therefore, energy) is lost at the battery impedance. Small batteries are thus greatly affected.

Given that capacitors are complementary to batteries [27,28], they have been harnessed to increase the power capabilities of batteries, even in commercial products [29]. Storage units comprising a battery and a capacitor operating in tandem have been proposed for pulsed loads [30–32]. Whenever the internal impedance of the supercapacitor is much lower than that of the battery, the equivalent impedance of the hybrid-storage unit is dominated by that of the supercapacitor. Thus, when high-impedance batteries are equipped with an appropriate capacitor in parallel, their peak-power capabilities are enhanced, the power losses are reduced, and the amount of energy that can be extracted is increased. In [33], run-time tests using a sensor node as the load and an ML2016 lithium battery with and without a 0.1 F tandem supercapacitor were performed achieving run-time extensions up to 33% when using different power conditioning alternatives.

References

1. M. Zakrzewski, S. Junnila, A. Vehkaoja, H. Kailanto, A. M. Vainio, I. Defee, J. Lekkala, J. Vanhala, and J. Hyttinen, Utilization of wireless sensor network for health monitoring in home environment, in *Proceedings of IEEE International Symposium on Industrial Embedded Systems*, SIES, Lausanne, Switzerland, pp. 132–135, 2009.
2. S. Lintelman, R. Robinson, L. Mingyan, L. Bushnell, R. Poovendran, and K. Sampigethaya, Secure wireless collection and distribution of commercial airplane health data, *IEEE Aerospace and Electronic Systems Magazine*, 24, 14–20, 2009.
3. G. Barrenetxea, F. Ingelrest, G. Schaefer, and M. Vetterli, Wireless sensor networks for environmental monitoring: The SensorScope experience, in *Proceedings of IEEE International Zurich Seminar on Communications*, Zurich, Switzerland, pp. 98–101, 2008.
4. Urbiotica homepage. Available on: www.urbiotica.com (accessed on August 2, 2012).
5. A. Sinha and A. Chandrakasan, Dynamic power management in wireless sensor networks, *IEEE Design & Test of Computers*, 18, 62–74, 2001.
6. V. Raghunathan, A. Kansal, J. Hsu, J. Friedman, and M. Srivastava, Design considerations for solar energy harvesting wireless embedded systems, in *Proceedings of Fourth International Symposium on Information Processing in Sensor Networks*, Los Angeles, CA, pp. 457–462, 2005.

7. P. D. Mitcheson, E. M. Yeatman, G. K. Rao, A. S. Holmes, and T. C. Green, Energy harvesting from human and machine motion for wireless electronic devices, *Proceedings of the IEEE*, 96, 1457–1486, 2008.
8. S. Dalola, M. Ferrari, V. Ferrari, M. Guizzetti, D. Marioli, and A. Taroni, Characterization of thermoelectric modules for powering autonomous sensors, *IEEE Transaction on Instrumentation and Measurement*, 58, 99–107, 2009.
9. T. Le, K. Mayaram, and T. Fiez, Efficient far-field radio frequency energy harvesting for passively powered sensor networks, *IEEE Journal of Solid-State Circuits*, 43, 1287–1302, 2008.
10. J. P. Thomas, M. A. Qidwai, and J. C. Kellogg, Energy scavenging for small-scale unmanned systems, *Journal of Power Sources*, 159, 1494–1509, 2006.
11. S. Roundy, D. Steingart, L. Frechette, P. Wright, and J. Rabaey, Power sources for wireless sensor networks, in *Proceedings of 1st European Workshop on Wireless Sensor Networks*, EWSN, Berlin, Germany, pp. 1–17, 2004.
12. S. F. J. Flipsen, Power sources compared: The ultimate truth? *Journal of Power Sources*, 162, 927–934, 2006.
13. N. Mokhoff. (2006, October 27). IEDM feels the power. Online EETimes Europe. Available: http://www.eetimes.com/electronics-news/4065336/IEDM-feels-the-power (accessed on August 2, 2012).
14. J. A. Hagerty, F. B. Helmbrecht, W. H. McCalpin, R. Zane, and Z. B. Popovic, Recycling ambient microwave energy with broad-band rectenna arrays, *IEEE Transaction on Microwave Theory and Techniques*, 52, 1014–1024, 2004.
15. S. Roundy, J. M. Rabaey, and P. K. Wright, *Energy Scavenging for Wireless Sensor Networks, with Special Focus on Vibrations*, 1st edn., Norwell, MA, Kluiwer Academic Publishers, 2004.
16. S. P. Beeby, M. J. Tudor, and N. M. White, Energy harvesting vibration sources for microsystems applications, *Measurement Science and Technology*, 17, R175, 2006.
17. R. J. M. Vullers, R. van Schaijk, I. Doms, C. Van Hoof, and R. Mertens, Micropower energy harvesting, *Solid-State Electronics*, 53, 684–693, 2009.
18. A. Kurs, A. Karalis, R. Moffatt, J. D. Joannopoulos, P. Fisher, and M. Soljacic, Wireless power transfer via strongly coupled magnetic resonances, *Science*, 317, 83–86, 2007.
19. K. V. Schuylenbergh and R. Puers, *Inductive Powering: Basic Theory and Application to Biomedical Systems*, Analog circuits and signal processing series, Springer, New York, 2009.
20. M. Winter and R. J. Brodd, What are batteries, fuel cells, and supercapacitors? *Chemical Reviews*, 104(10), 4245–4270, 2004.
21. D. Linden and T. B. Reddy, *The Handbook of Batteries*, 3rd edn, McGraw-Hill, New York, 2001.
22. Woodbank Communications, Electropaedia. Battery and Energy Technologies, Available on: http://www.mpoweruk.com/ (accessed on 2nd August, 2012).
23. N. J. Dudney, Thin film batteries for energy harvesting, in *Energy Harvesting Technologies*, S. Priya and D. J. Inman, Eds., New York: Springer Science+Business Media, pp. 355–363, 2009.
24. N. J. Dudney, Solid-state thin-film rechargeable batteries, *Materials Science and Engineering B*, 116, 245–249, 2005.
25. S. Buller, E. Karden, D. Kok, and R. W. De Doncker, Modeling the dynamic behavior of supercapacitors using impedance spectroscopy, *IEEE Transaction on Industry Applications*, 38, 1622–1626, 2002.
26. J. N. Marie-Francoise, H. Gualous, and A. Berthon, Supercapacitor thermal- and electrical-behaviour modelling using ANN, *IEE Proceedings of Electric Power Applications*, 153, 255–262, 2006.
27. G.A. Rincon-Mora and M. Chen. (2006). Squeezing operational life out of a shrinking energy capsule. Power Management Design Line. Available on: http://www.powermanagementdesignline.com/howto/194500129 (accessed on August 3, 2012).
28. T. Christen and M. W. Carlen, Theory of Ragone plots, *Journal of Power Sources*, 91, 210–216, 2000.

29. Tadiran Batteries. PulsesPlus Batteries for High Current Pulse Applications. Available on: http://www.tadiran.com/index.php/pulses-plus-lithium-cells (accessed on August 2, 2012).
30. T. A. Smith, J. P. Mars, and G. A. Turner, Using supercapacitors to improve battery performance, in *Proceedings of 33rd Annual IEEE Power Electronics Specialists Conference*, PESC, Cairns, Queensland, Australia, pp. 124–128, 2002.
31. R. A. Dougal, S. Liu, and R. E. White, Power and life extension of battery-ultracapacitor hybrids, *IEEE Transaction on Components and Packaging Technologies*, 25, 120–131, 2002.
32. C. E. Holland, J. W. Weidner, R. A. Dougal, and R. E. White, Experimental characterization of hybrid power systems under pulse current loads, *Journal of Power Sources*, 109, 32–37, 2002.
33. M. T. Penella and M. Gasulla, Runtime extension of low-power wireless sensor nodes using hybrid-storage units, *IEEE Transaction on Instrumentation and Measurement*, 59, 857–865, 2010.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

89

Wireless Sensing Technology

Gregory C. Willden
Southwest Research Institute

Ben A. Abbott
Southwest Research Institute

Ronald T. Green
Southwest Research Institute

89.1 Introduction and General Definition **89**-1
89.2 Wireless Sensor Networks **89**-2
89.3 Choosing Reasonable Wireless Sensor Technology Solutions **89**-2
89.4 Example System: Perch and Move **89**-3
89.5 Example System: Localization for Landslide Measurements **89**-3
89.6 Example System: Neutrally Buoyant Sensor Cave Mapping **89**-5
89.7 Example System: NBS Variant for Borehole-Accessed Caves **89**-8
89.8 Example System: NBS Variant for Culvert Inspection **89**-9
89.9 Conclusions and General Design Considerations **89**-10
References **89**-11

89.1 Introduction and General Definition

While specialized wireless technologies have been used to aid the packaging and deployment of sensor systems for decades—particularly in telemetry systems—support has moved to more generalized wireless sensor network (WSN) systems. Commonplace cell and Wi-Fi backbones provide robust options for data delivery, and these mechanisms are further augmented by simpler, lower cost, and lower power delivery technologies such as Bluetooth, IEEE 802.15.4, ZigBee, and 6LoWPAN. Of late, the smartphone and smart-grid markets have accelerated-related technologies. This rapid growth, high competition, and innovative solution space should be considered when designing sensing systems that require small size, low power, and distributed solutions with limited or no interconnection wires. Unfortunately, the lexicon of terminology used within the markets overlaps and lacks precision, so this chapter focuses more on concepts as opposed to definitions. And since there are countless possible uses of the technologies, the chapter utilizes examples of the real systems and the lessons learned to provide a framework for making design decisions.

Much of the genesis of current wireless sensor technology can be attributed to the smart dust and subsequent research directions pioneered by Defense Advanced Research Projects Agency (DARPA) (www.darpa.mil). The work evolved into the creation of the "mote"* (based on the meaning of mote as a small spec or grain of dust). The systems involved low-cost, battery-powered sensors equipped with wireless networking capability. Some formulations included the ability for sensors to transmit data on behalf of other sensors to create a wireless mesh network for increased reliability.

* Use of the word "mote" is based on the meaning as a small spec or grain of dust. Application to wireless sensor nodes can likely be attributed to the Berkeley team performing on the DARPA smart dust project, but subsequently the term was adopted by many organizations.

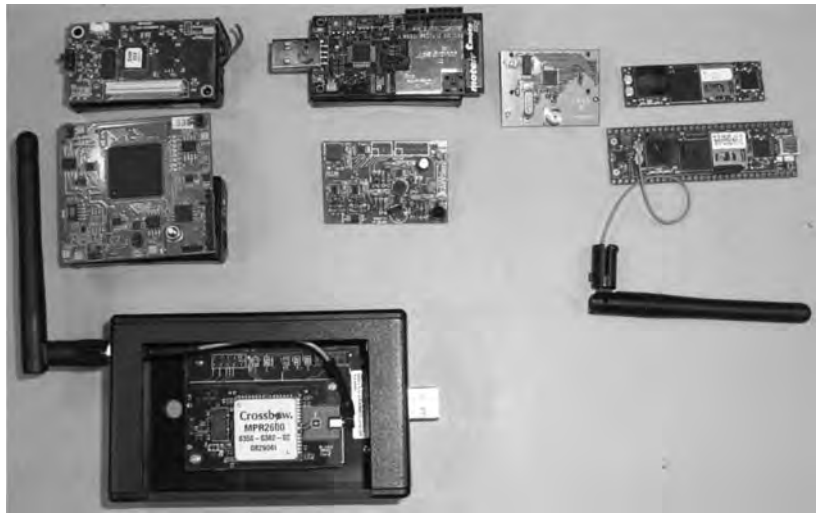


FIGURE 89.1 Example wireless sensor boards and processors (left to right): (upper left) crossbow mote WSN circa 2003, a spin-off of Berkeley mote development; (left middle) Xilinx Spartan-based application-specific mote expansion board; (bottom left) crossbow mote gateway module; (top) TMote Sky Berkeley mote; (middle) EasySen sensor expansion module for TMote Sky; (top) TI Chipcon CC2510 wireless processor evaluation module; (top right) Gumstix Overo computer on module; and (bottom right) Gumstix Pinto-TH Overo expansion module.

Generally speaking, a *wireless sensor node*

- Is battery powered
- Collects measurements
- Communicates wirelessly
- May employ mesh networking

89.2 Wireless Sensor Networks

WSNs are often formed by assembling a collection of wireless sensor nodes with a gateway node that provides data connectivity to the Internet, to the outside systems, or directly to the users in a closed system. A variety of technologies can be used in the design of wireless sensor nodes. Communication technologies vary in complexity, power requirements, maximum data rates, frequency bands of operation, and communication range. Approaches to increase system lifetime and minimize battery requirements are typically employed, and some systems employ energy-harvesting techniques that convert solar, wind, or kinetic energy to extend the life of the nodes. Processors used in wireless sensor systems are frequently highly integrated devices that combine onboard clocks, wireless communication, sensing electronics, and low-power modes into a single electronics chip or compact-printed circuit board. Figure 89.1 shows a few wireless sensor nodes of this class and examples of historical progression.

89.3 Choosing Reasonable Wireless Sensor Technology Solutions

At the height of the smart dust program* at DARPA, there was significant hype regarding the potential of this technology.† Many other research organizations, funding agencies, and universities joined in the race of working with motes as early adopters and applying them to a wide variety of projects.

* The smart dust project started in the late 1980s, and most of the work slowed by 2000.

† During this period of hype, the DARPA program management team commented to one of the authors that they hoped the technology became so ubiquitous that we would soon hear the phrase "Gee, the motes are thick today."

Online searches at the time returned headlines that boasted of how motes and smart dust would change our lives in areas as diverse as military combat, plant life, medical understanding, and pollution control. But today, we do not see the reality mirroring the magnitude of the dream as it had been envisioned. Minimum battery size, system power requirements, and packaging are among the limitations of the dream. While the technology has not lived up to the initial hype, the results of the smart dust program, and other related research, have produced valuable technological advances that should be considered when designing new sensor systems. Taking an objective look at successful wireless sensor designs and deployments will equip the system designer with a useful toolbox of capabilities and a measure of wisdom in how best to employ them to solve real-world sensing problems.

The following WSNs, which the authors developed or contributed to, illustrate design trade-offs and challenges associated with developing and deploying sensors in the real world. Each example concludes with a list of design trade-offs and how well they worked for the particular problem. We learned many things and hope you are able to utilize our lessons to build better systems of your own.

89.4 Example System: Perch and Move

The Perch and Move program [4] was funded by DARPA to demonstrate a remote surveillance and signal intelligence capability deployed on a portable autonomous helicopter. The wireless sensor portion of the program was designed to perform localization of handheld communication devices using radio-frequency (RF) signatures. The device was designed to meet strict size, weight, and power requirements and employed a low-cost Global Positioning Satellite (GPS) unit, a reverse-engineered radio for signal acquisition, and a digital signal processor for signal demodulation and analysis. Node-to-node communication was accomplished using a 9600 bit/s wireless modem. The GPS module provided Universal Time Coordinated (UTC) time and latitude, longitude, and altitude estimates. The module also provided a 1 pulse per second (PPS) signal accurate to approximately 40 ns. A least-mean-square algorithm was used to reduce l PPS jitter and increase the accuracy of time of arrival (TOA) calculations on intercepted signals. The system consisted of a group of homogenous sensors, some deployed on the ground and others on the airborne platform that jointly determined the location of a noncooperative transmitter using a time difference of arrival (TDOA) algorithm.

Radio-wave propagation modeling tools estimated the received signal strength and radio multipath effects for candidate sensor locations by generating city models from existing satellite imagery. Sensor battery life, vehicle flight times, platform stealth, noise, localization accuracy, and RF modeling all factored into a multi-constraint optimization for distributed coordination of sensor locations and flight paths. Design trade-offs and choices related to the development of the Perch and Move system are summarized in Table 89.1.

89.5 Example System: Localization for Landslide Measurements

Although monitoring of certain earth processes (e.g., volcanic eruption, flooding) may call upon very specific measurements (e.g., temperature, atmospheric or soil gas composition, water level), the precise knowledge of the location of sensor nodes is a universal data requirement and for some applications—such as surface displacement associated with a landslide—the only data requirement. According to the U.S. Geological Survey, landslides represent a widespread geologic hazard that results in hundreds of deaths and billions of dollars in property damage throughout the world each year.

WSN designers frequently pursue the ability for individual nodes to determine their own location by estimating node-to-node distances using the round-trip communication times to reference nodes at known locations. However, the primary limitations in this approach to the self-localization problem are the uncertainty created by the small frequency differences between the local oscillators on each node and the bandwidth of the communication signal. Typical localization accuracy for this approach is on the order of meters.

TABLE 89.1 Design Trade-Off Decisions for Perch and Move System

Design Element	Design Selection	Consequences
Power source	Lithium polymer battery	Pros: high power-to-weight ratio was required to extend limited flight distances and short station keeping times of the airborne platform to minimally acceptable levels Cons: cost and availability
Power conservation	Scheduling of sleep modes in an application-specific manner	Pros: fully on when likely to capture events of interest Cons: required application-specific information to design scheduling
Processing	Onboard data reduction	Pros: reduced communication requirements increased mission stealth Cons: increased size, weight, power, and cost; complex algorithm creation required
Networking	Commercial off the shelf (COTS) wireless modem	Pros: extremely reliable Cons: low data rate
Specialized hardware	Customized GPS for clock synchronization	Pros: made it possible to use distributed processing Cons: increased size, weight, power, and cost
Packaging and fabrication	COTS components with wiring interconnects	Pros: overall lower cost for small-volume deployments (<12 units), easily accessible components Cons: reliability suffers due to wiring interconnects, increased size, higher cost for larger volume deployments, and increased assembly time
Deployment	Attach preassembled sensor to existing delivery system	Pros: delivery is independent of sensor node Cons: some duplicate components (GPS)

The authors implemented a sensor network of spatially distributed nodes and deployed it on an active landslide area to measure the continuing motion and settling of the formation over a period of months [1,3]. The sensor nodes were built around the Texas Instruments CC2510 development kit that combined an 8051 microcontroller with a 2.4 GHz radio transceiver employing the 802.15.4 format. With a usable bandwidth of about 750 kHz, the one-way calibrated arrival accuracy is about 1 ms, which yields a range accuracy of about 300 m for any given processed burst. Additional averaging of bursts reduces this uncertainty by the square root of the number. For example, 10 cm accuracy may be achieved using nine million bursts, which is not a problem for stationary objects. However, there is a lower bound on the uncertainty due to (1) the ability to properly calibrate the system for the turnaround time from the responder and (2) the repeatability of the turnaround time. By avoiding baseband processing (perhaps tapping the system at the intermediate frequency), the turnaround repeatability is very good (i.e., picoseconds). For this reason, the developed system used separate RF circuitry to create a cooperative radar-like system that enabled a very accurate measurement of the changes in relative distances between nodes.

Because the primary objective of this effort was to build a network that measures position with about one centimeter accuracy, a simplified flooding-based communication and control program was developed such that all nodes allow control from any other node and flood messages to all of the nodes. This is a classic ad hoc networking approach that results in an extremely reliable and flexible delivery mechanism at the cost of excessive power use. However, under this scenario, the power used by the networking protocol is far less than that of the power required to make a distance measurement. Thus, the focus on optimizing this algorithm was not deemed worthwhile with respect to the research goals. The data collected by the system showed most of the nodes performed as designed and generated

TABLE 89.2 Design Trade-Off Decisions for Landslide Measurement System

Design Element	Design Selection	Consequences
Power source	Commercial-sealed gel cell battery	Pros: rechargeable, reliable, long life, low cost, easily acquired, relatively unaffected by temperature (at least over typical operating ranges) Cons: poor power density (large and heavy), disposal
Power conservation	Extensive use of deep sleep	Pros: current draw on the order of microamperes during sleep Cons: required relays for powering radar electronics, clock drift, no sampling while sleeping
Processing	Wireless processor for coordinating data acquisition and storage. Post-processing for data reduction	Pros: lower battery use, simpler to test and debug Cons: requires larger data storage per sensor node for unreduced sample storage, requires post-processing to verify system is functional; in situ testing requires extensive equipment and greater time than similar lab-based testing
Networking	Brute force network flooding, node locations allow for redundant coverage	Pros: extremely reliable Cons: increased time and power requirements; however, these may be offset by sleep duty cycle
Specialized hardware	Cooperative radar for node localization	Pros: sub-centimeter accuracies not obtainable through processing WSN communication channels Cons: increased size, weight, power, and cost; requires multiple antennas
Packaging and fabrication	Commercial watertight cases and COTS components with wiring interconnects	Pros: overall lower cost for small-volume deployments (<12 units), easily accessible components (some may be acquired near deployment sites) Cons: reliability suffers due to wiring interconnects, increased size, and higher cost for larger volume deployments
Deployment	Final assembly and camouflage performed during deployment	Pros: match the deployment to environmental conditions (vegetation, rock, land features, and color), protect the sensor components during transport to remote site, and adjust antenna height to increase communication reliability. Cons: longer deployment, more extensive in situ checkout needed

radar returns in the sub-centimeter range. Laboratory-based tests measured the units to generate measurements close to 3 mm in accuracy. Design trade-offs and choices related to the development of the landslide measurement system are summarized in Table 89.2.

89.6 Example System: Neutrally Buoyant Sensor Cave Mapping

Many parts of the world rely on karst aquifers, essentially interconnected water-filled caves and passageways, for their water supply needs. Conventional methods for characterizing karst conduits include tracer tests and mapping by cave divers. Dye tracers are useful for determining flow velocities but do not provide sufficient information about the size and complexity of the karst formation. The use of cave divers is very limited to those formations that are large enough to accommodate divers and slow enough to allow them to enter and exit safely and are located at reasonable depths and distances. Since automated sensors do not share those same limitations, the authors developed a sensor technology designed to travel with the flow of water and measure the size and morphology of caves through which it travels.

This patented sensor technology developed at Southwest Research Institute® (SwRI®) is referred to as the neutrally buoyant sensor, or NBS.

The NBSs were assembled using commercially available components such as ultrasound sensors and dual-axis magnetometers and accelerometers. A general-purpose wireless sensor node with onboard processor and memory unit coordinated the overall operation of the data collection components. The magnetometer provided the ability to gauge the orientation relative to magnetic north, and the accelerometer enabled the determination of motion dynamics (furnishing a form of inertial navigation) as the NBS traveled through conduits. Ultrasonic transducers characterized the conduit geometry and, ultimately, velocity. Six pairs of ultrasonic transducers were positioned on each axis around the circular NBS vessel to emit sonar "pings" outward to the conduit walls. Each pair consisted of a transmitter, which sent out ultrasound pulses normal to the sensor node, and a receiver, which recorded the reflected ultrasound pulse. Accurate distances to the surrounding conduit features are determined using the TOA of the first reflected pulse. The time of flight was converted to distance based on the density of water and air. Sonar ranging gives a representative characterization of the conduit geometry. The rate of displacement of distinctive wall features is used in velocity calculations. Post-processing of the data from all components enables the calculation of real-time velocity of the NBS as it travels along the conduit path, as well as calculation of the shape and size of the conduit. The orientation (magnetometer) data are used in combination with the sonar readings to detect features, derive velocity, and ultimately unravel the data into a chain of vectors corresponding to the path taken, velocity profile, and measured conduit shape. Similarly, the floating sonar coordinates are rectified to Cartesian position and distributed to the respective

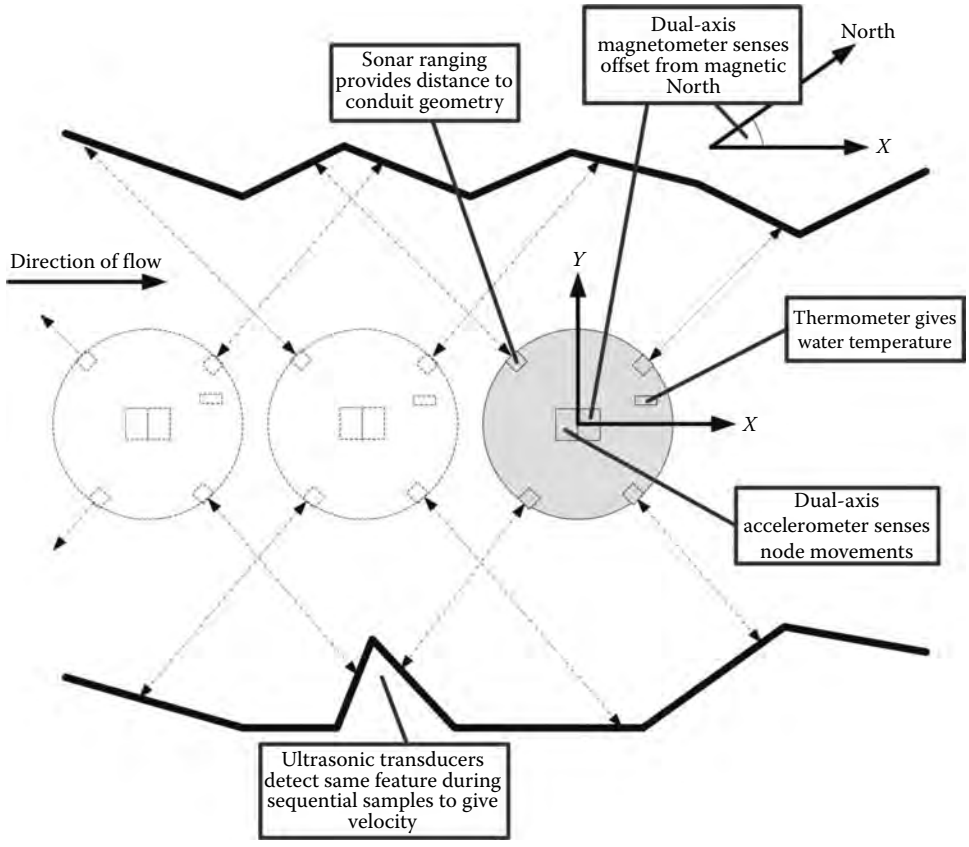


FIGURE 89.2 NBS concept of operations.

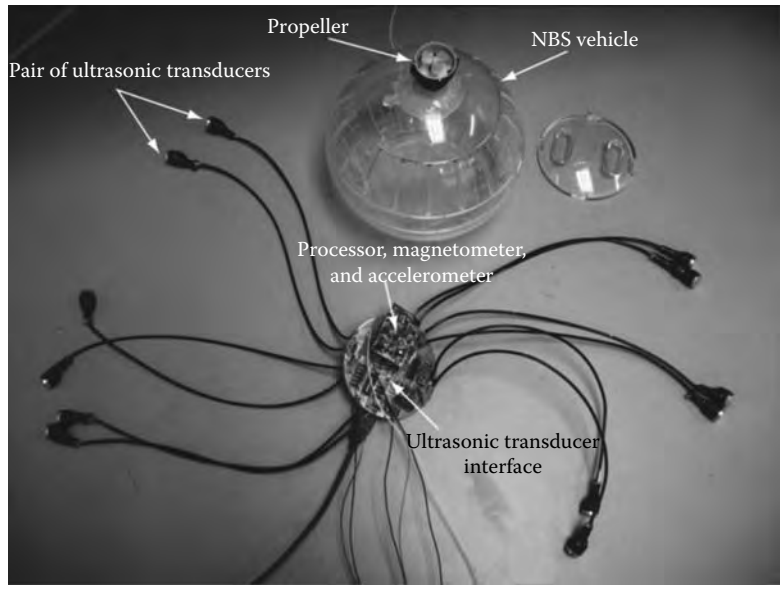


FIGURE 89.3 NBS electronics and housing.

vector segment. Accelerometers are used to detect if any periods of rapid movement occur since these could adversely affect the triangulation used in mapping. The conceptual operation of the NBS is illustrated in Figure 89.2. The prototype circuit assembly with enclosure, built from a "hamster ball" from the local pet store, is shown in Figure 89.3.

A deployment of the NBS in 2008 resulted in a transport trajectory that was sufficiently long and uninterrupted to allow collection of ultrasound, magnetometer, and accelerometer data of a 20 m long stretch of Honey Creek Cave in Texas. The detailed process for data reduction is outlined in [2]. The resulting 3D conduit image is shown in Figure 89.4. Design trade-offs and choices related to the development of the NBS system are summarized in Table 89.3.

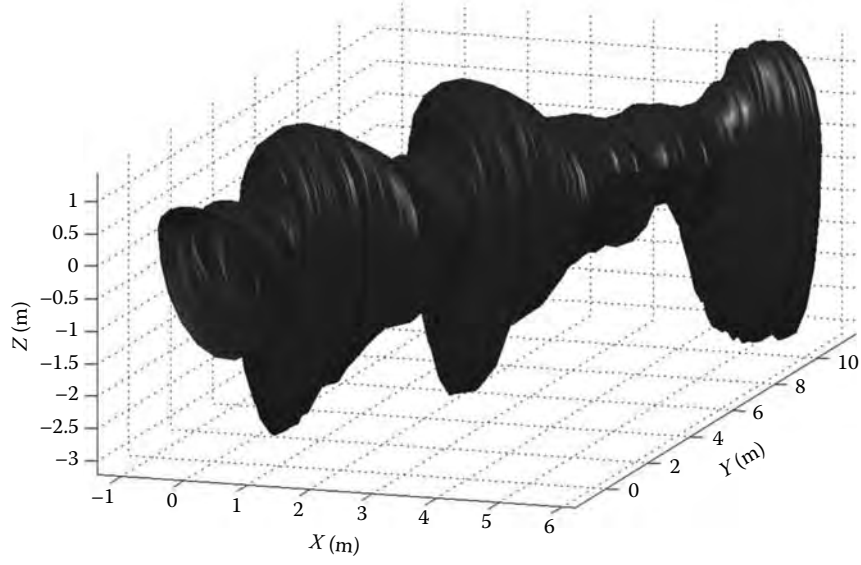


FIGURE 89.4 3D conduit imagery resulting from data collected during cave-mapping deployment.

TABLE 89.3 Design Trade-Offs for NBS

Design Element	Design Selection	Consequences
Power source	Single-use lithium cells	Pros: power density ratio Cons: environmental impact, in situ battery replacement
Power conservation	Low-power processor kept off until use	Pros: continuous data collection and communication Cons: reduced mission lifetime
Processing	Wireless processor for coordinating data acquisition and storage. Post-processing for data reduction	Pros: lower battery use, simpler to test and debug Cons: requires larger data storage per sensor node for unreduced sample storage, requires post-processing to verify system is functional
Networking	Simple point to point and relay	Pros: simple (specialized communications were needed to work in a cave) Cons: lack of robustness, lack of features, low data rate
Specialized hardware	Sonar transmit and receive circuitry-augmented WSN	Pros: required to collect desired measurements, primarily used WSN processor
Packaging and fabrication	Repurposed a commercially available housing, commodity automotive sonar transducers, wiring interconnects, custom waterproofing	Pros: overall lower cost for small-volume deployments (<12 units), tremendous cost savings from commodity sonar transducers Cons: reliability suffers due to wiring interconnects, higher cost for larger volume deployments, and many waterproofing challenges
Deployment	Magnetic relay for waterproof power switch, in situ buoyancy adjustment	Pros: low cost Cons: time to perform sensitive buoyancy adjustment

89.7 Example System: NBS Variant for Borehole-Accessed Caves

The U.S. Army Corps of Engineers (USACE) pursued a project to locate, characterize, and close off karst conduits that divert water around the Center Hill Dam in Tennessee. The project involved drilling regularly spaced boreholes deep into the rock on the south side of the dam with the goal of intersecting the conduits and then filling them with grout. Knowing the size and orientation of the conduits can guide subsequent drilling efforts and provide estimates of the amount of grout required to fill the voids. To that end, the USACE contracted with SwRI to develop a customized version of the NBS technology for mapping borehole-accessed caves.

The original NBS electronics package was redesigned to fit inside a section of two inch PVC tube with ultrasonic transducers oriented at 0°, 90°, 180°, and 270° angles around the tube. This tethered probe transmitted real-time data to the surface via a coaxial cable connected to the wireless radio antenna port. The tether cable was wound onto a cable reel instrumented with rotational encoders for measuring the depth of the probe. An antenna was connected to the top end of the tether cable to provide slip-ring functionality to the cable reel. A conscious design decision was made not to include a means of controlling the rotation of the probe as it descended and ascended through the cavity but to allow it to rotate freely as the cable allowed and measure the orientation with a magnetometer. This decision is in harmony with SwRI's basic low-cost approach to sensors and WSNs, that leveraging inexpensive COTS components enables more ubiquitous and advanced data collection for solving real-world problems. Design trade-offs and choices related to the development of this variant of the NBS system are summarized in Table 89.4.

TABLE 89.4 Design Trade-Offs for NBS Variant for Borehole-Accessed Caves

Design Element	Design Selection	Consequences
Power source	Single-use lithium cells	Pros: power density ratio Cons: in situ battery replacement
Power conservation	Low-power processor kept off until use	Pros: continuous data collection and communication Cons: reduced mission lifetime
Processing	Wireless processor transmitted data to control terminal over commodity coaxial tether cable via RF communications	Pros: continuous live data, ability to use extremely long tethers of commodity cable (RF through air would not work in long boreholes; conventional-wired networking protocols could not support distance requirements), extremely low cost, and memory requirements on WSN
Networking	Simple point to point	Pros: simple, robust networking
Specialized hardware	Sonar transmit and receive circuitry-augmented WSN, commodity 3-axis magnetometer	Pros: required to collect desired measurements, primarily used WSN processor, low cost
Packaging and fabrication	Specialized housing assembled from commodity PVC to match borehole dimensions, commodity coax cable dual purposed for tether and communications	Pros: overall lower cost for small-volume deployments (<12 units), tremendous cost savings from commodity components Cons: reliability suffers due to wiring interconnects, higher cost for larger volume deployments, and many waterproofing challenges; commodity tether strength is lower than classic downhole applications (some sensors were lost due to broken cables)
Deployment	Commodity motorized cable reel with depth measurement	Pros: low-cost expendable sensors and cable allowed measurements in higher-risk locations (competing teams were unwilling to risk loss of higher-cost sensors) Cons: lost probe resulted in minor delays

89.8 Example System: NBS Variant for Culvert Inspection

In its role of ensuring the safety of America's roads, the Federal Highway Administration (FHWA) sought out tools to facilitate and automate the inspection of culvert conduits. The previously described NBS projects offered a compelling alternative to manual inspection or expensive tethered robotic cameras. The FHWA funded development of a specialized version of the NBS system that included a multichannel, a high-speed analog-to-digital converter (ADC) for higher-resolution ultrasonic imagery, a 360° camera and lighting system for visual confirmation of sensor results, and integrated Wi-Fi for live streaming video and data retrieval. The combination of more channels of ultrasonic data, higher sampling rates, and streaming video required the use of more powerful processors and a full embedded Linux operating system. The embedded Linux provided full wireless networking support and digital signal processor (DSP)-accelerated video compression routines including the GStreamer libraries for network streaming video. The user controls the data acquisition and views the live results through a graphical user interface hosted on a ruggedized convertible laptop–tablet. This application is an example of how performance requirements can drive the solution beyond the capabilities of traditional WSN hardware and software environments. Design trade-offs and choices related to the development of the culvert inspection system are summarized in Table 89.5.

TABLE 89.5 Design Trade-Offs for Culvert Inspection System

Design Element	Design Selection	Consequences
Power source	Rechargeable lithium polymer battery	Pros: high power density Cons: cost of battery and charging infrastructure
Power conservation	Increased battery capacity to support compute requirements	Pros: supported high capability processing Cons: increased size and cost
Processing	Primary data reduction and processing is accomplished on the sensor	Pros: dramatically reduced data communication requirements Cons: increased size, weight, power, and cost
Networking	Commodity Wi-Fi	Pros: high throughput, simple, interoperability with commodity ruggedized laptops Cons: increased power requirements
Specialized hardware	360° camera, inertial measurement unit, high-speed ADC, power management	Pros: required to collect desired measurements, provides fusion of sonar and visual imagery Cons: increased size, weight, power, and cost
Packaging and fabrication	Custom circuit boards, custom housing	Pros: reliability Cons: higher per-unit cost for lower volumes means sensors are not expendable unless manufacturing volumes are high
Deployment	Commodity tether with Bluetooth depth/distance measurement	Pros: wireless communication supports broader range of deployment scenarios Cons: wireless may be limited in some scenarios

89.9 Conclusions and General Design Considerations

We have attempted to show the strengths and weaknesses of the WSN approach by examining the trade-offs and design choices from a series of successful deployments. WSN platforms provide a compelling option for low-power applications with straightforward measurement requirements across spatially separated deployment positions. However, as the compute requirements of the system increase, a hybrid approach is often required in which a digital signal processor or conventional processor is employed to augment the processing capability of the wireless platform. Hybrid approaches may also take advantage of the low-power WSN module to wake the expanded system components from extended sleep modes. The highlighted systems also demonstrated a high performance scenario when the WSN hardware should be abandoned in favor of more powerful systems.

In general, the major design factors in the WSN application space, in order of importance, are total power usage, communications rate, system reliability, and communication range. Each of these requirements is in conflict with the others when designing a system. For example, selecting a particular communication technology has far-reaching effects upon many other areas of the system, such as power, antenna size (packaging), algorithms that can be employed, and clock synchronization. In short, every aspect feeds back onto all the other aspects of the system and must be carefully considered.

A general summary of trade-offs and design considerations for wireless sensor systems is given in Table 89.6.

TABLE 89.6 Summary of Trade-offs and Design Considerations

Design Element	Design Considerations
Power source	A variety of battery technologies, capacities, power densities, voltages, rechargeable versus single use, etc. Energy harvesting (solar, wind, kinetic, etc.) may be employed for some applications. Some battery technologies are more sensitive than others to temperature effects.
Power conservation	Reducing the computation and communication duty cycle is not always the whole answer. Often the solution requires application-specific knowledge to balance the conflicting constraints of battery lifetime, communication rate, and system reliability and communication range.
Processing	WSNs that perform a lot of processing shared across nodes are engaging in a form of distributed processing using low-power processors with all the added complexities associated with these techniques, including additional communication requirements to support system coordination. The impact on system performance is analogous to that described by Amdahl's law for multiprocessors and parallel computing.
Networking	Many wireless networking technologies are available across different frequency ranges, power levels, communication ranges, and data throughput capabilities. Some of the technologies that may be considered are Wi-Fi (including low-power implementations), Bluetooth, 802.15.4, ZigBee, 6LoWPAN, ANT, and WirelessHART. The selection of a particular wireless technology is intimately connected with the conflicting design constraints of size, weight, power, cost, and communication requirements.
Specialized hardware	Wireless sensor processors and hardware generally support basic measurements, such as temperature, but solutions to real-world problems almost always require specialized hardware. The inclusion of specialized hardware is entirely dependent on the requirements of the proposed application.
Packaging and fabrication	The featured projects highlight various choices for system packaging, assembly, and fabrication. Solutions have included examples of both disposable and high value sensors, as well as varying levels of time and effort required for assembly. For low volume systems, a higher degree of manual assembly is acceptable, and even desirable, when it leads to lower total cost. Applications featuring higher sensor volumes require additional engineering at the design stage in order to minimize total system cost in the longer term. These opposing choices can also have a direct impact on system reliability since the possibility that a failure will occur increases in direct proportion with the number of interconnects and cables.
Deployment	Employing ultralow-cost components can have huge benefits in scenarios where many nodes can be deployed to provide a measure of redundancy whereby if one node fails, the system still succeeds. In some situations, the nodes are so inexpensive and the resulting data so valuable that individual nodes may be considered disposable and deployments can be designed to accommodate a certain level of loss by attrition. However, some deployments require a more powerful and/or expensive sensor such that individual node loss is deemed unacceptable or, at the very least, undesirable.

References

1. Kenney, J.D., D.R. Poole, G.C. Willden, et al. Precise positioning with wireless sensor nodes: monitoring natural hazards in all terrains, *2009 IEEE International Conference on Systems, Man, and Cybernetics*, San Antonio, TX. October 2009.
2. Kenney, J.D., G.C. Willden, R.N. McGinnis, et al. Development of neutrally buoyant sensors for mapping conduits in karst aquifers, *11th Multidisciplinary Conference on Sinkholes and the Engineering and Environmental Impacts of Karst,* Tallahassee, FL, September 2008.
3. Pilcher, M.E., B.E. Campion, and B.A. Abbott, Wireless system using continuous wave phase measurement for high-precision distance measurement, U.S. Patent No. U.S. 7,504,992, March 2009.
4. Van Rheeden, D.R., B.C. Brown, J.C. Price, et al. Automatic positioning of UAVs to optimize TDOA geolocation performance, *IEEE-AIAA 23rd Digital Avionics Systems Conference,* Salt Lake City, UT, October 2004.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

90

Telemetry

90.1 Introduction **90**-1
90.2 Multiple-Channel Telemetry **90**-4
Frequency Division Multiplexing • Subcarrier Modulation Schemes for FDM • Time Division Multiplexing • Pulse Code Modulation in Time Division Multiplexing
Defining Terms **90**-16
References **90**-17
Relevant Websites about Telemetry **90**-17

Albert Lozano-Nieto
Penn State University

90.1 Introduction

Telemetry is the science of gathering information at some remote location and transmitting the data to a convenient location to be examined and recorded. Telemetry can be done by different methods: optical, mechanical, hydraulic, electric, etc. The mechanical methods, either pneumatic or hydraulic, have acceptable results for short distances and are used in environments that have a high level of electromagnetic interference and in those situations where, for security reasons, it is not possible to use electrical signals, for example, in explosive environments. More recently, use of optical fiber systems allows the measurement of broad bandwidth and high immunity to noise and interference. Other proposed telemetry systems are based on ultrasound, capacitive or magnetic coupling, and infrared radiation, although these methods are not routinely used. The discussion in this chapter will be limited to the most-used systems: telemetry based on electric signals. The main advantage of electric over mechanical methods is that electrically based telemetry does not have practical limits regarding the distance between the measurement and the analysis areas and can be easily adapted and upgraded in already existing infrastructures. Electric telemetry methods are further divided depending on the transmission channel that they use as wire telemetry and wireless (or radio) telemetry. Wire telemetry is technologically the simplest solution. The limitations of wire telemetry are the low bandwidth and low transmission speed that it can support. However, it is used when the transmission wires can use the already existing infrastructure, as in most electric power lines that are also used as wire telemetry carriers. Wireless telemetry is more complex than wire telemetry, as it requires a final radio-frequency (RF) stage. Despite its complexity, it is widely used because it can transmit information over longer distances; thus, it is used in those applications in which the measurement area is not normally accessible. For most applications, the wireless nature of the system is implied when referring to telemetry. The largest area that led to the development of telemetry systems arose from defense and aerospace applications, specifically the development of rockets and ballistic missiles. Designers of these systems needed to monitor a large amount of data, while the rocket was in flight, and this task could only be achieved with sending the data wirelessly.

Figure 90.1 displays a generic telemetry system. It consists of (1) transducers to convert physical variables to be measured into electric signals that can be easily processed; (2) conditioning circuits to amplify the low-level signal from the transducer, limit its bandwidth, and adapt impedance levels;

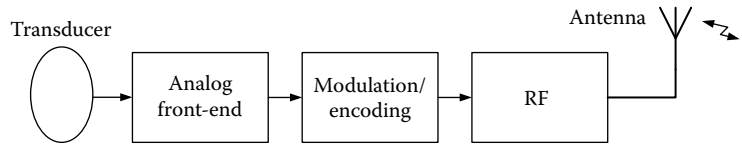


FIGURE 90.1 Block diagram for a basic wireless telemetry system.

(3) a signal-processing circuit that sometimes can be integrated in the previous circuits; (4) a subcarrier oscillator whose signal will be modulated by the output of the different transducers once processed and adapted; (5) a codifier circuit, which can be a digital encoder, an analog modulator, or a digital modulator, that adapts the signal to the characteristics of the transmission channel, which is a wire or an antenna; (6) a radio transmitter, in wireless telemetry, modulated by the composite signal; (7) an impedance line adapter, in case of wire transmission, to adapt the characteristic impedance of the line to the output impedance of the circuits connected to the adapter; and (8) for wireless communication, a transmitting antenna. It is necessary to point out that not all the blocks may be present in a specific system. The receiver end consists of similar modules. For wireless telemetry, these modules are (1) a receiving antenna designed for maximum efficiency in the RF band used, (2) a radio receiver with a demodulation scheme compatible with the modulation scheme, and (3) demodulation circuits for each of the transmitted channels. For wire telemetry, the antenna and the radio receiver are replaced by a generic front end to amplify the signal and adapt the line impedance to the input impedance of the circuits that follow. The transmission in wireless telemetry systems is done by sending a signal whose analog variations in amplitude or frequency are a known function of the variations of the signals from the transducers. The majority of current systems use digital data transmitting a finite set of symbols, each one representing one of the possible finite values of the composite signals at the time that it was sampled. The effective communication distance in a wireless system is limited by the power radiated by the transmitting antenna, the sensitivity of the receiver, and the bandwidth of the RF signal. As the bandwidth increases, the contribution of noise to the total signal also increases, and consequently more transmitted power is needed to maintain the same signal-to-noise ratio (SNR). This is one of the principal limitations of wireless telemetry systems. In some applications, the transmission to the receiver is done on baseband, after the conditioning circuits. The advantage of baseband telemetry systems is their simplicity, although because of the baseband transmission, they are normally limited to only one channel at low speeds.

For most applications, the systems need to acquire either different types of signals or the same type of data at different locations in the process that is being monitored. These different information signals can be transmitted using the same common carrier by multiplexing the data signals. Multiplexing allows different signals to share the same channel. Multiplexing techniques are usually considered either frequency division multiplexing (FDM) or time division multiplexing (TDM). In FDM, different subcarrier frequencies are modulated by the different measurement channel signals that cause the information spectrum to shift from baseband to the subcarrier frequency. Then, the subcarrier frequencies modulate the RF carrier signal, which allows the transmission of all desired measurement channels simultaneously. In TDM, the whole channel is assigned entirely to each of the measurement channels, although only during a fraction of the time. TDM techniques use digital modulation to sample the different measurement channels at different times. Then, these samples are applied sequentially to modulate the RF carrier. Figure 90.2 illustrates these concepts by showing frequency and time graphs for FDM and TDM, respectively.

Current uses of telemetry encompass a large and varied number of applications. Telemetry is widely used in space applications for either telemeasurement of a distant variable or telecommandment of actuators. In most of these types of applications, for example, in space telemetry, it is very important to design the telemetry systems to minimize the consumption of power [1], Some land mobile vehicles, such as trains, also use telemetry systems, either wireless or by using some of the existing power wires to transmit data to the central station and receive its commands [2]. In clinical

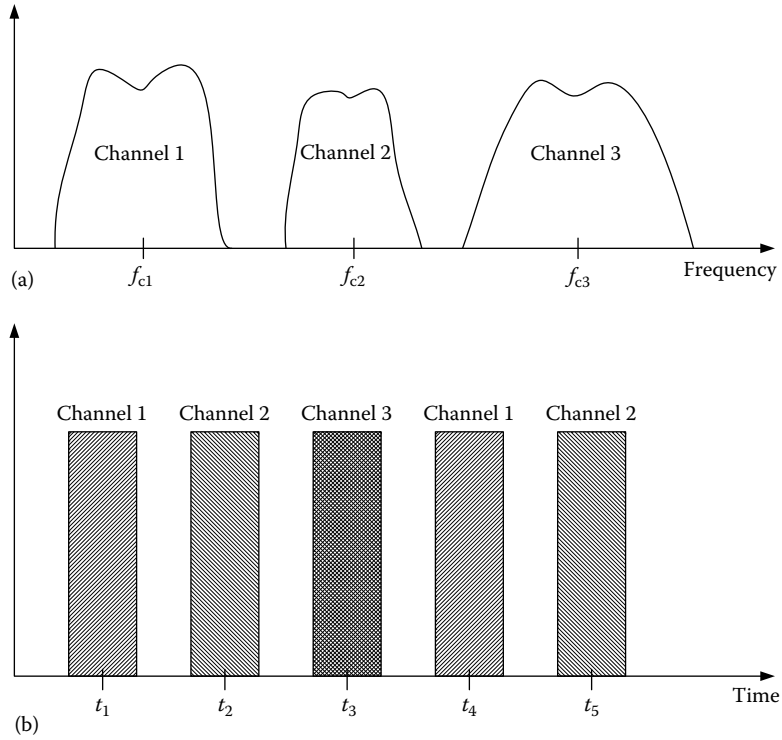


FIGURE 90.2 Basic characteristics of (a) FDM and (b) TDM signals. In FDM different channels are allocated at different subcarrier frequencies ($f_{c1}, f_{c2},\ldots$) while in TDM, only one channel is transmitted at a given time. The remaining channels are transmitted sequentially.

practice, the telemetry of patients increases their quality of life and their mobility, as patients do not need to be connected to a measurement system to be monitored. Several medical applications are based on implanting a sensor in a patient and transmitting the data to be further analyzed and processed either by radio [3] or by adapted telephone lines [4] from the receiving station. Optical sensors and fiber-optic communications are used in industry to measure in environments where it is not desirable to have electric signals such as explosive atmospheres [5]. The designer of a telemetry system needs also to keep in mind the conditions in which the system will have to operate. In most of the applications, the telemetry systems must operate repeatedly without adjustment and calibration in a wide range of temperatures. Finally, as different telemetry systems are developed, the need to permit tests to be made interchangeable at all ranges increases, which require compatibility of transmitting, receiving, and signal-processing equipment at all ranges. For this reason, the Department of Defense Research and Development Squad created the Guided Missiles Committee, which formed the Working Group on Telemetry. This later became the Inter-Range Instrumentation Group (IRIG) that developed telemetry standards. Today, the IRIG Standard 106-09 is the primary telemetry standard used worldwide by both government and industry.

Wireless telemetry systems need some coordination in order to use the radio spectrum in an efficient manner and more importantly, free of interference. This coordination takes shape in the form of standards, in particular the IRIG Standard 106-09 . While this chapter does not focus on the in-depth discussion of this or other standards, it discusses some of its more relevant and broader points. Most wireless telemetry applications occur in the UHF band. Table 90.1 shows the specific allocation frequencies within this frequency range as well as their unofficial name. The 1435–1535 MHz band is allocated in the United States for aeronautical telemetry including manned and unmanned aircraft, missiles,

TABLE 90.1 Telemetry Frequency Allocations in the UHF Band

Frequency Range (MHz)	Commonly Used Band Designation
1435–1525	Lower L-band
1525–1535	Lower L-band
2200–2290	Lower S-band
2310–2360	Upper S-band
2360–2390	Upper S-band

rocket sleds, and similar systems. The range from 1525 to 1530 MHz telemetry usage is on a secondary service, while the primary service is mobile satellite communications. The range from 1530 to 1535 MHz also supports telemetry as a secondary service, while the primary service is maritime mobile satellite communications. The band ranging from 2200 to 2290 MHz is assigned for space research, space operation, and satellites for Earth exploration including the telemetry systems associated to launch vehicles, research rockets, and space vehicles. The band ranging from 2310 to 2390 MHz is allocated to telemetry in a secondary basis, while the primary user of this band is the wireless communication system. Finally, the service allocation in the 2360–2390 MHz is for aeronautical telemetry similar to the services enumerated in the lower bands. The maximum power allowed for transmitters operating in these bands is 25 watts for both output power and effective isotropic radiated power.

90.2 Multiple-Channel Telemetry

Most of the industrial processes in which telemetry is used require the measurement of different physical variables to control the process, the measurement of only one physical variable at different locations, or normally a combination of both. In these multiple-channel measurements, baseband telemetry is not an option, as it would require building a different system for each channel. Multiple-channel telemetry is achieved by sharing a common resource (transmission channel), as is shown in Figure 90.3. The sharing of the transmission channel by all the measurement channels is designated by multiplexing. There are two basic multiplexing techniques: FDM and TDM. In FDM, different channels are assigned to different spectral bands and the composite signal is transmitted through the communication channel. In TDM, the information for different channels is transmitted sequentially through the communication channel.

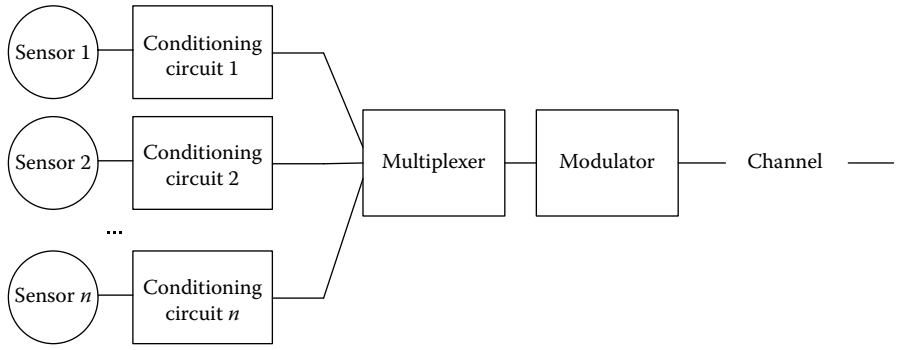


FIGURE 90.3 In multiple-channel telemetry, a common transmission channel is used to transmit the measured signals from different channels using different sharing schemes. The process is reversed at the receiver side.

90.2.1 Frequency Division Multiplexing

In FDM, shown in Figure 90.4a, each measurement channel modulates a sinusoidal signal of different frequency. These sinusoidal signals are called subcarriers. Each of the modulated signals is then low-pass filtered to ensure that the bandwidth limits are observed. After the filtering stage, all the modulated signals are fed into a summing block, producing what is known as a baseband signal. A baseband signal indicates here that the final carrier has not yet been modulated. The spectrum of the baseband signal is shown in Figure 90.4b, where it is possible to see how each measurement channel spectrum signal is allocated its own frequency. This composite signal finally modulates a carrier signal whose frequency depends on the transmission medium that is used. The signal is then fed into a transmission wire (similar to TV-broadcasting systems by cable) or, more commonly, into an antenna in the case of wireless telemetry systems. In wireless telemetry, the frequency of the carrier cannot be chosen arbitrarily, but is chosen in accordance with international agreements on the use of the electromagnetic spectrum. In the United States, the Federal Communications Commission (FCC) is the body that regulates the allocation of frequencies for different communication services. Table 90.2 shows the most common telemetry frequency bands and their intended use. Table 90.2 is for informational purposes only, and it is not a comprehensive guide to telemetry frequencies as these can change often.

At the receiver end, the carrier demodulator detects and recovers the composite baseband signal. The next step is to separate each of the subcarriers, by feeding the signal into a bank of parallel

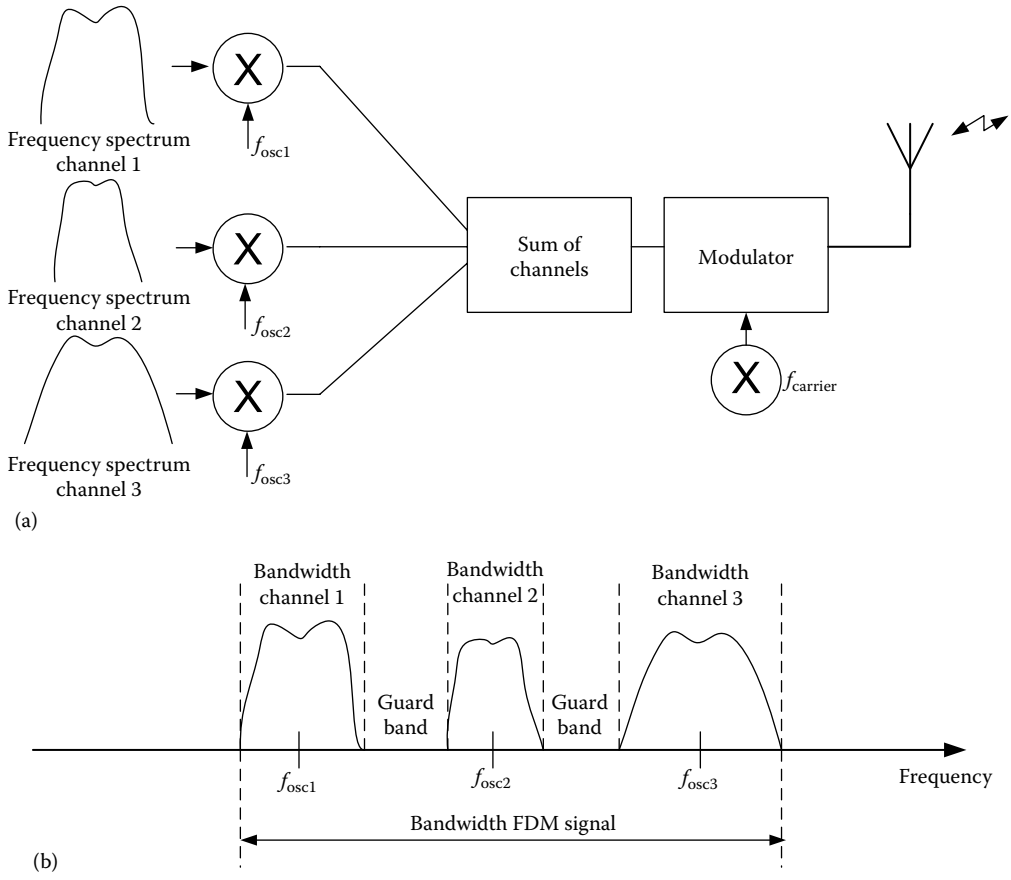


FIGURE 90.4 The different channels in an FDM system (a) are allocated at different subcarrier frequencies producing a composite signal shown in (b) that is later modulated by an RF frequency according to the transmission channel used. The guard bands limit the closeness of contiguous channels to avoid intermodulation and cross talk.

TABLE 90.2 Frequency Bands Allocated for General Telemetry Uses

Frequency Band MHz	Uses	Notes
72–76	Biotelemetry	Low-power devices; restricted by Part 15 of FCC rules
88–108	Educational	Four frequencies in this band; part 90 of FCC rules
154	Industry	Band in TV channels 7–13
174–216	Biotelemetry	Low-power operations restricted to hospitals
216–222	Multiple	BW < 200 kHz
450–470	General	Telemetry as secondary basis; limited to 2 W of RF
467	Industry	Business band; limited to 2 W of RF
458–468	Biotelemetry	Band in TV channels 21–29
512–566	Biotelemetry	Low-power operations restricted to hospitals
1427–1435	Fixed	Uses in land mobile services (telemetering and telecommand)
1435–1535	Aeronautical	
2200–2290	Mobile	

passband filters. Each channel is further demodulated, recovering the information from the transducer. The main practical problem of FDM systems is the cross talk between channels. Cross talk appears due to the nonlinearities of the electronic devices, which originates when the signal for one channel partially modulates another subcarrier in addition to the one assigned to that channel. Cross talk also originates when the spectra for two adjacent channels overlap. To avoid this effect, the subcarriers have to be chosen so that there is a separation (guard band) between the spectra of two contiguous channels. By increasing the guard band, the possibility of cross talk decreases, but the effective bandwidth also increases. The effective bandwidth equals the sum of the bandwidth of all channels, plus the sum of all the guard bands.

There are three alternative methods for each of the two modulation processes: the modulation of the measurement channel signals and the modulation of the composite signal. These methods are amplitude modulation (AM), frequency modulation (FM), and phase modulation (PM). The usual combinations are FM/FM, FM/PM, or AM/FM [6]. This chapter analyzes only the subcarrier modulation schemes, while the modulation for the RF signal is analyzed in *Electromagnetic, Optical, Radiation, Chemical, and Biomedical Measurement*.

90.2.2 Subcarrier Modulation Schemes for FDM

90.2.2.1 Subcarrier Modulation of Amplitude

In an AM subcarrier modulation scheme, the amplitude of a particular subcarrier signal is changed according to the value of the measured channel assigned to that frequency. The resulting AM signal is given by the expression

$$v(t) = A_c[1 + m(t)]\cos(\omega_c t) \tag{90.1}$$

where

A_c is the amplitude of the carrier
$m(t)$ the modulating signal
ω_c the frequency of the carrier

This type subcarrier modulation requires very simple transmitter and receiver circuits. The level of modulation is indicated by a parameter known as the percentage of modulation (P). For simple signals, P can be expressed as

$$P = m \times 100(\%) \tag{90.2}$$

In a more general way, the percentage of modulation (P) can be expressed as

$$P(\%) = \frac{A_{c(\max)} - A_{c(\min)}}{2A_c} \times 100 \tag{90.3}$$

where $A_{c(\max)}$ and $A_{c(\min)}$ are the maximum and minimum values for the carrier signal.

Figure 90.5 shows the spectrum of an amplitude-modulated signal, assuming that the modulating signal is a band-limited, nonperiodic signal of finite energy. The spectrum consists of two sidebands that are symmetrical in reference to the subcarrier. From Figure 90.5, it is also possible to see the main disadvantages of AM schemes. First, the bandwidth of the modulated channel is two times the bandwidth of the modulating signal, due to the two similar sidebands that appear. This results in an inefficient use of the spectrum. Second, the analysis of power for each of the components in Figure 90.5 shows that at least 50% of the transmitted power is used in transmitting the subcarrier, which is independent of the measured signal, as it does not contain any information. The remaining power is split between the two sidebands, which results in a maximum efficiency that it is theoretically possible to achieve of below 25%. The third main problem of AM is the possibility of overmodulation, which occurs when $m > 1$. Once a signal is overmodulated, it is not possible to recover the modulating signal with the simple circuits that are widely used for AM telemetry transmission.

The limitations of AM subcarrier modulation can be overcome using more efficient modulation techniques, such as double sideband (DSB), single sideband (SSB), and compatible single sideband (CSBB), which are also considered AM techniques. However, the complexity of these modulation systems and the cost associated with systems capable of recovering subcarrier signals modulated this way cause them not to be used in most commercial telemetry systems. Most of the available systems that based on AM subcarrier techniques use the traditional AM that has been described here, because its simplicity overcomes the possible problems of its use.

90.2.2.2 Subcarrier Modulation of Frequency

FM (or PM) is by far the most-used subcarrier modulation scheme in FDM telemetry systems. These angle modulations are inherently nonlinear, in contrast to AM. Angle modulation can be expressed as

$$v(t) = A\cos[\omega_c t + \varnothing(t)] \tag{90.4}$$

where $\phi(t)$ is the modulating signal, that is, the signal from the transducers after conditioning.

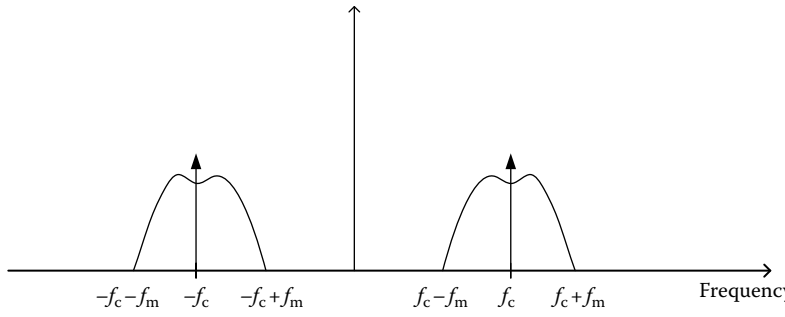


FIGURE 90.5 Resulting spectrum after AM. The spectrum of the baseband signal extends from 0 to f_m Hz. The resulting AM spectrum has doubled the required bandwidth, while only 0.25 of the total power is used in transmitting the desired information.

It is then possible to calculate the value of the instantaneous frequency as

$$f = \frac{1}{2\pi}\frac{d}{dt}[\omega_c t + \phi(t)] = \frac{\omega_c}{2\pi} + \frac{d}{dt}\phi(t) \tag{90.5}$$

This equation shows how the signal $v(t)$ is modulated in frequency. We can analyze two parameters that can be derived from the previous equations: frequency deviation and modulation index. Frequency deviation (f_m) is the maximum departure of the instantaneous frequency from the carrier frequency. The modulation index (β) is the maximum phase deviation. The following equations show how these parameters are related. The value of the instantaneous frequency (f) is [7]

$$f = \frac{\omega_c}{2\pi} + \frac{\beta\omega_m}{2\pi}\cos(\omega_m t) = f_c + \beta f_m \cos(\omega_m t) \tag{90.6}$$

The maximum frequency deviation is Δf and is given by

$$\Delta f = \beta f_m \tag{90.7}$$

Therefore, we can write the equation for the frequency modulated signal as

$$v(t) = A\cos\left[\omega_c t + \frac{\Delta f}{f_m}\sin(\omega_m t)\right] \tag{90.8}$$

The previous equation shows that the instantaneous frequency, f, lies in the range $f_c \pm \Delta f$. However, it does not mean that all the spectral components lie in this range. The spectrum of an angle-modulated waveform cannot be written as a simple equation. In the simplest case, when the modulating signal is a sinusoidal signal, a practical rule states that the bandwidth of an FM signal is twice the sum of the maximum frequency deviation and the modulating frequency. For modulating signals commonly found in measuring systems, the bandwidth is dependent upon the modulation index; that is, as the bandwidth allocated for each channel is limited, the modulation index will also be limited.

90.2.2.3 Frequency Division Multiplexing Telemetry Standards

The standards for telemetry systems using FDM are exclusively based on FM using two different approaches: proportional-bandwidth modulation (PBW) and constant-bandwidth modulation (CBW) as well as their combination.

In PBW, the bandwidth allocated to a channel is proportional to the subcarrier frequency. The standard allows three different subcarrier deviations: 7.5%, 15%, and 30%. Table 90.3 summarizes the most important parameters for the channels used by PBW.

TABLE 90.3 Main Characteristics of the Channels for PBM

Deviation	Number of Channels	Designation	Lowest Channel Center Frequency	Frequency Bandwidth Lowest Channel	Highest Channel Center Frequency	Frequency Bandwidth Highest Channel
7.5%	25	1 … 25	400 Hz	370–430 Hz	560 kHz	518–602 kHz
15%	12	A … L	22 kHz	18.7–25.3 kHz	560 kHz	476–644 kHz
30%	12	AA … LL	22 kHz	15.4–28.6 kHz	560 kHz	392–728 kHz

It is important to note that for the 7.5% deviation, the channels have been chosen so that the ratio between the upper deviation limit for a given channel and the lower deviation limit for the next channel is around 1.2. When using a deviation frequency of 15% (channels A through L), the upper deviation limit of one channel has a frequency higher than the lower deviation limit of the adjacent, higher channel. This results in the need to omit adjacent channels. A similar situation occurs when using a deviation frequency of 30% (channels AA through LL) although in this case, the highest deviation frequency is higher than the lowest deviation frequency of the two adjacent, higher channels, thus requiring the omission of these two channels.

CBW channels keep the bandwidth constant and independent of its carrier frequency. There are eight possible maximum subcarrier frequency deviations labeled A (for 2 kHz deviation) to H (for 256 kHz deviation). The deviation frequency doubles from one group to the next. There are 22 A-channels, whose center frequency range from 8 to 176 kHz. The separation between adjacent channels is a constant of 8 kHz. Table 90.4 shows a summary of the characteristics of CBW channels.

The standard also gives criteria for the use of the FDM standards. It focuses on the limits, most of the time dependent on the hardware used, and performance trade-offs such as data accuracy for data bandwidth that may be required in the implementation of the system. The subcarrier deviation ratio determines the SNR for a channel. As a rule of thumb, the SNR varies as the three-halves power of the subcarrier deviation ratio. On the other hand, the number of subcarrier channels that can be used simultaneously to modulate an RF carrier is limited by the channel bandwidth of the RF carrier as well as considerations of SNR. Given a limited RF bandwidth, as more channels are added to the FDM system, it is necessary to reduce the deviation ratio for each channel, which reduces the SNR for each channel. It is then very important to evaluate the acceptable trade-off between the number of subcarrier channels and the acceptable SNR values. A general equation that might be used to estimate the thermal noise performance of an FM/FM channel is the following [8]:

$$\left(\frac{S}{N}\right)_d = \left(\frac{S}{N}\right)_c \left(\frac{3}{4}\right)^{1/2} \left(\frac{B_c}{F_{ud}}\right)\left(\frac{f_{dc}}{f_s}\right)\left(\frac{f_{ds}}{F_{ud}}\right) \tag{90.9}$$

where

$(S/N)_d$ represents the SNR at the discriminator output
$(S/N)_c$ represents the SNR of the receiver
B_c is the intermediate-frequency (IF) bandwidth of the receiver
F_{ud} is the subcarrier discriminator output filter (at −3 dB)
f_s is the subcarrier center frequency
f_{dc} is the carrier peak deviation for the subcarrier considered
f_{ds} is the subcarrier peak deviation

TABLE 90.4 Parameters of CBW Channels in FDM Telemetry

Channel Denomination	Frequency Deviation, kHz	Lowest Channel Center Frequency, kHz	Highest Channel Center Frequency, kHz	No. of Channels	Separation between Channels, kHz
A	±2	8	176	22	8
B	±4	16	352	22	16
C	±8	32	704	22	32
D	±16	64	1408	22	64
E	±32	128	2816	22	128
F	±64	256	3840	15	256
G	±128	512	3584	7	512
H	±256	1024	3072	3	1024

According to the standard, the FM/FM composite FDM signal that is used to modulate an RF carrier can be of PBW format, CBW format, or a combination of both, with the only limitation that the guard bands between the channels used in the mixed format are equal or greater than the guard bands for the same channels in an unmixed format.

90.2.3 Time Division Multiplexing

TDM is a transmission technique that divides the time into different slots and assigns one slot to each measurement channel. In TDM, all the transmission bandwidth is assigned entirely to each of the measurement channels during a fraction of the time. After the signals from the measurement channels have been low-pass filtered, they are sequentially sampled by a digital switch that samples all the measurement channels in a period of time (T) that complies with the Nyquist criteria. Figure 90.6a shows a basic block diagram for an FDM system. The output of the sampler is a train of AM pulses that contains the individual samples for the channels framed periodically, as is shown in Figure 90.6b. Finally, the composite signal modulates an RF carrier. The set of samples from each one of the input channels is called a frame. For M measurement channels, the period between two consecutive pulses is $T_s/M = 1/Mf_s$, where T_s is the sampling period. The period between samples from the same channel is T_s. At the receiver end, by separating the digital signals into different channels by a synchronized demultipler and by low-pass filtering, it is possible to recover the original signal for each measurement channel.

TDM systems have advantages over FDM systems. First, FDM requires subcarrier modulators and demodulators for each channel, whereas in TDM only one multiplexer and one demultiplexer are required. Second, TDM signals are resistant to the error sources that originate cross talk in FDM: nonideal filtering and cross modulation due to nonlinearities. In TDM, the separation between channels depends on the sampling system. However, because it is impossible in practice to produce perfectly square pulses, their rise and fall times are different from zero. It is then necessary to provide guard time between pulses, similar to the band guards in FDM systems. Cross talk in TDM can be easily estimated assuming that the pulse decay is exponential with a time constant (τ) approximately equal to

$$\tau = \frac{1}{2\pi B} \tag{90.10}$$

where B is the −3 dB channel bandwidth. The cross talk (k) between channels can be approximated as

$$k = -54.5T_g \text{ (dB)} \tag{90.11}$$

where T_g is the minimum time separation between channels, called guard time.

A common situation in measurement systems occurs when the M signals that need to be measured have very different speeds. The channel sampling rate is determined by the fastest signal, thus needing an M-input multiplexer capable of handling signals at that sampling frequency. A convenient solution is to feed several slow signals into one multiplexer, then combine its output with the fast signal in a second multiplexer [6].

90.2.3.1 Subcarrier Modulation Schemes for Time Division Multiplexing

In this process, the signal that results after the multiplexing and sampling process modulates a train of pulses. The most common methods for subcarrier modulation are pulse amplitude modulation (PAM), pulse duration modulation (PDM), and pulse position modulation (PPM). Figure 90.7 illustrates these three modulation schemes, where the pulses are shown square for simplicity. The parameter that is modulated (amplitude, duration, or relative position) changes proportionally to the amplitude of the sampled signal. However, in PAM and PDM, the values have an offset, so that when the value of the sample

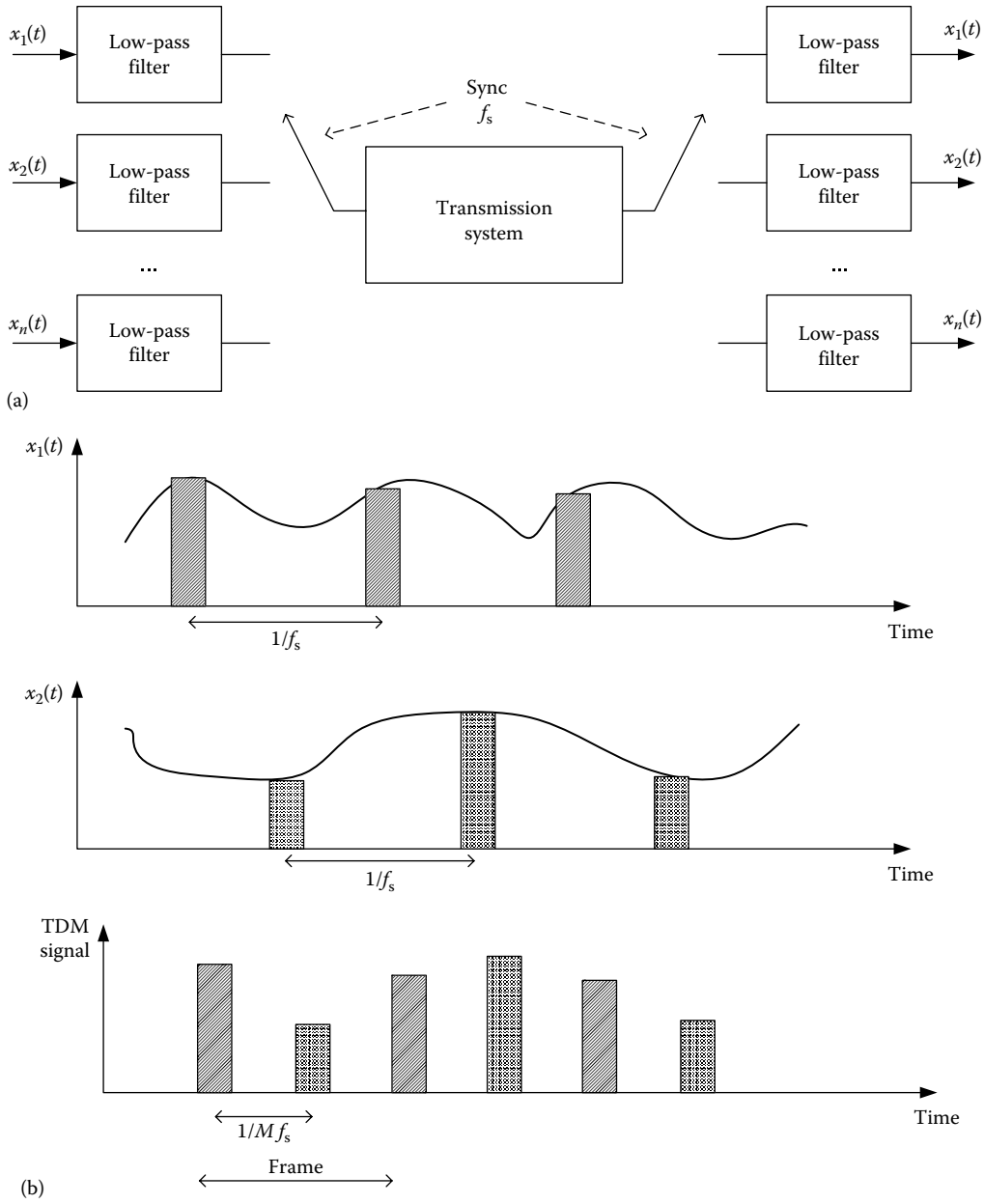


FIGURE 90.6 (a) Block diagram of a TDM system based sequentially sampling M different channels at a sampling frequency f_s and sending the information for each channel sequentially. (b) Time representation of the TDM system.

is zero, the pulse amplitude or the pulse width is different from zero. The reason for these offsets is to maintain the rate of the train of pulses constant, which is very important for synchronization purposes. The common characteristics of the different analog modulation schemes for pulses in TDM are (1) a modulated signal spectrum with a large low-frequency content, especially close to the sampling frequency; (2) the need to avoid overlaying between consecutive pulses in order to conserve the modulation parameters; and (3) the possibility of reconstructing the original samples from the modulated signal through low-pass filtering after demultiplexing. The reduction of noise depends on the bandwidth of the modulated signal, with this being the principal design criterion.

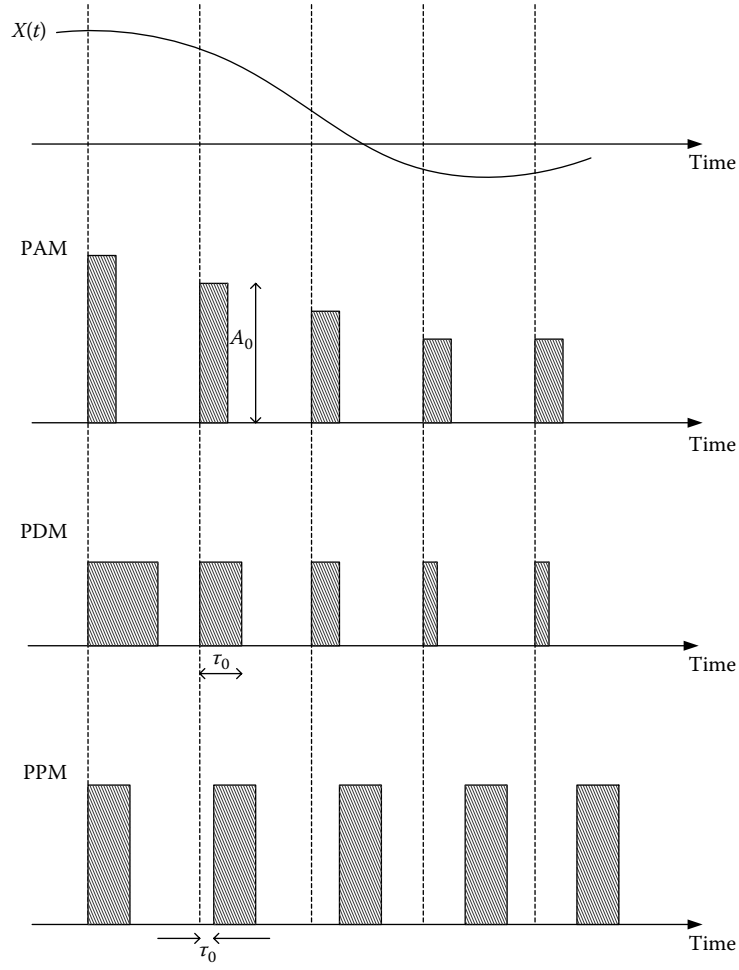


FIGURE 90.7 Different modulation schemes used in TDM. The variations in amplitude of the signal $x(t)$ are transmitted as amplitude variations of pulses (PAM), duration changes of pulses (PDM), or changes in the relative position of the pulses (PPM). In all the cases, level 0 is transmitted by a pulse whose amplitude (A_0), duration (τ_0), or relative position (τ_0) is different from 0.

90.2.3.1.1 Pulse Amplitude Modulation

PAM waveforms are made of unipolar, nonrectangular pulses whose amplitudes are proportional to the values of the samples. It is possible to define the modulation index using similar criteria as in analog AM. Similarly, in PAM the modulation index is limited to values less than 1.

90.2.3.1.2 Pulse Duration Modulation

PDM is made of unipolar, rectangular pulses whose durations or widths depend on the values of the samples. The period between the center of two consecutive pulses is constant. The analysis of the resulting spectrum shows that it is possible to reconstruct the samples by low-pass filtering [9].

90.2.3.1.3 Pulse Position Modulation

PPM is closely related to PDM, as PPM can be generated through PDM. In PPM the information resides on the time location of the pulses rather than in the pulses by themselves. It is then possible to transmit very narrow pulses to reduce the energy needed; this energy reduction is the most important advantage of PPM.

90.2.4 Pulse Code Modulation in Time Division Multiplexing

All the previously analyzed subcarrier modulation schemes in telemetry systems are based on an analog signal that modulates either an analog carrier or a train of pulses. Pulse code modulation (PCM) is different: it is a digital modulation in which the measured signal is represented by a group of codified digital pulses. Two variations of PCM that are also often used are delta modulation (DM) and differential pulse code modulation (DPCM). In analog modulation schemes, the modulating signal from the transducer can take any value between the limits. If noise alters the modulating signal, it is impossible to decide its real value. Instead, if not all the values in the modulating signal are allowed and the separation between the allowed levels is higher than the expected noise values, it is then possible to decide which were the values sent by the transmitter. This immunity against noise makes PCM systems one of the preferred alternatives for telemetry. Figure 90.8 shows the basic elements of a PCM telemetry system. A PCM encoder (or PCM commutator) converts the input data into a serial data format suitable for transmission through lines by wireless techniques. At the receiving end, a PCM decoder (or PCM decommutator) converts the serial data back into individual output data signals. PCM systems transmit data as a serial stream of digital words. The PCM encoder samples the input data and inserts the data words into a PCM frame. Words are assigned specific locations in the PCM frame, so the decoder can recover the data samples corresponding to each input signal. The simplest PCM frame consists of a frame synchronization word followed by a string of data words. The frame repeats continually to provide new data samples as the input data change. Frame synchronization enables the PCM decoder to locate the start of each frame easily.

The IRIG standard also defines the characteristics of PCM transmission for telemetry purposes; in particular, the pulse train structure and system design characteristics. The PCM formats are divided into two classes for standards purposes: class I and class II. The simpler types are class I, whereas the more complex types are class II. Some of the characteristics of class II systems are bit rates greater than 5 Mbit/s, word lengths in excess of 16 bits, fragmented words, unevenly spaced subcommutation, format changes, tagged data formats, asynchronous data transmission, and merger of multiple format types, among others. Table 90.5 provides a brief summary of relevant PCM specifications.

The following PCM codes, shown in Figure 90.9, are recognized by the IRIG standards: NRZ-L (nonreturn-to-zero level), NRZ-M (nonreturn-to-zero mark), NRZ-S (nonreturn-to-zero space), BiØ-L (biphase level), BiØ-M (biphase mark) and BiØ -S (biphase space). The standard also recommends that the transmitted bit stream be continuous and contain sufficient transitions to ensure bit acquisition and continued bit synchronization. Bit rates should be at least 10 bits/s. If the bit rate is above 5 Mbit/s, the PCM system is classified as class II. In reference to the word formats, the standard defines a fixed format as one that does not change during transmissions with regard to the frame structure, word length or location, commutation sequence, sample interval, or measurement list. Individual words may vary in length from 4 bits to not more than 16 bits in class I and not more than 64 bits in class II. Fragmented words, defined as a word divided into not more than eight segments and placed in various locations within a minor frame, are only allowed in class II. All word segments used to form a data word are constrained to the boundaries of a single minor frame.

The frame structure allowed by the standards for PCM telemetry specifies that data are formatted into fixed frame lengths that contain a fixed number of equal-duration bit intervals. A minor frame is defined as the data structure in time sequence from the beginning of a minor frame synchronization

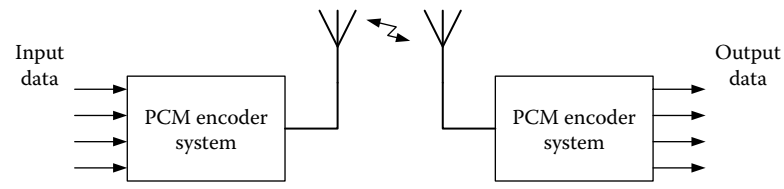


FIGURE 90.8 Block diagram showing a basic PCM link for telemetry.

TABLE 90.5 Summary of Relevant PCM Specifications

Specification	Class I	Class II
Class format support	Class I (simple formats) supported on all ranges	Class II (complex formats) requires concurrence of range involved
Primary bit representation (PCM codes)	NRZ-L, NRZ-M, NRZ-S, RNRZ-L, BiØ-L, BiØ-M, BiØ-S	Same as class II
Bit rate	10 bps–5 Mbps	10 bps to >5 Mbps
Bit rate accuracy and Stability	0.1%	Same as class I
Bit jitter	0.1 bit	Same as class I
Bit numbering	MSB = bit number 1	Same as class I
Word length	4–16 bits	4–64 bits
Fragmented words	Not allowed	Up to 8 segments each; all segments in the same minor frame
Minor frame length	<8192 bits or <1024 words (includes synchro)	<16,384 bits (includes synchro)
Major frame length	<256 minor frames	Same as class I
Minor frame numbering	First minor frames in each major frames number 1	Same as class I
Format change	Not allowed	Frame structure is specified by frame format identification (FFI) word in every minor frame

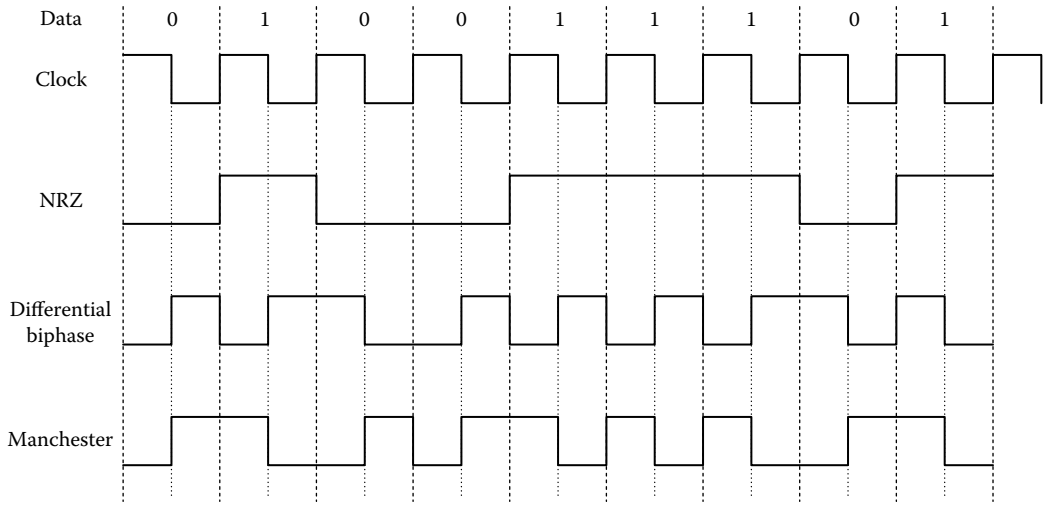


FIGURE 90.9 IRIG standards. *NRZ* (nonreturn to zero): "1" represented by a high logic level and "0" represented by a low logic level. *Differential biphase*: "1" represented by a change in level at the start of the clock period and "0" represented by no change of level at the start of the clock period. There is always a change of level at the middle of the clock period. *Manchester*: "1" represented by a high-to-low transition at the middle of the clock period and "0" represented by a low-to-high transition at the middle of the clock period.

pattern to the beginning of the next minor frame synchronization pattern. The minor frame length is the number of bit intervals from the beginning of the frame synchronization pattern to the beginning of the next synchronization pattern. The maximum length of a minor frame will not exceed 8,192 bits or 1,024 words in class I and will not exceed 16,384 bits in class II. Minor frames consist of the synchronization pattern, data words, and subframe synchronization words if they are used. The standard allows the use of words of different length if they are multiplexed in a single minor frame. Figure 90.10 shows a graphical representation of a PCM frame structure. Major frames contain the number of minor frames

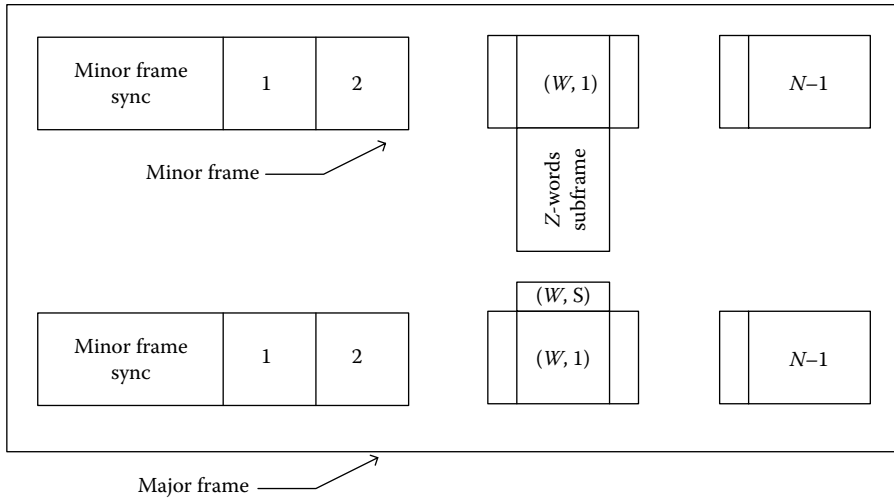


FIGURE 90.10 Structure of a PCM frame. The maximum length of a minor frame is 8192 bits or 512 for class I and 16,284 bits for class II. A major frame contains $N \times Z$ words, where Z is the number of words in the maximum subframe and N is the number of words in the minor frame. Regardless of its length, the minor frame synchronism is considered as one word. W is the word position in the minor frame, while S is the word position in the subframe.

required to include one sample of every parameter in the format. Their length is defined as minor frame length multiplied by the number of minor frames contained in the major frame. The maximum number of minor frames per major frame is limited to 256.

The IRIG standard also gives recommendations for maximal transmission efficiency in PCM telemetry. The IF bandwidth for PCM telemetry data receivers should be selected so that 90%–99% of the transmitted power spectrum is within the receiver 3 dB bandwidth. The IF also has effects on the bit error probability (BEP) according to the following equation for NRZ-L PCM/FM [10]:

$$\text{BEP} = 0.5 * e^{k\,SNR} \tag{90.12}$$

where

$k \approx -0.7$ for IF bandwidth equal to bit rate
$k \approx -0.65$ for IF bandwidth equal to 1.2 times bit rate
$k \approx -0.55$ for IF bandwidth equal to 1.5 times bit rate

Figure 90.11 shows a graphical representation of the bit error rate versus the SNR for these three values of k.

Other data codes and modulation techniques have different BEP versus SNR performance characteristics, but in any case, they will have similar trends.

The standard also specifies the recommended frame synchronization patterns for general use in PCM telemetry. There are different lengths for synchronization patterns, but in all of them, the 111 is the first bit sequence transmitted. The patterns for lengths 16–30 were selected in order to minimize the probability of false synchronization over the entire pattern overlap portion of the ground station frame synchronization [9]. The spectral density (S) for the NRZ and BiØ codes are

$$\text{NRZ Codes} \quad S = \frac{\sin^2(\pi f T)}{(\pi f T)^2} \tag{90.13}$$

$$\text{Biphase Codes} \quad S = \frac{\sin^4(\pi f T/2)}{(\pi f T/2)^4} \tag{90.14}$$

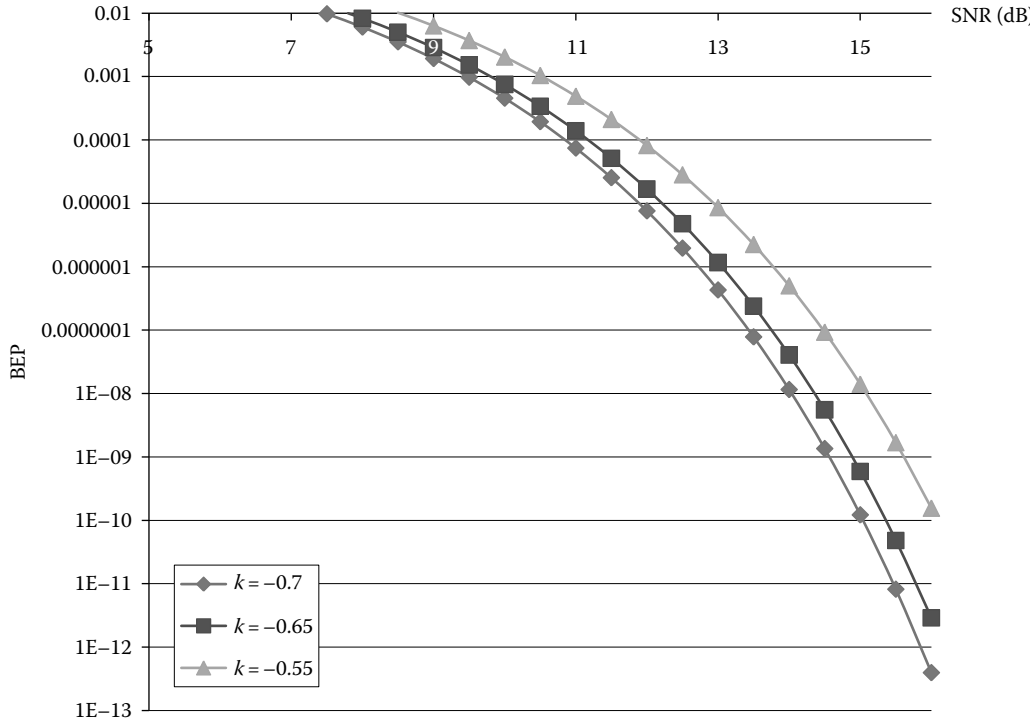


FIGURE 90.11 Bit error rate versus SNR for three values of the parameter k.

The calculation of spectral densities allows the determination of the BEP for the previous type of codes assuming perfect bit synchronization. These calculations show that for the same SNR, the lowest BEP is achieved for NRZ-L and Bi codes, followed by NRZ and BiØ mark and space codes and finally for random NRZ-L codes (RNRZ-L).

Defining Terms

Bandwidth: The range of frequencies occupied by a signal.

Carrier: A frequency that is modulated by a signal containing information.

Channel: A subcarrier that carries information.

Constant-bandwidth (CBW) channel: A channel whose bandwidth is independent of its carrier frequency.

Deviation ratio: The ratio of the maximum carrier frequency deviation to the maximum data frequency deviation.

Frequency deviation: The difference between the center frequency of a carrier and its upper or lower deviation limit.

Frequency division multiplexing (FDM): A composite signal consisting of a group of subcarriers arranged so that their frequencies do not overlap or interfere with each other.

Frequency response: The highest data frequency that can be carried by the channel.

IRIG: Inter-Range Instrumentation Group of the Range Commanders Council (RCC).

Proportional-bandwidth (PBW) channel: A channel whose bandwidth is proportional to its carrier frequency.

Remote switching: Telemetry consisting only of yes/no or on/off orders.

Signaling: Telemetry consisting of binary information.

Subcarrier: A carrier combined with other carriers to create a composite signal.

Subcarrier bandwidth: The difference between the upper and lower frequencies of a modulated carrier.

References

1. B.P. Dagarin, R.K. Taenaka, and E.J. Stofel, Galileo probe battery system, *IEEE Aerospace Electron. Syst. Mag.*, 11 (6): 6–13, 1996.
2. M.W. Pollack, Communications-based signaling: advanced capability for mainline railroads, *IEEE Aerospace Electron. Syst. Mag.*, 11 (11): 13–18, 1996.
3. M.C. Shults, R.K. Rhodes, S.J. Updike, B.J. Gilligan, and W.N. Reining, A telemetry-instrumentation system for monitoring multiple subcutaneously implanted glucose sensors, *IEEE Trans. Biomed. Eng*. 41: 937–942, 1994.
4. M. Rezazadeh and N.E. Evans, Multichannel physiological monitor plus simultaneous full-duplex speech channel using a dial-up telephone line, *IEEE Trans. Biomed. Eng.*, 37: 428–432, 1990.
5. Y.S. Trisno, P. Hsieh, and D. Wobschall, Optical pulses powered signal telemetry system for sensor network application, *IEEE Trans. Instrum. Meas.*, 39: 225–229, 1990.
6. R. Pallás-Areny and J.G. Webster, *Sensors and Signal Conditioning*, 2nd edn., New York: John Wiley & Sons, 486–489, 2001.
7. H. Taub and D.L. Schilling, *Principles of Communication Systems*, 2nd edn., New York: McGraw-Hill, 1986.
8. K.M. Uglow, Noise and bandwidth in FM/FM radio telemetry, *IRE Trans. Telemetry Remote Control*, pp. 19–22, 1957.
9. J.L. Maury and J. Styles, Development of optimum frame synchronization codes for Goddard space flight center PCM telemetry standards, *Proceedings of the National Telemetering Conference*, Houston, TX, 1964.

Relevant Websites about Telemetry

http://www.irig106.org/: IRIG 106 is a comprehensive telemetry standard to ensure interoperability in aeronautical telemetry applications.

http://irig.org/: IRIG ORG is an informational website that is available to all who are interested in quick and easy access to telemetry or IRIG information.

http://www.irig106.org/docs/123-09/IRIG_106-07_Chapter_10_Programming_Handbook.pdf: *IRIG Standard Programming Handbook*.

http://www.telemetry.org: International Telemetry Conference held annually in the United States.

http://www.telemetry-europe.org/: European Society of Telemetry.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

X

Control and Human Factors

91 **PID Control** *F. Greg Shinskey* **91**-1
Introduction • Open and Closed Loops • Mode Selection • Controller Hardware • Tuning Controllers • Defining Terms • References • Further Readings • Partial List of Suppliers

92 **Optimal Control and the Software** *Halit Eren* **92**-1
Introduction • Cost Functions • Calculus of Variations • Riccati Equation • State-Feedback Matrix • Software Support and Applications • Bibliography • Partial List of Providers

93 **Electropneumatic and Electrohydraulic Instruments: Modeling** *M. Pachter and C.H. Houpis* **93**-1
Introduction • Background • Hydraulic Actuator Modeling • EHA Actuator Modeling • Actuator Compensator Design • Conclusion • Nomenclature • References • Partial List of Vendors and Suppliers

94 **Explosion-Proof Instruments** *Sam S. Khalilieh* **94**-1
Introduction • Fundamentals of Explosion Protection • Classification of Hazardous Areas • Enclosure Types and Requirements • Protection Methodologies • Defining Terms • References • Further Readings • Partial List of Manufacturers and Suppliers

95 **Measurement and Identification of DC Brush and Brushless Stepping Motors** *Stuart Schweid, Robert Lofthus, and John McInroy* **95**-1
Introduction • Hybrid Stepping Motors • DC Brush and Brushless Motors • References • Further Readings • Partial List of Manufacturers and Suppliers

96 **Human Factors in Displays** *Jeffrey D. Onken, Barrett S. Caldwell, and Steven A. Murray* **96**-1
Introduction • Fundamentals of Light Measurement • Fundamentals of Vision • Visual Performance • Display Performance Considerations • Display Content Considerations • Cathode-Ray Tube Displays • Liquid Crystal Displays • Plasma Displays • Electroluminescent Displays • Light-Emitting Diode Displays • Electronic Paper/Electronic Ink • Defining Terms • References • Further Information



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

91

PID Control

91.1 Introduction **91**-1
91.2 Open and Closed Loops **91**-1
Open-Loop Responses • Closed-Loop Responses
91.3 Mode Selection **91**-4
Proportional Control • Integral Control • PI Control • PD Control • PID Controllers
91.4 Controller Hardware **91**-6
Pneumatic Controllers • Electronic Controllers • Digital Controllers
91.5 Tuning Controllers **91**-7
Manual Tuning • Autotuning • Self-Tuning
Defining Terms **91**-8
References **91**-9
Further Readings **91**-9
Partial List of Suppliers **91**-9

F. Greg Shinskey
Process Control Consultant

91.1 Introduction

Process control plays an essential role in the safe manufacture of quality products at market demand while protecting the environment. Flow rates, pressures, and temperatures within pipes and vessels, inventories of liquids and solids, and product quality are all examples of measured variables that must be controlled to meet the above objectives. While there are several means available for controlling these variables, the proportional–integral–derivative (PID) family of controllers has historically carried the major share of this responsibility and, because of their simplicity and reliability, will continue to do so in the future.

The acronym PID stands for the three principal modes of the controller, each of which bears a mathematical relationship between the controlled variable c and the manipulated variable m driven by the controller output. The proportional mode relates changes in m to changes in c through a proportional gain. The integral mode moves the output at a rate related to the deviation of c from its desired value, known as the set point or reference r. Finally, the derivative mode moves the output in response to the time derivative or rate of change of c. Interestingly, the mathematical relationships were actually determined *after* controllers had already been created to solve process-control problems. The integral mode was initially called *automatic reset*, and the derivative mode *hyper-reset* or *pre-act* [1].

91.2 Open and Closed Loops

Figure 91.1 describes the process and controller in functional blocks connected in a loop. Inputs to the process are manipulated m and load q variables, usually rates of flow of material or energy into or out of the process. The load may be a single variable or an aggregate, either independent or

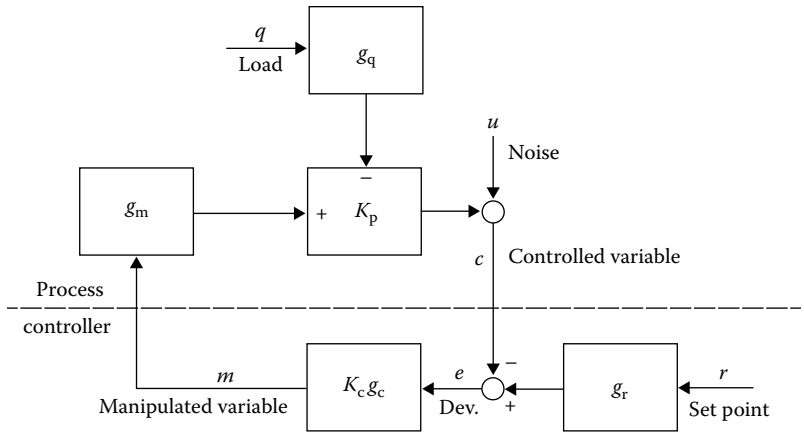


FIGURE 91.1 Process and controller connected in a loop.

manipulated or controlled by another controller. If independent, it is often unmeasured, with its value inferred by the level of controller output required to maintain the controlled variable at set point. Noise u is shown affecting the controlled variable directly, typically caused by local turbulence in flow, pressure, and liquid-level measurements or by nonuniformity of streams whose composition is measured.

91.2.1 Open-Loop Responses

In the absence of automatic control, the controlled variable is subject to variations in the load and to manual adjustments to m intermittently introduced by operators. These cause variations in c following both the steady-state and dynamic functions appearing in Figure 91.1:

$$c = K_p\left(mg_m - qg_q\right) + u \tag{91.1}$$

where

K_p is the steady-state gain
g_m and g_q are the dynamic-gain vectors in the manipulated and load paths, respectively
u is the noise level

The vectors have both magnitude and phase angle that are functions of the frequency or period of the signal passing through. In the *open loop*, variations in m are likely to be steps introduced by operators, but variations in q could take any form—steps, ramps, random variations, or cycles—depending on the source of the disturbance. Steps are easily introduced manually and contain all the frequencies from zero to infinity; consequently, they are useful for evaluating loop response and testing the effectiveness of controllers. Noise is typified by random variations in a frequency range above the bandwidth of the control loop.

The dynamic elements typically consist of dead time and lags. However, liquid level is the *integral* of the difference between inflow and outflow [2], in which case K_p in Figure 91.1 is replaced by an integrator. Any difference between inflow and outflow will then be integrated, causing the liquid level to continue rising or falling until the vessel limit is reached. This process has no self-regulation and therefore cannot be left indefinitely in an open loop.

91.2.2 Closed-Loop Responses

In the *closed loop*, c responds to load and set point as follows:

$$c = \frac{rg_r}{1 + 1/K_p g_m K_c g_c} - \frac{qg_q}{g_m K_c g_c + 1/K_p} + u \tag{91.2}$$

where

K_c is the proportional gain of the controller
g_c is the dynamic gain of its integral and derivative modes
g_r that of its set-point filter

For the typical *self-regulating* process in the open loop, c will take an exponential path following a step change in load as shown by the dashed curve in Figure 91.2. If the loop is closed, the controller can move m to return c to r along the solid curve in Figure 91.2, an optimum trajectory having a minimum integrated absolute error (IAE) between c and r. The leading edge of the curve depends on g_q and the trailing edge depends on g_m and the PID settings. If the PID settings are optimized for load response, they will usually cause c to overshoot a change in set point r, as shown by the dashed curve in Figure 91.3. If the PID settings are then readjusted to reduce set-point overshoot, load response is extended. A preferred solution is to insert a lead–lag filter in the set-point path, which produced the solid curve in Figure 91.3.

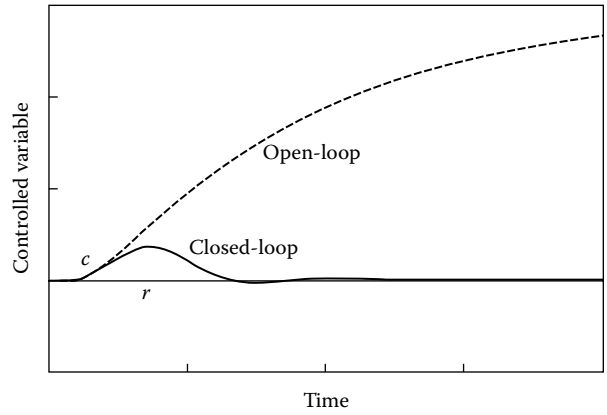


FIGURE 91.2 Closed-loop control minimizes error.

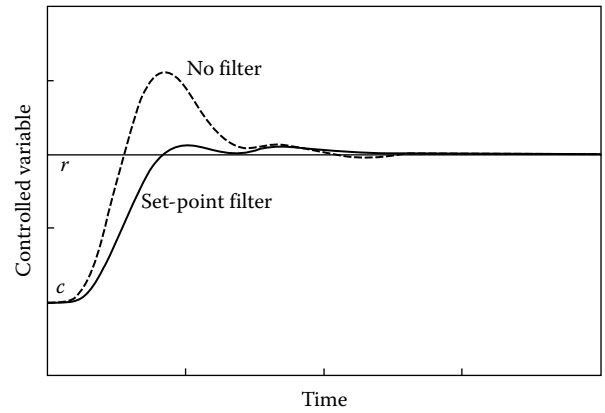


FIGURE 91.3 A lead–lag filter reduces overshoot.

91.3 Mode Selection

The family of PID controllers includes proportional (P), integral (I), proportional–integral (PI), proportional–derivative (PD), and two PID controllers. Each has its own advantages and limitations and therefore its preferred range of applications. Each is outlined along with an application.

91.3.1 Proportional Control

In a *proportional* controller, the deviation between c and r must change for the output to change:

$$m = b \pm K_c(r - c) \tag{91.3}$$

where b is an adjustable bias, also known as "manual reset," and the sign of the deviation is selected to provide negative feedback. Some controllers have proportional gain adjustable as percent proportional band P, where $K_c = 100/P$. If m must move to a different steady-state value because of a change in load or set point, the deviation will also change. Therefore, the proportional controller allows a steady-state offset between c and r whenever m does not equal b, which can only be eliminated by manually resetting b. Proportional control is recommended for liquid level, where K_c can be set quite high without loss of stability and there is no economic penalty for offset.

Proportional control of liquid level is actually preferred when manipulating the flow leaving a tank as the feed to a critical downstream process. Setting K_c slightly above 1.0 will keep the level in the tank for all rates of inflow while minimizing the rate of change of outflow. This application is called averaging level control [2].

91.3.2 Integral Control

The *integral* mode eliminates offset by driving the output at a rate proportional to the deviation:

$$\frac{dm}{dt} = C \pm \frac{r - c}{I} \tag{91.4}$$

where

t is time
d is the differential operator
I is the integral time setting
C is the constant of integration, i.e., the initial value of the controller output when placed on automatic

Some controllers have integral time adjusted as integral gain or "reset rate" $1/I$ in inverse time as "repeats per minute." Integration produces a phase lag of 90° between input and output, which increases the response time and period of oscillation of the loop. The integral controller *cannot be* used for liquid level, because two integrators in series form an unstable closed loop [2]. Its use is limited to optimizing the set points of other closed loops that are already stable [2].

91.3.3 PI Control

The PI controller combines the proportional and integral modes:

$$m = C \pm K_c\left[r - c + \frac{1}{I}\int (r - c)dt\right] \tag{91.5}$$

The deviation and its integral are added vectorially, producing a phase lag falling between 0° and 90°. This combination provides stability with elimination of offset, making it the most common controller used in the fluid-processing industries. It is used almost universally, even in those applications where other controllers are better suited.

91.3.4 PD Control

The addition of derivative action to a proportional controller adds response for the rate of change of the controlled variable:

$$m = b \pm K_c\left(r - c - D\frac{dc}{dt}\right) \tag{91.6}$$

where D is the derivative time setting. Derivative action is preferably applied only to c as indicated and not to the set point, which would only increase set-point overshoot. A pure derivative function has a dynamic gain increasing indefinitely with frequency—to avoid instability within the controller, it is usually limited to about 10 by filtering. This provides an optimum combination of responsiveness and noise rejection. Still, the high-frequency gain amplifies noise in flow and liquid-level measurements enough to prevent the use of the derivative mode on those loops; temperature measurements are usually noise-free, allowing it to be used to advantage there. The phase lead provided by the derivative mode reduces the period of oscillation and settling time of a loop and improves stability.

PD controllers are recommended for batch processes, where operation begins with c away from r and ends ideally with $c = r$ and flows at zero. The PD controller must be biased for this final output state, with D adjusted to eliminate overshoot, which for a zero-load process is permanent [2].

91.3.5 PID Controllers

The phase lead of derivative action more than offsets the phase lag of integral in a properly adjusted PID controller, resulting in a net phase lead and typically half the IAE of a PI controller applied to the same process. There are two types of PID controllers in common use, which differ in the way the modes are combined [3]. Early controllers combined PD and PI action in series, multiplying those functions rather than adding them. These interacting PID controllers remain in common use, even implemented digitally. Noninteracting controllers combine the integral and derivative modes in parallel, producing a more mathematically pure PID expression:

$$m = C \pm K_c\left[r - c - \frac{1}{I}\int(r - c)dt - D\frac{dc}{dt}\right] \tag{91.7}$$

The interacting controller can be described in the same form, but the coefficients of the individual terms are different [3]. The effective integral time of the interacting controller is $I + D$, its effective derivative time is $1/(1/I + 1/D)$, and its proportional gain is augmented by $1 + D/I$. Thus, the principal result of mode interaction is to require different PID settings for the two controllers applied to the same process.

A block diagram of a commercial interacting PID controller appears in Figure 91.4 with transfer functions described in Laplace transforms whose operator s is equivalent to the differential operator d/dt. Integration is achieved by positive feedback of the output through a first-order lag set at the integral time. The feedback signal f is taken downstream of the high and low limits and may be replaced with a constant or other variable to stop integration, in which case, the controller behaves as PD with remote bias. This feature is extremely valuable in preventing reset windup, which occurs whenever a controller with integral action remains in automatic while the loop is open, and results in a large overshoot when the loop is then closed [3]. Most noninteracting controllers lack this feature.

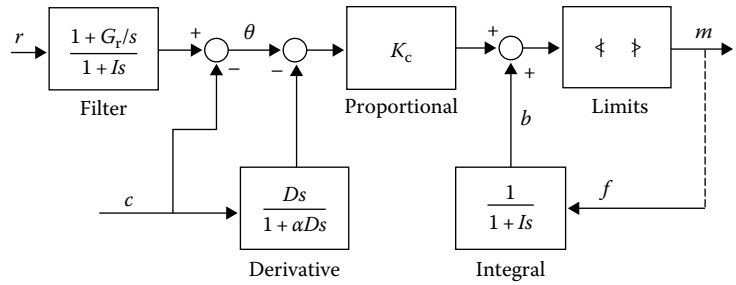


FIGURE 91.4 Commercial interacting PID controller.

Derivative action combines a differentiator Ds with a lag (filter) of time constant αD, where α represents the inverse of the high-frequency gain limit of typically 10. Figure 91.4 also shows a lead–lag set-point filter having a lag time constant of integral time and lead time of $G_r I$, which produces a gain of G_r to a step in set point. This gain is adjustable from 0 to 1, with 1 eliminating the filter and zero imposing a first-order lag. The effect of this lag is equivalent to eliminating proportional action from r in Equations 91.5 and 91.7, an optional feature in some PID controllers. The gain adjustment offers more flexibility in optimizing set-point response as in Figure 91.3; G_r is set around 0.3 for lag-dominant processes such as those described in Figures 91.2 and 91.3 and closer to 1 for dead time–dominant processes [3].

91.4 Controller Hardware

Pneumatic proportional controllers were first used early in this century, with integral and then derivative functions added around 1930. Pneumatic controllers are still in use in hazardous areas, in remote gas fields where they are operated by gas, and for simple tasks such as regulating temperatures in buildings. Electronic analog controllers began to replace pneumatic controllers in the late 1950s, but their functions were implemented using digital microprocessors beginning around 1980. Digital control initially began in mainframes in the 1960s but now is available in programmable logic controllers (PLC), personal computers (PCs), and in distributed control systems (DCS).

91.4.1 Pneumatic Controllers

The simplest pneumatic controllers are proportional units used to control heating, ventilating, and air-conditioning (HVAC) in buildings. However, complete PID controllers with functionality similar to that shown in Figure 91.4 are also available, used for such demanding tasks as temperature control of batch exothermic reactors and regulation of flows and pressures on offshore oil platforms. There are panel-mounted units available as well as weatherproof models used for field installation. Most have automanual transfer stations with bumpless transfer between the two modes of operation. Their principal limitations are a speed of response limited by transmission lines (lengths beyond 30 m are unsuitable for flow and pressure control) and lack of computing capability (required for nonlinear characterization and automatic tuning).

91.4.2 Electronic Controllers

The first electronic controllers mimicked their pneumatic counterparts while eliminating transmission lags. Eventually remote automanual transfer and remote tuning were added, and microprocessors brought these controllers calculation and logic functions, signal conditioning and characterization, and autotuning and self-tuning features as well. Multiple controllers are even available in a single station for implementing cascade and feedforward systems. The bandwidth of electronic *analog* controllers extends

to 10 Hz and in some units even further. *Digital* controllers execute their calculations intermittently rather than continuously; most are limited to 10 Hz operation, which with digital filtering results in an effective dead time of 0.1 s, reducing bandwidth to about 1.3 Hz.

Electronic controllers are used extensively in field locations and dedicated to control of individual units such as pumps, dryers, wastewater-treatment facilities—wherever only a few loops are required. Peer-to-peer communication is available in some controllers for incorporation into networks, and most can communicate with PC where data can be displayed on a graphic interface. Configuration (selection of scales, control modes, alarms, and other functions) can be done either via keys on the controller faceplate or through a PC.

91.4.3 Digital Controllers

While the stand-alone digital controller evolved from electronic analog models, centralized digital controllers originated with digital computers. Mainframe computers were first used for direct digital control (DDC) where valves were manipulated by the computer, and for supervisory control, where the computer positioned set points of analog controllers. Gradually minicomputers and microprocessors replaced the mainframe, with functions becoming *distributed* among field input–output modules, workstations, control modules, etc., in clusters and other configurations such as DCS. PCs are used for some workstations and even for direct control in some plants.

Where many loops are controlled by a single processor, the interval between samples is likely to be 0.5 s or even longer. Users tend to keep expanding the functions demanded of the processor, thereby extending the interval between samples. This reduces the bandwidth of DCS controllers to values too low for combustion and compressor controls. Some digital PID algorithms produce an *incremental* output Δm, requiring an integrator downstream to produce m. These are not available in proportional or PD action because they have no fixed bias b, but only a constant of integration C that is subject to change whenever the controller is operated manually.

The PLC was originally a replacement for relay logic. Eventually PID loops were added, which generally operate much faster (e.g., 100 Hz) than other digital controllers. However, many have nonstandard algorithms and lack the functionality of other PID controllers, such as proportional and PD control, windup protection, derivative filtering, set-point filtering, nonlinear characterization, and autotuning.

91.5 Tuning Controllers

A controller is only as effective as its *tuning*: the adjustment of the PID settings relative to the process parameters to optimize load and set-point response as described in Figures 91.2 and 91.3. Tuning is required when the controller is first commissioned on a loop and may have to be repeated if process parameters change appreciably with time, load, set point, etc. Tuning requires the introduction of a disturbance in either the open or closed loop and interpretation of the resulting response. *Autotuning* essentially automates manual procedures, whereas *self-tuning* can recognize and correct an unsatisfactory response without testing.

91.5.1 Manual Tuning

Ziegler and Nichols [1] developed the first effective tuning methods, and these are still used today. Their open-loop method steps the controller output to produce a response like the broken curve in Figure 91.2. The apparent dead time (delay) in the response and the steepest slope of the curve are then converted into appropriate settings for PI and PID controllers using simple formulas. The open-loop method is most accurate for lag-dominant processes, but the closed-loop method is more broadly applicable. It is based on inducing a uniform cycle by adjusting the gain of a controller with only its

proportional mode effective (D is set to 0 and I at maximum). The period of the cycle and the controller gain are then used to calculate appropriate PID settings.

The open-loop method has been extended to other processes [3] but remains limited to the accuracy of step-test results. Fine-tuning must be done with the loop closed: With the deviation zero, place the controller in manual, step the output, and immediately transfer to automatic—this simulates a closed-loop load change. The resulting response should appear like the solid curve in Figure 91.2, having a first peak, which is symmetrical, followed by a slight overshoot and well-damped cycle. If the overshoot is excessive, I and/or D time should be increased; in the case of undershoot, they should be decreased. (To simplify tuning of PID controllers, I and D can be moved together, keeping the I/D ratio at about 4 for noninteracting controllers and 3 for interacting controllers.) If damping is light, lower the proportional gain; if recovery is slow (producing an unsymmetrical first peak), raise the proportional gain. The set point should *not* be stepped for test purposes unless set-point response is important, as for flow controllers.

91.5.2 Autotuning

Some autotuning controllers use a step test in the open loop, implementing Ziegler–Nichols rules. A single pulse produces more accurate identification, and a doublet pulse is better still [3]. A closed-loop method replaces proportional cycling with relay cycling, where the output switches between high and low limits as c crosses r [3]. The period of the resulting cycle is used to set I and D and its amplitude relative to the distance between output limits to set K_c. The principal limitation of the step and relay autotuning methods is that only two response features are used to identify a complex process. The estimated PID settings may not fit the specific process particularly well, resulting in instability in some loops and sluggishness in others. The autotuning function cannot monitor the effectiveness of its work, as self-tuning does.

91.5.3 Self-Tuning

A self-tuning controller need not test the process, but simply observe its closed-loop response to set-point and load changes with its current PID settings. It then performs fine-tuning as described earlier to bring the overshoot or damping or symmetry of the response curve closer to optimum. This may require several iterations if the PID settings are far from optimum, but can eventually converge on optimum response, and readjust as necessary whenever process parameters change. Without a test disturbance, mischaracterization is possible, especially for a sinusoidal disturbance, where detuning may result when tightening would give better control.

Defining Terms

Averaging level control: Allowing the liquid level in a tank to vary in an effort to minimize changes in its outflow.

Closed-loop response: The response of a controlled variable to changes in set point or load with the controller in the automatic mode.

Derivative action: The change in controller output responding to a rate of change in the controlled variable.

Dynamic gain: The ratio of the change in output from a function to a change in its input that varies with time or with the frequency of the input.

Integral action: The rate of change of controller output responding to a deviation between the controlled variable and set point.

Load: A variable or aggregate of variables that affects the controlled variable.

Manipulated variable: A variable changed by the controller to move the controlled variable.

Noise: A disturbance having a frequency range too high for the controller to affect.

Offset: Steady-state deviation between the controlled variable and set point.

Open-loop response: The response of a controlled variable to process inputs in the absence of control action.

Proportional action: The change in controller output responding directly to a change in the controlled variable.

Proportional band: The percentage change in the controlled variable required to drive the controller output full scale.

Self-regulation: The property of a process through which a change in the controlled variable affects either the flow into or out of a process in such a way as to reduce further changes in the controlled variable.

Tuning: Adjusting the settings of a controller to affect its performance.

Windup: Saturation of the integral model of a controller that results in the controlled variable subsequently overshooting the set point.

References

1. J. G. Ziegler and N. B. Nichols, Optimum settings for automatic controllers, *Trans. ASME*, 64, 759–768, 1942.
2. F. G. Shinskey, *Process Control Systems*, 4th edn., New York: McGraw-Hill, pp. 22–25, 25–28, 192–194, 388–390, 1996.
3. F. G. Shinskey, *Feedback Controllers for the Process Industries*, New York: McGraw-Hill, pp. 68–70, 176–178, 157–164, 148–151, 155–156, 1994.

Further Readings

Astrom, K. J. and T. Hagglund, *Advanced PID Control*, Triangle Park, NC: ISA Publication, 2006.

Finn Peacock, PID loop tuning, http://www.pidtuning.net/ (accessed on July 24, 2012).

PID Algorithm, Straight line control, LLC, http://straightlinecontrol.com/pid_algorithms.html (accessed on July 25, 2012).

Partial List of Suppliers

Applied Processor & Measurement, 8201 Old Post Rd. E., East Amherst, NY 14051, Phone: 716-741-1141, Fax: 716-741-1142, www.appliedprocessor.com

Autonics, 1353 Armour Blvd., Mundelein, IL 60060, Phone: 847-680-8160, 866-862-2602 (toll free), Fax: 847-680-8155, www.autonicsonline.com

Control Warehouse, 356 Cypress Rd., Ocala, FL 34472, Phone: 888-212-5531 (toll free), Fax: 352-687-8925, www.controlswarehouse.com

Divelbiss, 9778 Mt. Gilead Rd., Fredericktown, OH 43019, Phone: 740-694-9015, 800-245-2327 (toll free), Fax: 740-694-9035, www.divelbiss.com

Miller Energy, Inc., 3200 S. Clinton Ave., South Plainfield, NJ 07080, Phone: 908-755-6700, 800-631-5454 (toll free), Fax: 908-755-0312, www.millerenergy.com

WATTCO, 500 S. Federal Hwy., P.O. Box 1811, Hallandale, FL 33009, Phone: 514-488-9124, 800-492-8826 (toll free), Fax: 514-488-9126, www.wattco.com/



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

92

Optimal Control and the Software

92.1 Introduction **92**-1
92.2 Cost Functions **92**-3
92.3 Calculus of Variations **92**-6
92.4 Riccati Equation **92**-8
92.5 State-Feedback Matrix **92**-11
92.6 Software Support and Applications **92**-14
Bibliography **92**-15
Partial List of Providers **92**-15

Halit Eren
Curtin University

92.1 Introduction

Optimal control maximizes (or minimizes) the value of a function chosen as the *performance index* or *cost function* from an operational control system. Optimal control theory, on the other hand, is the mathematics of finding parameters that causes the performance index to take an extreme value subject to system constraints.

Optimal control theory is based on calculus variations. It is a mathematical approach for determining control policies in control systems. Some of the first and main contributors of optimal control theory are Lev Pontryagin and his colleagues. The theory made a leap forward by the contribution of Richard Bellman in the United States. Optimal control is applied in many fields, such as satellites and aerospace, aircrafts and spacecrafts, chemical engineering, communications engineering, robots and robotics, power systems, electrical drives, computers and computer systems, management science, finance, and many other areas. In many applications, simple interconnections of control devices and controllers to the processes do not provide the most economical operations. Therefore, optimization is a powerful technique to obtain the best results from a known process. If the process is operating under a steady-state condition, optimization considers the process stationary, and it is only concerned with the operating points. When the process is stationary, the resulting optimum operating point can easily be maintained by setting the predetermined set points and precalculated control parameters. Nevertheless, if the process changes from time to time, new optimum set points for the system need to be determined for each change.

The performance of systems are optimized for many reasons such as improving the quality, increasing production, decreasing waste, obtaining greater efficiency, maximizing the safety, and saving time and energy. In many optimization problems, boundary conditions are imposed by the system for safety in operations, availability of minimum and maximum power, limitations in storage capacity, capability of the operating machinery, temperature, speed, force, acceleration, and indeed any hard physical reasons.

Many different types of mathematical methods may be employed for the solution of optimal control problems depending on the nature of the problem. For example, if the performance index and constraints can be formulated as linear algebraic functions of the controlled variables, then selection of the linear programming may be the best way to go. *Simplex methods* may provide a good way of solving the linear programming problem. If the equations describing the system are nonlinear, then the solution may involve nonlinear techniques or linearization of the problem in some subregions. If the problem involves determination of the largest value of a function with two or more variables, then the *steepest ascent (or descent) methods,* sometimes termed as *hill-climbing* or *gradient methods*, may be used.

Hill-climbing methods still are popular as they are easy to understand and to use. They are briefly discussed here to lay a good background for the sound understanding of optimal control theory.

An illustration of the hill-climbing technique in the region of absolute minima is given in Figure 92.1. This figure also illustrates a common pitfall in control theory, the local maxima and local

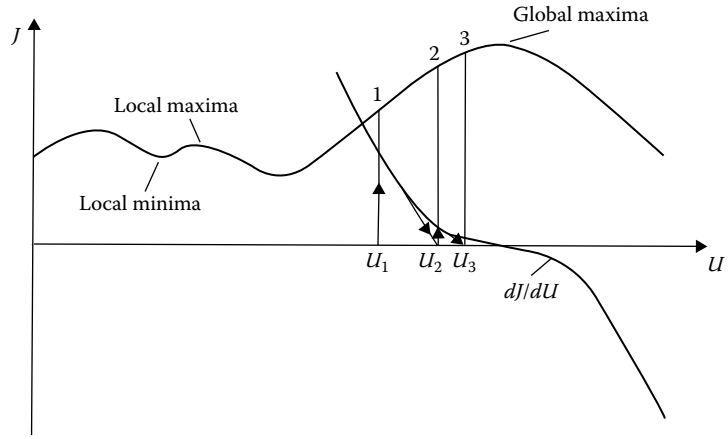


FIGURE 92.1 Hill-climbing technique illustrated in the region of absolute maxima. The sign of the first derivative of the function changes as the derivative passes through the maximum point. This figure shows the local maxima and minima that should be taken care of in optimal control design.

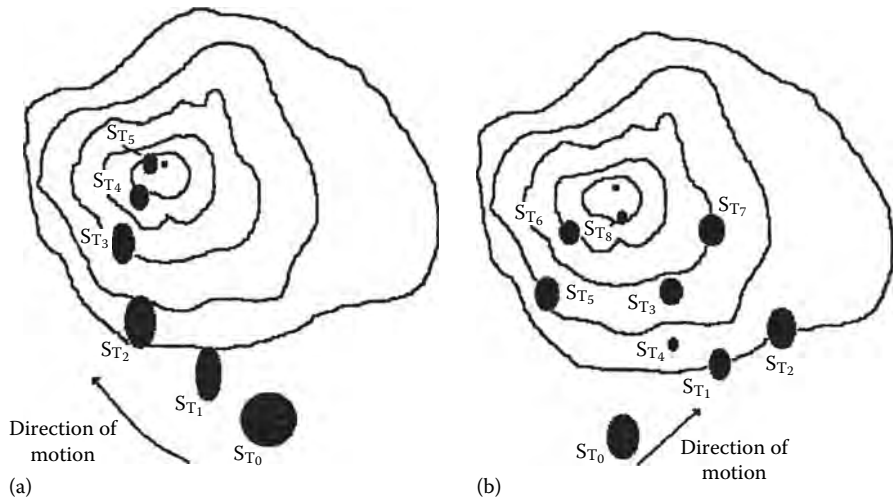


FIGURE 92.2 Examples of hill-climbing method: (a) gradient and (b) random walk. All these methods use an initial estimated point where the first derivative is guessed to be zero. This point is then used, in the direction of steepest ascent or steepest descent as the case may be, for the next point until the absolute maximum or minimum is found.

minima. As it can be seen, in this figure, the sign of the first derivative of the function changes as the derivative passes through the maximum point. This method commonly uses an initial estimated point where the first derivative is considered (guessed) to be zero. This point is then used, in the direction of steepest ascent or descent as the case may be, for the next point until the absolute maxima or minima is found without being trapped in local extreme points. There are many different methods to find maxima or minima by hill-climbing methods, and Figure 92.2 illustrates some examples of implementation.

Apart from hill-climbing techniques, there are many other and relatively new methods available for the solutions of modern optimal control problems. Some of the methods involve complex hyperspaces (many variables) having stringent constraints. If a good mathematical model describing the system is available, then various optimization methods may be used for finding the optimum conditions on the model itself rather than seeking optimum conditions on the actual process. This indicates that availability of the model determines the type of optimization method that can be employed.

92.2 Cost Functions

In designing optimal control systems, rules may be set for determining the control decisions, subject to certain constraints, such that some measure of deviations from an ideal case is minimized. That measure is usually provided by a selected *performance index* or *cost function*. Performance index is a function whose value indicates how well the actual performance of the system matches the desired performance. Appropriate selection of the performance index is important since it determines the nature and complexity of the problem. That is, the optimal control outcome will be dependent on the selected performance index. The choice of the performance index depends on system specifications, physical realizability, and the restrictions on the controls to be used. In general, the choice of an appropriate performance index involves a compromise between a meaningful evaluation of system performance and availability of a feasible mathematical description, which may be linear or nonlinear, or stationary or time varying.

The performance index is often selected by the engineer to make the system behave in a desired fashion. By definition, a system whose design minimizes the selected performance index is optimal. It is important to point out that a system optimal for a performance index may not be optimal under another performance index. In practical systems, due to possible complexities and cost of implementation, it may be better to employ approximate optimal control laws that are not rigidly tied to a single performance index.

Before starting optimization of any system, it is necessary to formulate the system by having information on system parameters and describing equations, constraints, class of allowable control vectors, and the selected performance indexes. Then the solution can be proceeded by determining the optimal *control vector* $\mathbf{u}(k)$ within the class of allowable vectors. The control vector $\mathbf{u}(k)$ will be dependent on factors such as the nature of performance index, constraints, initial values of state, initial outputs, and the desired state as well as the desired outputs. If analytical solutions are impossible or too complicated, then alternative computational solutions may be employed.

In simple cases, errors between the desired and actual responses can be selected as the performance index to be minimized. As an example, different descriptions of error between actual and desired responses are depicted in Figure 92.3. The aim is to keep the errors as small as possible. The time integral of an error gives the severity of the error. However, since the positive and negative errors mathematically cancel each other, absolute values must be used as

$$\text{Integral absolute error IAE} = \int_0^T |e(\tau)|\, d\tau \tag{92.1}$$

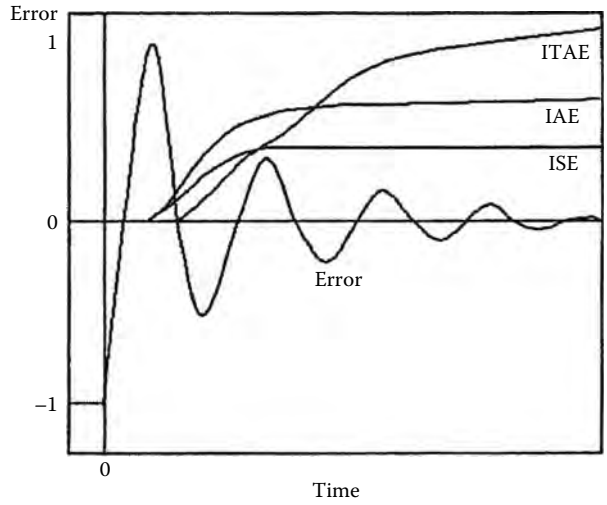


FIGURE 92.3 Description of errors between actual and desired responses. In simple optimal control applications, errors between the desired and actual responses can be selected as the performance indexes. The aim is to keep the errors as small as possible.

In many cases, it is better to use integral squared errors, which take into account large errors rather than regular ones

$$\text{Integral squared error ISE} = \int_0^T e^2(\tau)d\tau \tag{92.2}$$

Time-dependent functions can be implemented to capture errors occurring late in time rather than normal transients:

$$\text{Integral time absolute error ITAE} = \int_0^T |e(\tau)|\tau d\tau \tag{92.3}$$

and

$$\text{Integral time squared error ITSE} = \int_0^T e^2(\tau)\tau d\tau \tag{92.4}$$

The mean square error, MSE, is also commonly used since slight modification leads to inclusion of statistical techniques and analysis of random noise

$$\text{Mean square error MSE} = \lim_{T\to\infty} \frac{1}{T}\int_0^T e^2(\tau)\tau d\tau \tag{92.5}$$

Modern optimal control theory is developed within the *state-space* framework, and performance indexes are more complex and comprehensive, some of which are briefly visited as follows.

Suppose that the control command of a system is expressed in vectorial form as **u** and the state of the system is described by **x** (Oxford English Dictionary describes *state* as a condition with respect to circumstances, attributes, structure, form phase, or the like). Further, suppose that the rate of chance of state $\dot{\mathbf{X}}$ is a function of state **x**, control command **u,** and time *t*:

$$\dot{\mathbf{X}} = f(\mathbf{x}, \mathbf{u}, t) \quad \mathbf{x}(0) = \mathbf{x}_0 \tag{92.6}$$

Then a control law $\mathbf{u}(\mathbf{x}, \mathbf{t})$ or a control history $\mathbf{u}(t)$ can be determined such that a performance index or a scalar functional

$$J(u) = \int_0^T g(\mathbf{x}(\tau), \mathbf{u}(\tau), \tau) d\tau \tag{92.7}$$

This takes a minimum value out of all other possibilities and still holds as in Equation 92.6. A boundary relationship $\mathbf{x}(T) = \mathbf{x}_f$ must also be met as a constraint.

The most common form of the performance index, $J(u)$, is the minimum time control that can be expressed as

$$J(u) = \int_0^T d\tau = T \tag{92.8}$$

Many different criteria are also used, such as minimum fuel consumption, minimum energy consumption, and few other quadratic forms:

$$J(u) = \int_0^T |u(\tau)| d\tau \tag{92.9}$$

$$J(u) = \int_0^T u^2(\tau) d\tau \tag{92.10}$$

$$J(u) = \int_0^T (qx^2(\tau) + ru^2(\tau)) d\tau \tag{92.11}$$

A general term for the continuous time performance index leading to optimal control is expressed as

$$J(u(t)) = \int_0^T g(x(\tau), u(\tau)) d\tau \tag{92.12}$$

The performance indices are minimized for the constraints
$\dot{\mathbf{X}}(t) = f(\mathbf{x}(t), \mathbf{u}(t), t)$, for $t \in (t_0, t_f)$ and $\mathbf{x}(t)$ is an admissible state and $\mathbf{x}(t) \in \mathbf{X}(t)$ and $\forall t \in (t_0, t_f)$ are satisfied.

Slight variations of Equations 92.9 through 92.12 lead to mathematics of the discrete time or digital versions of the optimal control.

92.3 Calculus of Variations

Calculus variations are suitable to solve linear or nonlinear optimization problems with linear or nonlinear boundary conditions. Basically, it is a collection of many different analytical methods, and they are discussed differently from book to book. In here, a typical approach that leads to more general and widely used modern theories is introduced.

Consider a dynamic system operating in a time interval $t_0 < t < t_f$:

$$\dot{\mathbf{X}} = f\left[\mathbf{x}, \mathbf{u}, t\right] \tag{92.13}$$

where the initial state $\mathbf{x}_0$ is given.

The system has n state and m control variables. The scalar function to be optimized is

$$J = k\left[\mathbf{x}_f, t_f\right] + \int \mathbf{L}\left[\mathbf{x}, \mathbf{u}, t\right] dt \tag{92.14}$$

Define a *scalar*, such as Hamiltonian

$$H = H\left[\mathbf{x}, \mathbf{u}, \lambda t\right] = L\left[\mathbf{x}, \mathbf{u}, t\right] + \sum \lambda_i f_i \tag{92.15}$$

λ is known as the *Lagrange multiplier*. Also, we need to define a modified objective function:

$$J = k\left[\mathbf{x}_1, t_1\right] + \int H dt \tag{92.16}$$

The resulting solution is optimal when

$$\lambda = \frac{-\delta H}{\delta x} = \frac{-\delta L}{\delta x} - \sum \frac{-\lambda_i \delta f_i}{\delta x} \tag{92.17}$$

$$\lambda_f = \frac{\delta k}{\delta x} \tag{92.18}$$

and

$$\frac{\delta H}{\delta u} = \frac{\delta L}{\delta u} + \sum \frac{\lambda_i \delta f_i}{\delta u} = 0 \quad \text{for} \quad t_0 < t < t_f \tag{92.19}$$

The solutions of Equations 92.17 and 92.19 in the aforementioned form are difficult to obtain. Nevertheless, based on the these ideas, more general theories can be developed such as Pontryagin's maximum principle.

An advancement over the calculus variations is the Pontryagin's maximum principle, which offers easier solutions and expands the range of applicability to bounded control problems. In simple form, this principle may be explained as follows:

Given the system

$$\mathbf{x}_i = f_i[\mathbf{x}, \mathbf{u}, t] \tag{92.20}$$

with an objective function

$$J = \sum k_i x_i(t_f) \tag{92.21a}$$

the maximum principle states that if the control vector **u** is optimum, then the Hamiltonian

$$H = \sum p_i f_i \tag{92.21b}$$

is also maximized with respect to the control vector **u** over the set intervals:

$$p_i = \sum p_j \frac{\delta f_j}{\delta x_j} \tag{92.21c}$$

$$p_i(t_f) = k_i \tag{92.21d}$$

where p is the adjoint or costate vector, and

$$\frac{\delta H}{\delta p_i} = f_i[\mathbf{x}, \mathbf{u}, t] = x_i \tag{92.21e}$$

$$\frac{\delta H}{\delta p_i} = \sum \frac{p_i \delta f_j}{\delta x_j} = -p_i \tag{92.21f}$$

The initial values of state vector **x** provide the remaining constants necessary to solve these equations for determining an optimum control vector. This method is also applicable to systems with complex performance indexes.

Another expansion of calculus variations is the *Kalman filter*. Kalman is essentially a general filtering technique that can be applied to solutions of problems such as optimal estimation, prediction, noise filtering, stochastic optimal control, and designing of optimal controllers. The method has the advantage of providing estimates of variables in the presence of noise for both stationary and nonstationary processes. Most other types of noise can be treated by this method if they can be translated to Gaussian form. The technique can also be employed in systems with forcing disturbances and containing more than one noise source. Generally, systems that have either quadratic objective functions or uncorrelated Gaussian noise as the input are suitable for this technique.

A means of applying the Kalman filter to a nonlinear system is to find a good estimate of the system and then use it to define a new set of linear *state equations* that approximates the system to linear form at a normal operating point. Then the filter can be applied to the new set of linear equations. Applications of Kalman filter are limitless in instrumentation and measurement systems and other optimal control problems. It can easily be programmed on digital computers.

92.4 Riccati Equation

Some of the models in optimization problems resemble the models of traditional control theory and practice. In these models, the process and control variables are vector-valued and constrained by linear plant equations. The cost functions are in the form of quadratic costs. These type of systems are termed as *linear/quadratic* systems or LQ systems. The theory and models of LQ systems are well developed for both deterministic and stochastic cases.

Express a quadratic optimal control problem as

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) \quad \mathbf{x}(0) = \mathbf{x}_0 \tag{92.22}$$

where

$\mathbf{x}(k)$ is the state vector (n-vector)
$\mathbf{u}(k)$ is the control vector (r-vector)
$\mathbf{A}$ is the $n \times n$ nonsingular vector
$\mathbf{B}$ is the $n \times r$ matrix

The aim is to find the optimal control sequence $\mathbf{u}(0), \mathbf{u}(1), \ldots, \mathbf{u}(N-1)$ that minimizes a quadratic performance index

$$J = \frac{1}{2}\mathbf{x}^{\star}(N)\mathbf{S}\mathbf{x}(N) + \frac{1}{2}\sum \mathbf{x}^{\star}(k)\mathbf{Q}\mathbf{x}(k) + \mathbf{u}^{\star}(k)\mathbf{R}\mathbf{u}(k) \tag{92.23}$$

where

$\mathbf{x}^{\star}$ and $\mathbf{u}^{\star}$ are transpose of $\mathbf{x}$ and $\mathbf{u}$ matrixes, respectively
$\mathbf{Q} = n \times n$ positive definite or positive semidefinite symmetric matrix
$\mathbf{R} = r \times r$ positive definite or positive semidefinite symmetric matrix
$\mathbf{S} = n \times n$ positive definite or positive semidefinite symmetric matrix

Matrices **Q, S,** and **R** are selected to weight the relative importance of the performance measures caused by the state vectors $\mathbf{x}(k)$, the final state $\mathbf{x}(N)$, and control vectors $\mathbf{u}(k)$, respectively, for $k = 0, 1, 2, \ldots, N-1$.

The initial state of the system is arbitrary but the final state $\mathbf{x}(N)$ may be fixed. If the final state is fixed, then the term 1/2 $\mathbf{x}^{\star}(N)$ $\mathbf{S}\mathbf{x}(N)$ may be removed from the performance index, and the terminal state $\mathbf{x}_f$ may be imposed. If the final state is not fixed, then the term 1/2 $\mathbf{x}^{\star}(N)$ $\mathbf{S}\mathbf{x}(N)$ represents the weight of the performance measure to the final state.

There are many different ways of solving the aforementioned equations, one of which makes use of the concept of Lagrange multipliers. With the aid of Lagrange multipliers, the performance index may be modified as

$$\mathbf{L} = \frac{1}{2}\mathbf{x}^{\star}(N)\mathbf{S}\mathbf{x}(N) + \frac{1}{2}\sum \mathbf{x}^{\star}(k)\mathbf{Q}\mathbf{x}(k) + \mathbf{u}^{\star}(k)\mathbf{R}\mathbf{u}(k) + \lambda^{\star}(k+1)$$
$$\left[\mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) - \mathbf{x}(k+1)\right] + \left[\mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) - \mathbf{x}(k+1)\right]^{\star} \lambda^{\star}(k+1) \tag{92.24}$$

It is known that minimization of the function **L** is equivalent to minimization of performance index **J** under the same constraints.

In order to minimize, **L** needs to be differentiated with respect to vectors $\mathbf{x}(k)$, $\mathbf{u}(k)$, and $\boldsymbol{\lambda}(k)$, and the results must be set to zero. The partial differentiation of the function **L** with respect to variables gives the following:

$$\frac{\delta \mathbf{L}}{\delta \mathbf{x}(k)} = 0 \quad \mathbf{Q}\mathbf{x}(k) + \mathbf{A}^{*}\lambda(k+1) - \lambda(k) = 0 \tag{92.25}$$

$$\frac{\delta \mathbf{L}}{\delta \mathbf{x}(N)} = 0 \quad \mathbf{S}\mathbf{x}(N) - \lambda(N) = 0 \tag{92.26}$$

$$\frac{\delta \mathbf{L}}{\delta \mathbf{u}(k)} = 0 \quad \mathbf{R}\mathbf{u}(k) + \mathbf{B}^{*}\,\lambda(k+1) = 0 \tag{92.27}$$

and

$$\frac{\delta \mathbf{L}}{\delta \lambda(k)} = 0 \quad \mathbf{A}\mathbf{x}(k-1) + \mathbf{B}\mathbf{u}(k-1) - \mathbf{x}(k) = 0 \tag{92.28}$$

A close inspection of the preceding formulae indicates that the last equation is simply the state equation for $k = 1, 2, 3, \ldots, N$. And also, the value of Lagrange multiplier can be determined by Equation 92.16. The Lagrange multiplier is often termed as *covector* or *adjoint vector*.

Rewriting Equations 92.25 and 92.27,

$$\lambda(k) = \mathbf{Q}\mathbf{x}(k) + \mathbf{A}^{*}\lambda(k+1) \tag{92.29}$$

$$\mathbf{u}(k) = -\mathbf{R}^{-1}\mathbf{B}^{*}\,\lambda(k+1) \tag{92.30}$$

The state Equation 92.22 can be expressed as

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) - \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{*}\,\lambda(k+1) \tag{92.31}$$

In order to obtain the solution to the minimization problem, we need to solve Equations 92.29 and 92.31 simultaneously as a two-point boundary-value problem. The solutions of these two equations in this form, and the optimal values of the state vector and the Lagrange multiplier vector lead to the values of control vector $\mathbf{u}(k)$, which will optimize the open-loop control system. However, for closed-loop control systems, the Riccati transformations must be applied:

$$\mathbf{u}(k) = -\mathbf{K}(k)\mathbf{x}(k) \tag{92.32}$$

where $\mathbf{K}(k)$ is the $r \times n$ feedback matrix. Now the Riccati equation in feedback form can be obtained by assuming that $\boldsymbol{\lambda}(k)$ is written as

$$\lambda(k) = \mathbf{P}(k)\mathbf{x}(k) \tag{92.33}$$

where $\mathbf{P}(k)$ is an $n \times n$ matrix. Substituting Equation 92.33 into 92.29 and 92.31 gives

$$\mathbf{P}(k)\mathbf{x}(k) = \mathbf{Q}\mathbf{x}(k) + \mathbf{A}^* \mathbf{P}(k+1)\mathbf{x}(k+1) \tag{92.34}$$

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) - \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^* \mathbf{P}(k+1)\mathbf{x}(k+1) \tag{92.35}$$

In writing Equations 92.34 and 92.35, the Lagrange multiplier $\boldsymbol{\lambda}(k)$ has been eliminated. This is an important step in solving two-point boundary-value problems. By further manipulations, it is possible to show that

$$\mathbf{P}(k) = \mathbf{Q} + \mathbf{A}^* \mathbf{P}(k+1)\mathbf{A} - \mathbf{A}^* \mathbf{P}(k+1)\mathbf{B}\left[\mathbf{R} + \mathbf{B}^* \mathbf{P}(k+1)\mathbf{B}\right]^{-1} \mathbf{B}^* \mathbf{P}(k+1)\mathbf{A} \tag{92.36}$$

Equation 92.36 is known as the Riccati equation.

From Equations 92.26 and 92.33, writing $\lambda(N) = \mathbf{S}\mathbf{x}(N) = \mathbf{P}(N)\mathbf{x}(N)$ gives

$$\mathbf{P}(N) = \mathbf{S} \tag{92.37}$$

Hence, the Riccati equation can be solved backward from $k = N$ to $k = 0$, starting from the known values of $\mathbf{P}(N)$.

The optimal control vector $\mathbf{u}(k)$ can now be calculated from Equations 92.29, 92.30, and 92.35 as

$$\mathbf{u}(k) = -\mathbf{R}^{-1}\mathbf{B}^* \lambda(k+1) = -\mathbf{R}^{-1}\mathbf{B}^* \left[\mathbf{A}^*\right]^{-1}\left[\mathbf{P}(k) - \mathbf{Q}\right]\mathbf{x}(k) = -\mathbf{K}(k)\mathbf{x}(k) \tag{92.38}$$

where

$$\mathbf{K}(k) = -\mathbf{R}^{-1}\mathbf{B}^* \left[\mathbf{A}^*\right]^{-1}\left[\mathbf{P}(k) - \mathbf{Q}\right] \tag{92.39}$$

It is worth noting that optimal control vector may be obtained in slightly different forms by slightly different manipulations of the aforementioned equations, such as

$$\mathbf{u}(k) = -\mathbf{R}^{-1}\mathbf{B}^* \left[\mathbf{P}^{-1}(k+1) + \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^*\right]^{-1} \mathbf{A}\mathbf{x}(k) \tag{92.40}$$

Equations 92.38 and 92.39 indicate that optimal control law requires feedback of the state vector with time-varying gain $\mathbf{K}(k)$. Figure 92.4 illustrates the optimal control scheme of a system based on the quadratic performance index. In practical applications, the time varying $\mathbf{K}(k)$ is calculated before the process begins. Once the state matrix $\mathbf{A}$, control matrix $\mathbf{B}$, and weighting matrices $\mathbf{Q}$, $\mathbf{R}$, and $\mathbf{S}$ are known, the gain $\mathbf{K}(k)$ may be precomputed off-line to be used later. The control vector $\mathbf{u}(k)$ at each stage can be determined immediately by premultiplying the state vector $\mathbf{x}(k)$ by the known gain $\mathbf{K}(k)$.

From the above equations, the minimum value of the performance index can also be calculated, by using the initial values, as

$$J_{\min} = \frac{1}{2}\mathbf{x}^*(0)\mathbf{P}(0)\mathbf{x}(0) \tag{92.41}$$

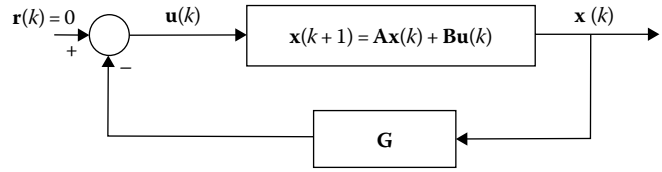


FIGURE 92.4 Optimal control scheme based on the quadratic performance index. The control law requires feedback of the state vector with time-varying gain $\mathbf{K}(k)$. The gain $\mathbf{K}(k)$ is calculated for use later, before the process begins.

The *steady-state* solutions of Riccati equations are necessary when dealing with time-invariant (steady-state) optimal controls. There are many ways of obtaining steady-state Riccati solutions; an example is given as follows:

$$\mathbf{P} = \mathbf{Q} + \mathbf{A}^* \mathbf{PA} - \mathbf{A}^* \mathbf{PB}(\mathbf{R} + \mathbf{B}^* \mathbf{PB})^{-1}\mathbf{B}^* \mathbf{PA} \tag{92.42}$$

The steady-state value of $\mathbf{K}$ may be found as

$$\mathbf{K} = (\mathbf{R} + \mathbf{B}^* \mathbf{PB})^{-1}\mathbf{B}^* \mathbf{PA} \tag{92.43}$$

Here, the optimal control law for the steady-state operation may be expressed as

$$\mathbf{u}(k) = -(\mathbf{R} + \mathbf{B}^* \mathbf{PB})^{-1}\mathbf{B}^* \mathbf{PAx}(k) \tag{92.44}$$

This section is presented for discrete time optimal control systems rather than continuous systems, due to recent widespread use of computers and microprocessors as online and off-line control tools. Since the principles are the same, solutions can easily be extended for continuous time systems with minor modifications.

92.5 State-Feedback Matrix

The state-feedback method is another design technique that allows the designer to locate the poles of the system wherever they are needed. This type of approach is termed as the *pole-placement method* where the term "pole" refers to the poles of the closed-loop transfer function as in Figure 92.5. In this method, it is assumed that state variables are measurable and are available for feedback. If the state is available, the system is said to be deterministic; it is noise-free and all the parameters are known. If the state is not available, then methods such as measurement feedback laws or state estimators may be selected.

In many applications, instead of state variables, it is more convenient to use state estimates coming from an observer or Kalman filter. In this case, there are three well-known methods available, these being the certainty equivalent, the separation, and the dual control. Certainty equivalence has considerable advantages over others since they lead to deterministic optimal control laws such Pontryagin's principle or the conditional mean of the state such as the Kalman filter. It leads to a practical controller that may be built as a filter and optimal law in cascade.

State-feedback controllers are relatively easy to implement. For example, in the pole-placement method, the relationship of the feedback control $\mathbf{u}$ to state $\mathbf{x}$ for *linear systems* is

$$\mathbf{u}(k) = -\mathbf{Kx}(k) \tag{92.45}$$

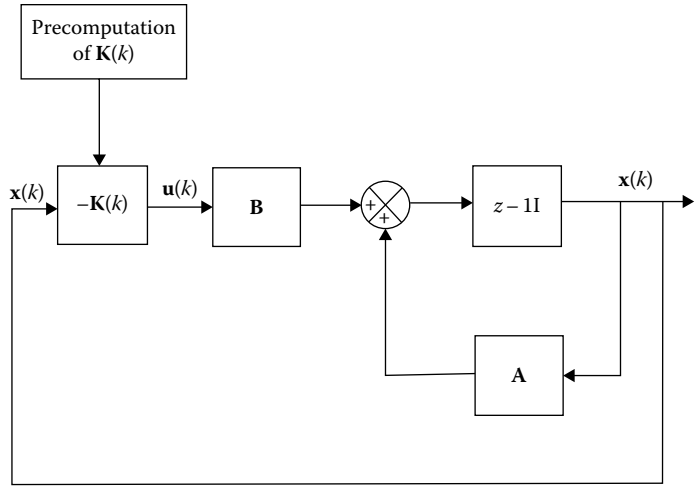


FIGURE 92.5 Pole-placement design of closed-loop transfer function. This method assumes that state variables are measurable and are available for the feedback.

In linear quadratic cases, the gain **K** is time varying:

$$\mathbf{u}(k) = -\mathbf{K}(k)\mathbf{x}(k) \tag{92.46}$$

where **K** is a $(n \times r)$ feedback matrix. This is correct at least in two cases:

1. In linear systems with no noise, pole-placement controllers, and an observer as the state estimator, the control algorithms for

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k)$$

$$\mathbf{y}(k) = \mathbf{C}\mathbf{x}(k)$$

   are

$$\hat{\mathbf{x}}(k+1) = (\mathbf{A} - \mathbf{G}\mathbf{C})\hat{\mathbf{x}}(k) + \mathbf{G}\mathbf{y}(k) + \mathbf{B}\mathbf{u}(k) \tag{92.47}$$

$$\mathbf{u}(k) = \mathbf{K}\hat{\mathbf{x}}(k) \tag{92.48}$$

   where the eigenvalues of $(\mathbf{A} - \mathbf{G}\mathbf{C})$ and $(\mathbf{A} + \mathbf{B}\mathbf{K})$ are selected to meet the design specifications.
2. In linear systems, influenced by Gaussian white noise, in which the control is optimal according to a quadratic criterion, the system

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) + \mathbf{G}\mathbf{v}(k) \tag{92.49}$$

$$\mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) + \mathbf{w}(k) \tag{92.50}$$

has control law of the form

$$\hat{\mathbf{x}}\left(\frac{k+1}{k}\right) = \mathbf{A}\hat{\mathbf{x}}(k) + \mathbf{B}\hat{\mathbf{u}}(k) \tag{92.51}$$

$$\hat{\mathbf{x}}(k+1) = \mathbf{A}\hat{\mathbf{x}}\left(\frac{k+1}{k}\right) + \mathbf{G}(k+1)\mathbf{y}(k+1) - \mathbf{C}\hat{\mathbf{x}}\left(\frac{k+1}{k}\right) \tag{92.52}$$

$$\hat{\mathbf{u}}(k) = \mathbf{K}(k)\hat{\mathbf{x}}(k) \tag{92.53}$$

where $\mathbf{K}(k)$ and $\mathbf{G}(k)$ are optimal gains for the deterministic optimal controls and the Kalman filters, respectively.

In practice, not all the state variables are easily accessible, and in general only the outputs of the system are measurable. Therefore, when feedback from the state variables is required in a given design, it is necessary to observe the states from information contained in the output as well as the input variables. The subsystem that performs the observation of the state variables based on the information received from the measurements of inputs and outputs is called the state observer. Figure 92.6 shows the block diagram of such a system.

Suppose a state-feedback gain $\mathbf{K}$ has been selected so that the eigenvalues of

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) \tag{92.54}$$

$$\mathbf{u}(k) = -\mathbf{K}(k)\mathbf{x}(k) + \mathbf{r}(k) \tag{92.55}$$

are located at $\lambda_1, \lambda_2, \ldots, \lambda_n$. Furthermore, assume that gain $\mathbf{G}$ in the identity observer

$$\hat{\mathbf{x}}(k+1) = (\mathbf{A} - \mathbf{GC})\hat{\mathbf{x}}(k) + \mathbf{G}\mathbf{y}(k) + \mathbf{B}\mathbf{u}(k) + \mathbf{G}\mathbf{y}(k) \tag{92.56}$$

is chosen such that eigenvalues of the observers are $\mu_1, \mu_2, \ldots, \mu_n$.

When the observer state estimate $\mathbf{x}(k)$ is used instead of state $\mathbf{x}(k)$, the resulting system has a state dimension $2n$, modeled by

$$\begin{bmatrix} \mathbf{x}(k+1) \\ \hat{\mathbf{x}}(k+1) \end{bmatrix} = \begin{bmatrix} \mathbf{A} & -\mathbf{BK} \\ \mathbf{GC} & \mathbf{A} - \mathbf{GC} - \mathbf{BK} \end{bmatrix} \begin{bmatrix} \mathbf{x}(k) \\ \hat{\mathbf{x}}(k) \end{bmatrix} + \begin{bmatrix} \mathbf{B} \\ \mathbf{B} \end{bmatrix} \mathbf{r}(k) \tag{92.57}$$

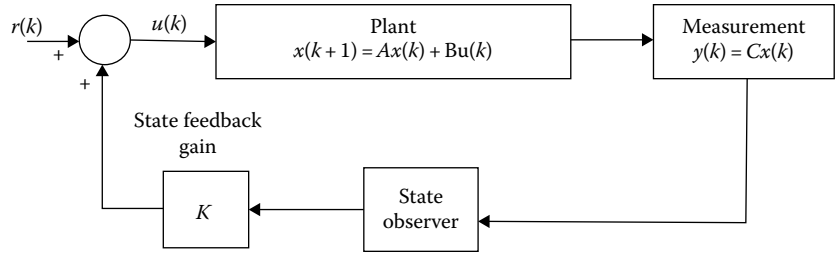


FIGURE 92.6 Use of state observer. When feedback from the state variables are required, it may be possible to observe the states from information contained in the output as well as the input variables based on the information received from the measurements.

A similarity transform

$$\begin{bmatrix} \mathbf{x}(k) \\ \mathbf{e}(k) \end{bmatrix} = \mathbf{P} \begin{bmatrix} \mathbf{x}(k) \\ \hat{\mathbf{x}}(k) \end{bmatrix} \tag{92.58}$$

where

$$\mathbf{P} = \begin{bmatrix} \mathbf{I} & 0 \\ \mathbf{I} & -\mathbf{I} \end{bmatrix} \tag{92.59}$$

corresponds to a change of variables to $\mathbf{x}(k)$ and $\mathbf{e}(k) = \mathbf{x}(k) - \mathbf{x}(k)$ converts this to

$$\begin{bmatrix} \mathbf{x}(k+1) \\ \mathbf{e}(k+1) \end{bmatrix} = \begin{bmatrix} \mathbf{A} - \mathbf{BK} & \mathbf{BK} \\ 0 & \mathbf{A} - \mathbf{GC} \end{bmatrix} \begin{bmatrix} \mathbf{x}(k) \\ \mathbf{e}(k) \end{bmatrix} + \begin{bmatrix} \mathbf{B} \\ 0 \end{bmatrix} \mathbf{r}(k) \tag{92.60}$$

This system has the same eigenvalues as the original because of the nature of the similarity transforms and the eigenvalues are the solution of

$$\det(\lambda \mathbf{I} - (\mathbf{A} - \mathbf{BK}))(\lambda \mathbf{I} - (\mathbf{A} - \mathbf{GC})) = 0 \tag{92.61}$$

This requires that one of

$$\det(\lambda \mathbf{I} - (\mathbf{A} - \mathbf{BK})) = 0 \tag{92.62}$$

$$\det(\lambda \mathbf{I} - (\mathbf{A} - \mathbf{GC})) = 0 \tag{92.63}$$

must hold. These are also the eigenvalues of the design for the pole-placement feedback state controller $\lambda_1, \lambda_2, \ldots, \lambda_n$ and of the design of the state observer $\mu_1, \mu_2, \ldots, \mu_n$, respectively. This indicates that the use of observer does not move the designed poles from the pole-placement algorithm.

The algorithms for pole-placement observer are well known, and they are part of many control design packages. Steady-state Kalman gains and optimal control gains are also commonly available in such programs. Codes for Kalman filters and LQ controllers are also easy to write.

Note: For further readings on this topic, please refer to the sources in the bibliography.

92.6 Software Support and Applications

There are not manufacturers of optimal controllers. Most mathematical packages such as *MATLAB*® Optimization Toolbox, CAD/CAM algorithms, majority of *artificial intelligence and neural network* software tools, and other simulation and design packages offer solutions to optimal control problems. Optimization tools are also part of specialized design packages addressing specific applications, such as FPGA, a language-based design, technology-specific optimizations; OPTCON, optimal control algorithms for linear stochastic models; DISNEL; and MIMO.

MATLAB has additional packages, such as the PROPT, and aims to solve dynamic optimization problems. The problems require mathematical description as a state-space model using ordinary differential

equations or differential algebraic equations. PROPT makes use of Chebyshev or Gauss point collections for solving optimal control problems. Once the initial and final conditions are known, using cost functionals leads to state trajectories.

Another product of MATLAB is an object-oriented software termed DIDO. It has been developed for solving hybrid optimal control problems and based on pseudospectral optimal control theory that mimics control problems in a manner as if being written on a piece of paper. It eliminates the need for solving the costates. Pontryagin's minimum principle is used as a backup tool to check the results.

Sparse Optimal Control Software (SOCS), developed by the Boeing Company, comes in two packages. They can be applied in very large problems, such as trajectory optimization, chemical processes, and machine tools. The software can be implemented on desktop computers and is also capable of solving problems that can contain up to 100,000 variables and constraints. It is capable of solving linear or nonlinear problems.

There are many commercially available and custom-built software packages, some of which are the following: BARON, a global optimization solver; BT, for minimization; CO (constrained optimization); DOC (design optimization control program); DOT (design optimization tool); GENESIS, a structural optimization software; LSNNO, for large-scale optimization; MCS, for global optimization; OPTIMA library, for optimization and sensitivity analysis; TNPACK, for nonlinear unconstrained minimization; TOMLAB, for MATLAB Optimization; and many others. More information can be found in websites, such as NEOS WIKI (http://www.neos-guide.org/NEOS/index.php/Optimization_Software_Guide).

Most of the software packages are used in optimal problem solving in industrial and nonindustrial areas. MATLAB PROPT, for example, finds applications in http://tomdyn.com/matlab_optimal_control_examples.html for the following: aerodynamic trajectory determination, bang-bang control systems, chemical engineering, dynamic systems, flight path tracking, large-scale linear optimal problems, pendulum control, path tracking of robots, mechanical engineering, turbo generators, marine population dynamics, nondifferentiable applications, parameter estimation, radio telescopes, drug displacement, missile intercept, satellite control, space shuttles, and many others.

Bibliography

Cakaj, S. *Modeling, Simulation and Optimization: Tolerance and Optimal Control*, InTech, Vienna, Austria, 2010.

Evans, L.C. *Introduction to Mathematical Optimal Control Theory: Version 0.2*, available at http://math.berkeley.edu/~evans/control.course.pdf (accessed on February 16, 2012).

Grimble, M.J. *Optimal Control Applications and Methods*, Wiley, Chichester, U.K., Online Library. http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1099-1514, (accessed on February 29, 2012).

Hocking, L.M. *Optimal Control: An Introduction to the Theory and Applications*, Oxford Press, Oxford, U.K., 2001.

Hull, D.J. *Optimal Control Theory for Applications*, Springer, New York, 2003.

Naidu, D.S. *Optimal Control Systems*, CRC Press, Boca Raton, FL, 2003.

Optimization Software Guide, Neos Wiki. http://www.neos-guide.org/NEOS/index.php/Optimization_Software_Guide (accessed on February 20, 2012).

PROPT MATLAB Optimal Control Software. http://tomdyn.com/index.html (accessed on February 28, 2012).

Weber, R. *Optimization and Control*, University of Cambridge, Cambridge, U.K., 2010.

Partial List of Providers

Centurion Controls, 102 Hamilton Industrial Court, Wentzville, MO 63385, Phone: 866-646-4383, Fax: 435-487-9250, www.centurioncontrols.com

Certeon, 4 Van De Graaf Dr., Burlington, MA 01803, Phone: 781-425-5200, Fax: 781-425-5210, www.certeon.com

Innovative Custom Software, Inc., 1323 Bond St., No. 103, Naperville, IL 60563, Phone: 630-892-5022, Fax: 630-748-0153, www.icsinc.net

Managing Automation, 5 Penn Plaza Fl 9, New York, NY 10001, Phone: 212-629-1577, www.managingautomation.com/maonline/

Sterling, 2900 S. 160th St. New Berlin, WI 53151, Phone: 262-641-8610, Fax: 262-641-8653, www.sterlco.com/

Trola Industries, 2360 N. George St., York, PA 17406-3202, Phone: 717-848-3700, Fax: 717-848-6993, www.trolaindustries.com

Zepnick Solution, 1310 Brookfield Ave., Green Bay, WI 54313, Phone: 920-662-1682, Fax: 920-662-1685, www.zepnick.com

93

Electropneumatic and Electrohydraulic Instruments: Modeling

93.1 Introduction **93**-1
93.2 Background **93**-1
Actuator Modeling Requirements
93.3 Hydraulic Actuator Modeling **93**-2
93.4 EHA Actuator Modeling **93**-4
Motor • Pump and Fluid • Piston and Flight Control Surface • Piston and Flight Control Surface Dynamics • Quantitative Feedback Theory (QFT)
93.5 Actuator Compensator Design **93**-10
93.6 Conclusion **93**-11
Nomenclature **93**-11
References **93**-12
Partial List of Vendors and Suppliers **93**-12

M. Pachter
Air Force Institute of Technology

C.H. Houpis
Air Force Institute of Technology

93.1 Introduction

In control systems, the output of the controller (also of the equalizer or compensator), which operates on the command (and feedback) input signals, is invariably sent to an actuator. The latter is a power element critically positioned at the plant input and, to a large extent, determines the performance of the overall control system. Thus, the importance of the accurate modeling of the actuator element with respect to achieving and maintaining the desired control system performance over its operating range cannot be overemphasized. The impact of items such as the actuator phase-lag characteristics, parameter variations and aging, sensor noise, and operating scenario (e.g., in a flight control context, it is referred to as *flight conditions*) should be explicitly considered in a control system design process. To achieve a robust control system, where these items are taken into account, requires a good understanding and an accurate model of the actuation element. This chapter discusses the actuator modeling process.

93.2 Background

Power-to-weight considerations have driven actuator design to pneumatic, hydraulic, electrohydraulic, and recently, in flight control applications, electrohydrostatic configurations. In this chapter, the point of view is taken that the actuator element constitutes a feedback control system in its own right. Thus, the actuator comprises of an "amplifier" element (the servo valve), a power element (the "ram"), an actuator displacement sensor, and a controller about which more will be said in the sequel. Furthermore, it is

important that the actuator subsystem be considered in the context of the complete control system, e.g., a flight control system. This point of view is promulgated throughout this chapter. First, a relatively simple hydraulic actuator is considered. The focus then shifts to the modern electrohydrostatic actuator (EHA) technology that, in the flight control application, offers a high degree of combat survivability and easier maintainability since its components are collocated with the actuator. Since the EHA does not require long hydraulic lines, required maintenance time and equipment can be reduced. This also substantially reduces the profile exposed to hostile fire. In addition, actuator failures due to various causes and/or aging need to be taken into account in designing the overall control system. The design of a robust control system, utilizing an actuator, requires an accurate actuator model in order to satisfy the system performance specifications. The EHA and its flight control application are used as the "vehicle" for presenting the actuator modeling process.

93.2.1 Actuator Modeling Requirements

When an actuator needs to be replaced in a control system, for reasons cited previously, by one "off the shelf," the robustness of the overall control system must be unaffected by this replacement unit. Thus, the actuator design process entails

1. The derivation of an accurate mathematical model for the actuator.
2. Sensitivity analyses on the actuator control systems. This should include sensitivity to variations in load, component efficiencies, physical plant parameters, and actuator sensor noise.
3. The identification of a reasonable set of plant variations for the actuator, based on the sensitivity analysis, which is required for designing a robust actuator control system that will meet the specifications in the presence of predicted variations.

93.3 Hydraulic Actuator Modeling

The modeling of a purely hydraulic actuator employed in an irreversible flight control system is presented first. Based upon the notation shown in Figure 93.1, the following relationships are obtained:

$$x_1 = \delta_{ec} a \tag{93.1a}$$

$$\frac{x_v - x_f}{b + c} = \frac{x_v - x_1}{c} \tag{93.1b}$$

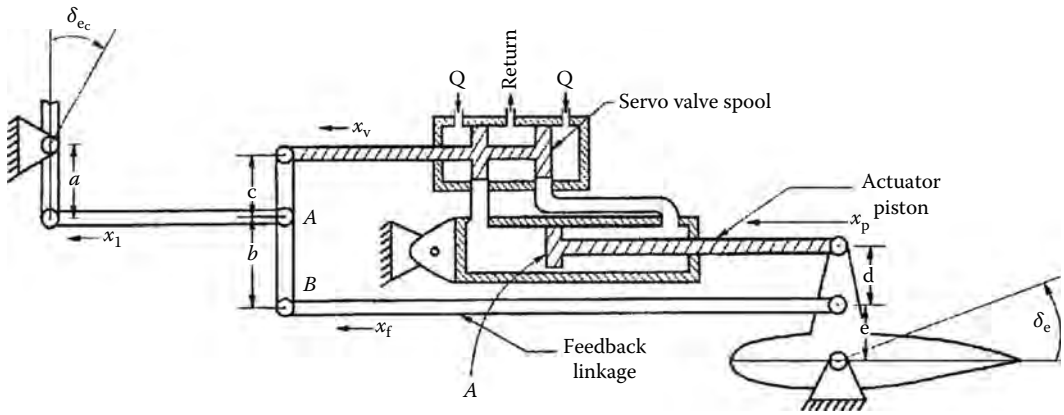


FIGURE 93.1 Irreversible control system.

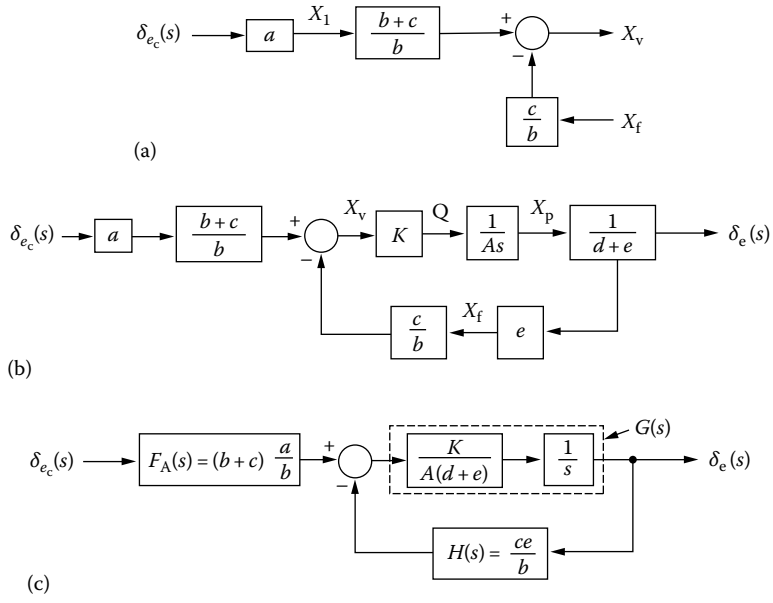


FIGURE 93.2 (a) Actuator block diagram, (b) block diagram of actuator feedback control system, and (c) simplified block diagram.

that yield

$$x_v b + x_f c = x_1(b + c) \tag{93.2}$$

From these equations, Figure 93.2a is obtained. Next, the relationships

$$\frac{dx_p}{dt} = \frac{Q}{A}, \quad Q = Kx_v, \quad \delta_e = \frac{x_p}{d+e}, \quad x_f = e\delta_e \tag{93.3}$$

are utilized to yield the block diagram of a typical feedback control system (shown in Figure 93.2b) that represents the irreversible hydraulic actuator under consideration. This block diagram is simplified to the one shown in Figure 93.2c, which yields the first-order transfer function of the actuator:

$$G(s) = \frac{\delta_e(s)}{\delta_{e_c}(s)} = \frac{K}{\tau s + 1} \tag{93.4}$$

where the gain $K = [a(b + c)]/ce$, the time constant $\tau = [Ab(d + e)]/cKe$ of the actuator, and $0.025 < \tau < 0.1$ s. For a step-input command, a typical actuator time response is shown in Figure 93.3.

The more-detailed dynamic modeling of the valve leads to higher-order hydraulic actuator models. For example, the transfer function of the currently used actuator of the F-16 aircraft is of fourth order [1], and its frequency response is compared with its first-order approximation in Figure 93.4. It is indeed important to use the full-blown high-order actuator dynamic model in robust control system design. As Figure 93.4 shows, the simplified first-order model fits well the high-order model in the low-frequency range. Unfortunately, a significant phase angle difference exists at the higher frequencies where the actuator will be operating, for the higher frequencies are within the bandwidth of modern robust ("high gain") control system compensators.

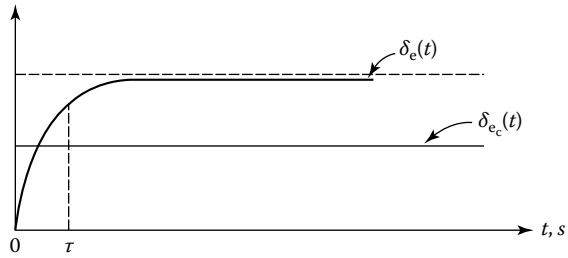


FIGURE 93.3 Time response of an actuator.

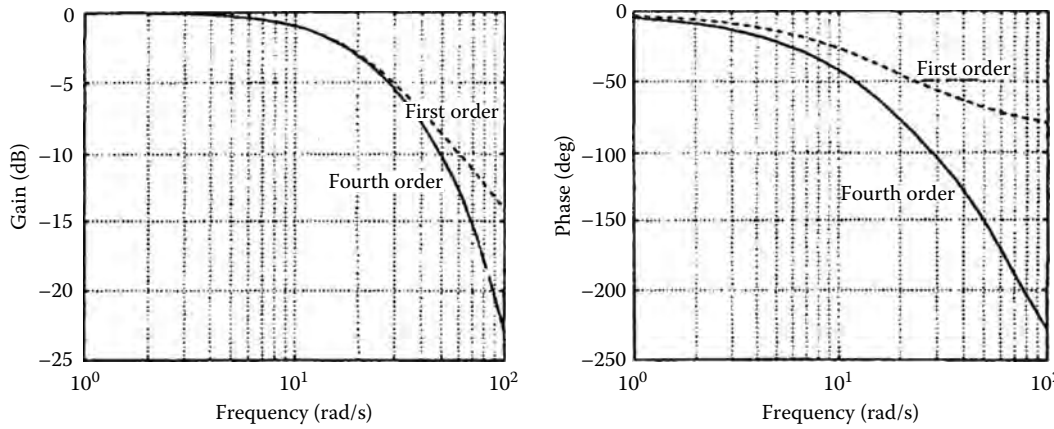


FIGURE 93.4 Comparison of the first- and fourth-order actuator models.

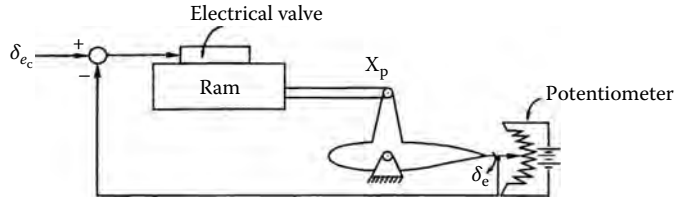


FIGURE 93.5 Electrohydraulic actuator.

The next step in the actuator state of development is then an electronic "fly-by-wire" mechanization of the feedback loop. This is shown in Figure 93.5 where an actuator is shown that employs an electric (solenoid controlled) valve and a potentiometer to measure the control surface deflection (actuator displacement) and where the power element of the actuator, the RAM, is a hydraulic cylinder.

93.4 EHA Actuator Modeling

The novel EHA actuator (see, e.g., Ref. [2]) offers a high degree of maintainability and survivability because all the actuator elements are collocated and a central hydraulic pump with the attendant high-pressure hydraulic plumbing is not needed. The modeling of the modules of the EHA is outlined first.

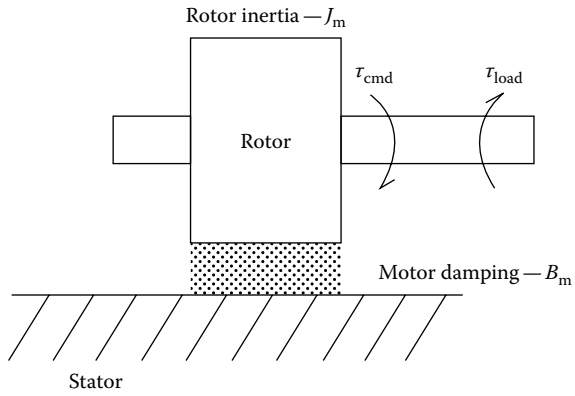


FIGURE 93.6 Motor model.

93.4.1 Motor

Electric motors with the rotor moment of inertia J_m and electromechanical damping B_m (see Figure 93.6) are subject to variations in output torque and subsequent fluctuations in rotor speed ω_m. Since the torque due to the load counteracts the torque generated by the motor, perturbation may also be caused by variations of the load torque. This relationship is expressed as

$$\tau_e(s) = \tau_{cmd}(s) - \tau_{load}(s) \tag{93.5}$$

where τ_{load} is the load torque due to the differential pressure of the fluid in the pump. This results in a first-order transfer function

$$\frac{\Omega_m(s)}{\tau_e(s)} = \frac{1}{J_m s + b_m} \tag{93.6}$$

93.4.2 Pump and Fluid

Electric motors have a limited torque-to-mass ratio, due to the finite and limited magnetic flux density that can be generated [3]. High-pressure hydraulic systems, with the system pressure of 2000–5000 psi, can generate high forces resulting in more compact and higher torque-to-mass ratios than electric motors. Generally, high-pressure hydraulic systems are stiffer against the load than electric motors. Hence, the EHA utilizes a DC motor to pump high-pressure fluid into the piston chamber. The DC motor internal to the EHA converts the electric power into mechanical power. It is the pump that converts this mechanical power into hydraulic power. The hydraulic power, acting against the piston, is converted to mechanical power capable of moving the load, e.g., large flight control surfaces.

The flow rate Q_m generated by the pump is proportional to the motor speed, i.e.,

$$Q_m(s) = \frac{D_m}{2\pi}\Omega_m(s) \tag{93.7}$$

where D_m represents the pump displacement constant. The flow rate of the hydraulic fluid is primarily dependent on two factors: change in chamber volume and change in pressure due to the compressibility of the fluid effect [4]. The chamber volume changes as the piston moves through the chamber at speed dx_p/dt. The flow rate due to the changes in chamber volume is then expressed as $\pm A dx_p/dt$.

Secondary fluid dynamic effects include fluid compressibility, internal leakage flow, and external leakage flow. These secondary effects are modeled by a first-order equation as follows:

$$\delta Q(s) = Q_m(s) - AsX_p = [K_s s + C_t]P(s) \tag{93.8}$$

that results in the transfer function

$$\frac{P(s)}{\delta Q(s)} = \frac{1}{K_s s + C_t} \tag{93.9}$$

93.4.3 Piston and Flight Control Surface

93.4.3.1 Piston and Flight Control Surface Fundamentals

The pressure developed by the pump and fluid acts on the piston surface (see Figure 93.7), causing the RAM to extend or retract. This force then generates a torque through a hinge to deflect the control surface. This torque has to overcome two load components: control surface inertia and aerodynamic loads. The control surface inertia is primarily due to the fact that the control surface has certain size and mass. The aerodynamic load only occurs in flight, when the air pressure over the control surface applies aerodynamic forces to it. The aerodynamic load is determined by three factors: the surface area of the flight control surface, the aerodynamic loading that varies with altitude and airspeed, and the relative angle of the surface to the wind. The surface angle to the wind depends on the angle of surface deflection and on the angle of attack of the aircraft.

93.4.4 Piston and Flight Control Surface Dynamics

93.4.4.1 Flexible Hinge Joint Model

The magnitude of F_A acting on the piston is equal to $\wp A$, where $\wp$ is the differential pressure developed by the pump and fluid and A is the surface area of the piston. Thus, the force created by the pump and fluid can be expressed as

$$F_A = \wp A \tag{93.10}$$

The piston dynamics, with the piston mass M_p and piston damping B_p, can be described by the second-order model:

$$F_A - F_p = [M_p s^2 + B_p s]X_p \tag{93.11}$$

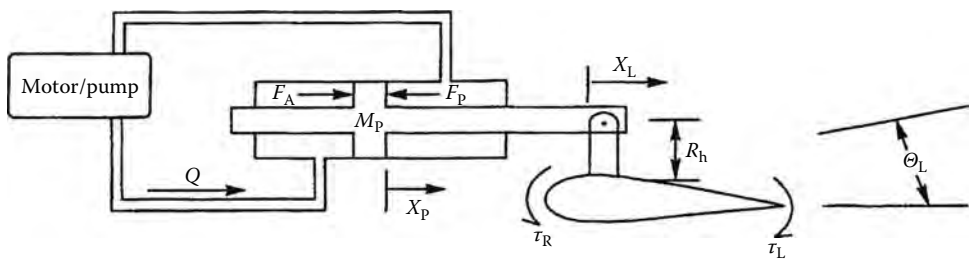


FIGURE 93.7 Simplified actuator control (not drawn to scale).

The resulting torque acting on the flight control surface due to this force imbalance can be described by

$$\tau_R = \frac{K_h}{R_h}(X_P - X_L) \tag{93.12}$$

where
K_h is the hinge stiffness constant
R_h is the hinge length

The stabilator inertia acts against the torque generated by the actuator, such that

$$\tau_R - \tau_L = [J_L s^2 + B_L s]\theta_L \tag{93.13}$$

where
τ_L is the torque created by the aerodynamic load and stabilator inertia
J_L and B_L represent the mass properties of the flight control surface

93.4.4.2 Stiff Hinge Joint Model

Equations 93.11 through 93.13 represent a rather complex model of load dynamics. The complexity of the model can be reduced if the linkage between the actuator and the flight control surface is considered rigid. This is a valid assumption, since the natural frequency of the hinge for a well-designed actuation system is much greater than the actuator bandwidth of concern. Hence, for a rigid linkage, the piston and the load dynamics are still expressed as in Equations 9.11 and 9.13.

$$F_A - F_P = [M_P s^2 + B_P s]X_P \tag{93.14}$$

and

$$\tau_R - \tau_L = [J_L s^2 + B_L s]\theta_L \tag{93.15}$$

Dividing Equation 93.15 by R_h yields

$$F_R - \frac{\tau_L}{R_h} = \left[\frac{J_L s^2 + B_L s}{R_h}\right]\theta_L \tag{93.16}$$

Assuming rigidity of the hinge assembly, $F_R \approx F_P$. Hence, adding Equations 93.14 and 93.16 results in

$$F_A - \frac{\tau_L}{R_h} = [M_P s^2 + B_P s]X_P + \left[\frac{J_L s^2 + B_L s}{R_h}\right]\theta_L \tag{93.17}$$

where τ_L is defined as

$$\tau_L = (\text{Load}_{\text{aero}})\,\theta_L \tag{93.18}$$

Since rigidity implies that $X_p = X_L$ and $\theta_L = X_p/R_h$, Equation 93.17 is further reduced to

$$F_e = F_A - \frac{\tau_L}{R_h} = \left[\left(M_P + \frac{J_L}{R_h^2}\right)s^2 + \left(B_P \frac{B_L}{R_h^2}\right)s\right]X_P \tag{93.19}$$

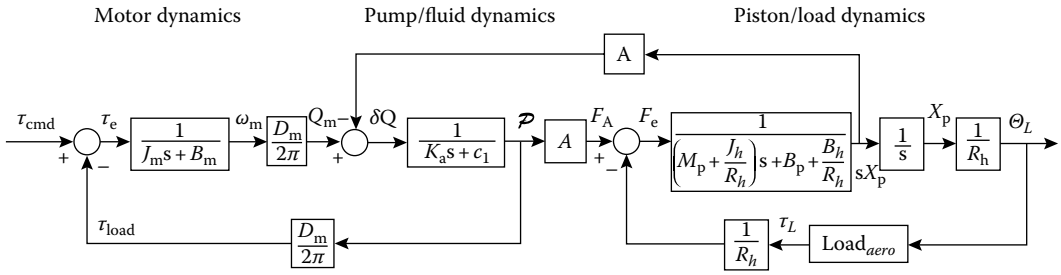


FIGURE 93.8 Bare EHA block diagram.

or expressed in a transfer function form:

$$\frac{X_p(s)}{F_e(s)} = \frac{1}{s\left[\left(M_P + (J_L / R_h^2)\right)s + \left(B_P(B_L / R_h^2)\right)\right]} \tag{93.20}$$

Equation 93.20 represents a simplified model of the load dynamics. A complete simplified design model of the bare EHA, without the controller, is shown in Figure 93.8 using the individual component models.

93.4.4.3 Hinge Moments

The aerodynamic loads mentioned previously describe the torque opposing the piston motion. Thus, load generated by the flow of air above and below the flight control surface applies a torque to the hinge assembly, which in turn adds back pressure to the piston. The hinge moment is modeled as follows:

$$\tau_L = \bar{q} S_t R_h \left(Ch_\alpha \frac{\alpha(s)}{\theta_L(s)} + Ch_\delta \right) \theta_L \tag{93.21}$$

where

S_t denotes the surface area of the control surface
$\bar{q}$ denotes the dynamic pressure
Ch_α and Ch_δ denote the hinge moment coefficients of the control surface with respect to angle of attack and surface deflection δ

The aerodynamic load dynamic pressure $\bar{q}$ is a function of airspeed U_0 and air density [2] ρ, as expressed by

$$\bar{q} = \frac{1}{2} \rho U_0^2 \tag{93.22}$$

The air density in a standard atmosphere drops exponentially with increasing altitude. The dynamic pressure $\bar{q}$ increases as the Mach number gets higher and altitude decreases. The hinge moment for modern fighters with all movable tails, where $Ch_\alpha \approx Ch_\delta$, is modeled by

$$\tau_L = \bar{q} S_t R_h Ch_\delta \left(\frac{\alpha(s)}{\theta_L(s)} + 1 \right) \theta_L \tag{93.23}$$

93.4.4.4 Aircraft Short-Period Approximation

If the aircraft forward speed is assumed constant (i.e., the speed perturbations $u \approx 0$), the X force equation can be neglected since it does not significantly contribute to the short-period oscillation [5]. Thereby, the short-period approximation of the longitudinal channel of the aircraft that is relevant to the actuator dynamics is extracted and is written as

$$\begin{bmatrix} \dot{\alpha} \\ \dot{q} \end{bmatrix} = \begin{bmatrix} z_\alpha & z_q \\ m_\alpha & m_q \end{bmatrix} \begin{bmatrix} \alpha \\ q \end{bmatrix} + \begin{bmatrix} z_\delta \\ M_\delta \end{bmatrix} [\delta_e] \tag{93.24}$$

This yields a second-order minimum-phase transfer function of the form

$$\frac{\alpha(s)}{\delta(s)} = \frac{-K(s+a)}{(s+b)(s+c)} \tag{93.25}$$

For the frequency of interest, the short-period approximation closely resembles the full state model. As seen in Figure 93.9, where the solid lines represent the full state model frequency response of the vehicle and + lines represent the short-period model frequency response of the vehicle, phase and attenuation characteristics are closely matched at high frequency. The approximation, as seen in the figure, is not valid for frequencies below 0.5 rad/s due to the effects from the slow longitudinal mode, which is referred to as the phugoid.

93.4.5 Quantitative Feedback Theory (QFT)

Robust quantitative feedback theory (QFT) [3,6] compensators are ideal for obtaining the desired response from an actuator. The term *robust* in control theory implies that a system under control, in this case the actuator, remains stable throughout its operating envelope, rejects disturbances, and results in minimum degradation in the performance specifications. A QFT-designed feedback system, which uses

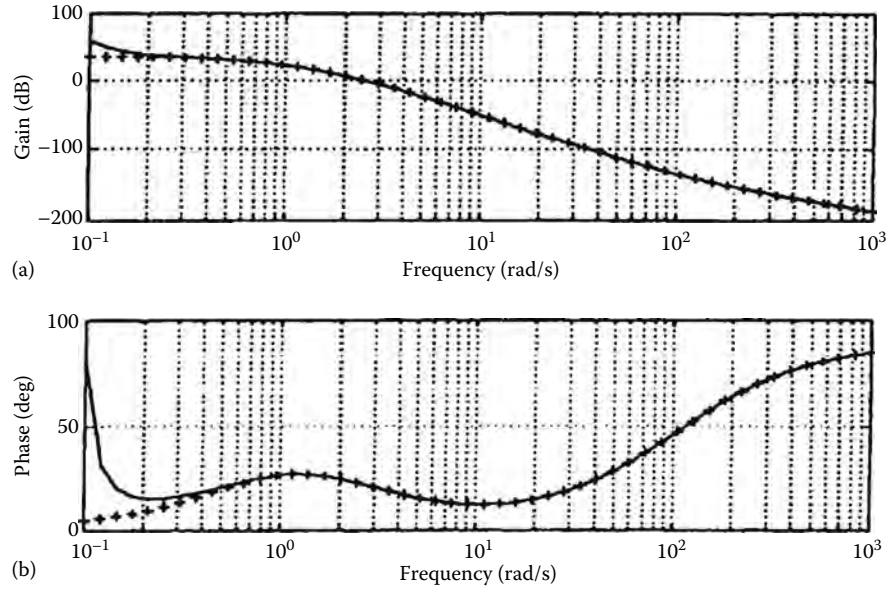


FIGURE 93.9 Comparison of full- and short-period approximation for longitudinal equations of motions: (a) magnitude response and (b) phase response.

output feedback, assures that the output tracks the input values despite parametric uncertainties or disturbances. The output of a unity-gain feedback system with uncertain plant models can vary depending on the plant conditions; however, the output of a robust QFT feedback control system with uncertain plant models will not vary much. The unity-gain system, however, cannot control the tracking response, which necessitates the inclusion of a prefilter.

Plant (specifically, actuator) uncertainties may be caused by manufacturing tolerances. For example, a component A is specified to be 10 ± 0.1 mm long. In a large production run, the length of component A may be anywhere between 9.9 and 10.1 mm. This tolerance range may be substantial enough to produce noticeable variations in the plant model. Plant uncertainties may also be caused by the environment, specifically the system (actuator) operating conditions. For example, the parameters of an aircraft model display large variations depending on altitude and airspeed. Finally, the performances of various components decay over time, introducing an additional element of uncertainty.

The QFT design paradigm accounts for the plant variations in the design procedure. It is a linear design technique for designing a linear robust controller for linear or nonlinear control systems [3,6]—in this case, the actuator control system. As long as all the QFT design requirements are met, the output responses are guaranteed to conform to the specification boundaries. The initial step in the design process is to establish the upper and lower performance tracking boundaries. The maximum allowable disturbance is used to determine the disturbance boundary. The phase margin angle, gain margin, or maximum peak value is used to establish the stability boundary. These combined boundaries establish the limitation of the system performance. The next step is to analyze the system to determine the parameters that will cause noticeable variations to the plant. The system, often modeled only as a single plant transfer function in other design methods, is now modeled more comprehensively as a *set* of plant transfer functions. The variations now form a closed region in the Nichols magnitude and phase chart, called the template; larger template sizes indicate a higher degree of plant uncertainty. Templates at different frequencies differ in size and shape.

The most important aspect of modeling is the thorough analysis of the bare actuator plant, in order to find the sources of plant variation. For an aircraft application, the three largest sources of variations are the fluctuations in aerodynamic load on the flight control surfaces, the motor torque, and the hydraulic pump fluid pressure.

93.5 Actuator Compensator Design

QFT compensators, which are designed in the frequency domain, are ideally suited for the actuator control system. The QFT compensator processes its inputs, which are the commanded actuator displacement and the measured current actuator position, and generates a commanded signal to, e.g., the solenoid valve in a hydraulic actuator or a current for the electric DC motor in the EHA actuator. In order to proceed with the actuator QFT compensator design, the block diagram as shown in Figure 93.7 is transformed to the standard inner and outer QFT loop structures shown in Figure 93.10 [2,3,6]. For this design problem, the external disturbance $d_2(t)$ can be ignored, i.e., $d_2(t) = 0$, whereas the $d_1(t)$ disturbance models a possible measurement bias. The inner loop controls the angular rate of the motor, while

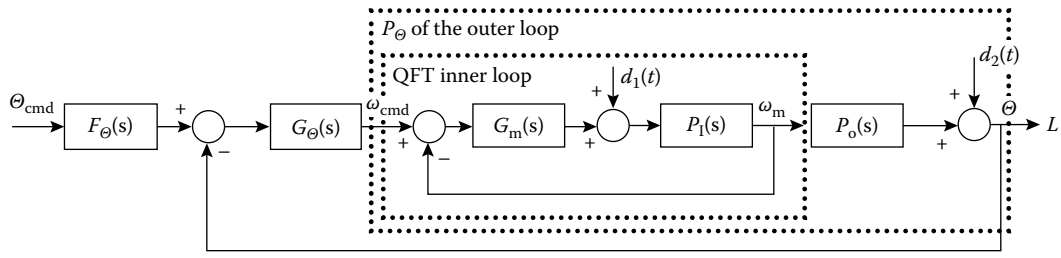


FIGURE 93.10 Two feedback loops in QFT two degrees of freedom structure.

the outer loop controls the surface deflection. Since tracking in the inner loop is less important, for its function is to reduce the uncertainty level of the outer loop, the inner loop prefilter can be set to unity, and the tracking performance enforcement is relegated to the prefilter F_θ of the outer loop.

93.6 Conclusion

The modeling of a hierarchy of hydraulic actuators has been presented. The modeling of modern EHAs is emphasized, and dynamic models of its subcomponents are carefully developed. A systems approach to actuator modeling and design is taken, and the feedback structure is highlighted. In this chapter, it is stressed that because of the complex feedback interactions at work, proper actuator modeling requires consideration of the latter in the full context of the control system in which it is to be employed. Thus, in this presentation, the aerospace application is considered. At the same time, the actuator is being viewed not as a component of a control system but as a control system in its own right. Hence, it is shown that simple mechanical feedback linkages that fulfilled the role of a controlling element for the actuator can be replaced by a more flexible, high-performance, robust, full-blown electronic actuator compensator. Concerning electropneumatic instruments and sensors, these are critical elements in systems where the flow of liquids and/or gases is being controlled, e.g., in the chemical industries. In aerospace, electropneumatic sensors are used to measure airspeed, altitude, and angle of attack and sideslip angles of the aircraft; these functions are collocated in the air data computer (ADC). Currently, modern sensors use feedback action for accuracy enhancement, and as such, the design of modern electropneumatic instruments and sensors is conceptually similar to the design of EHAs. Detailed information on actuators and sensors may be found in Refs. [4,7].

Nomenclature

A	ram cross-sectional area
a,b,c,d,e,R_h	length of linkages
B_p	piston damping
C_h	hinge moment coefficient
D_m	pump displacement constant
F	force
J	moment of inertia
K	gain
K_h	hinge stiffness constant
M_p	mass of piston
M,z,m	aerodynamic stability derivatives
$\wp$	differential pressure developed by the pump
P	pressure
$\overline{q}$	dynamic pressure
q	pitch rate
Q	flow rate
s	Laplace variable
S_t	aerodynamic surface area
U,u	airspeed
x,X	displacement
α	angle of attack of aircraft
δ	deflection angle
Ω, ω	rotor speed
ρ	air density
θ	deflection angle
τ	torque

References

1. O.R. Reynolds, M. Pachter, and C.H. Houpis, Full envelope flight control system design using quantitative feedback theory, *AIAA J. Guidance Control Dyn*., 19, 1023–1029, 1996.
2. K.H. Kang, Electro-hydrostatic actuator controller design using quantitative feedback theory, MS thesis, Graduate School of Engineering, Air Force Institute of Technology, Wright-Patterson AFB, OH, December 1994, AFIT/GE/ENG/94D-18.
3. C.H. Houpis, M. Pachter, S. Rasmussen, and R. Sating, Quantitative feedback theory for the engineer, Wright Laboratory Technical Report, WL-TR-95-3061, Wright Laboratory, Wright-Patterson AFB, OH, June 1995. (Available from the National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22151, document number AD-A297574.)
4. C.W. de Silva, *Control Sensors and Actuators*, Englewood Cliffs, NJ: Prentice-Hall, 1989.
5. J.H. Blakelock, *Automatic Control of Aircraft and Missiles*, New York: John Wiley & Sons, 1991.
6. J.J. D'Azzo and C.H. Houpis, *Linear Control System Analysis and Design—Conventional and Modern*, 4th edn., New York: McGraw-Hill, 1988.
7. E.H.J. Pallet and S. Coyle, *Automatic Flight Control*, Oxford, U.K.: Blackwell, 1993.

Partial List of Vendors and Suppliers

ControlAir, Inc., 8 Columbia Dr., Amherst, NH 03031-2315, Phone: 603-886-9400, 800-216-3636 (toll free), Fax: 603-889-1844, www.controlair.com

Equilibar, 320 Rutledge Rd., Fletcher, NC 28732, Phone: 828-650-6590, Fax: 801-504-4439, www.equilibar.com

Nexen Group, Inc., 560 Oak Grove Pkwy., Vadnais Heights, MN 55127-8509, Phone: 651-484-5900, 800-843-7445 (toll free), Fax: 651-286-1099, www.nexengroup.com/nexen/index.jsp

Ross Controls, 32900 N. Avis Dr., Madison Heights, MI 48071, Phone: 248-397-1958, 800-438-7677 (toll free), commerce.rosscontrols.com

94

Explosion-Proof Instruments

94.1 Introduction **94**-1
94.2 Fundamentals of Explosion Protection **94**-1
94.3 Classification of Hazardous Areas **94**-3
94.4 Enclosure Types and Requirements **94**-5
94.5 Protection Methodologies **94**-6
Intrinsic Safety • Making Field Devices Intrinsically Safe • Ignition Curves • Certification and Approval • Explosion-Proof Fundamentals • Purging and Pressurization
Defining Terms **94**-16
References **94**-16
Further Readings **94**-17
Partial List of Manufacturers and Suppliers **94**-17

Sam S. Khalilieh
Earth Tech

94.1 Introduction

Where hazardous atmospheres can exist, electricity should be a primary concern of every engineer and system designer. Hazardous atmospheres can exist not only in the more common surroundings of industrial, chemical, and environmental facilities but also in many less obvious environs where dust is present, where gas can accumulate, and where combustible gas-forming reactions occur. To minimize risks in such areas, it is necessary to design specific hazard-reducing electric systems. Most electric equipment is built to specific standards aimed to reduce the incidence of fires and human casualties. The majority of such incidents can be attributed to poor or defective installations, improper use of approved equipment, deteriorated equipment, and accidental applications. In combination with an explosive atmosphere, these factors can result in extremely dangerous conditions. Designing an electric system for a hazardous location requires careful planning, research, engineering, and ingenuity in using proper protection techniques to develop better applications and classifications that reduce hazards.

94.2 Fundamentals of Explosion Protection

The safety of personnel and equipment present in a hazardous area should never be taken for granted. In 1913, a massive methane gas explosion in a coal mine in Glamorganshire, South Wales, claimed the lives of 439 mine workers. After months of research and studies, a group of experts and scientists concluded that the explosion was caused by a small amount of electric energy stored in the circuit. This small amount of energy combined with the presence of an explosive gas and air mixture and the absence of proper protection proved to be fatal for the mine workers.

To understand the dangers associated with electric equipment in hazardous areas, one must first understand the basics. Chemically speaking, oxidation, combustion, and explosions are all exothermic reactions where heat is given off at different reaction speeds. For these reactions to occur, three components must be present simultaneously in certain concentrations. These components are (1) fuel (liquid, gas, or solids), (2) a sufficient amount of oxygen (air), and (3) an ignition source (electric or thermal).

Some of the ignition sources that can be potentially hazardous include (1) hot surfaces (motor windings, heat trace cable, light fixtures), (2) electric sparks and arcs (when circuits are opened and closed, short circuits), (3) mechanical sparks (friction, grinding), (4) electrostatic discharge (separation process in which at least one chargeable substance is present), and (5) radiation, compression, and shock waves.

When dealing with electric equipment in hazardous locations, it is important to understand and to be familiar with the following terms:

1. *Flash point*—the minimum temperature at normal air pressure at which a combustible or flammable material releases sufficient vapors ignitable by an energy source. Depending on the flash point (FP), flammable liquids are divided into four classes of hazard:
   a. AI (FP < 21 °C)
   b. AII (21 °C < FP < 55 °C)
   c. AIII (55 °C < FP < 100 °C)
   d. B (FP < 21 °C at 15 °C dissolving in water)
2. *Ignition temperature*—the minimum temperature under normal operating pressure at which a dangerous mixture ignites independently of the heating or heated element.
3. *Flammable limits*—the upper explosive limit (UEL) or the maximum concentration ratio of vapor to air mixture above which the propagation of flame does not occur when exposed to an ignition source. Here, the mixture is said to be "too rich" to burn. The lower explosive limit (LEL) is the minimum concentration ratio of vapor to air mixture below which the propagation of flame does not occur when exposed to an ignition source. Also, here, the mixture is said to be "too lean" to explode. Significant attention must be given to LEL, since it provides the minimum quantity of gas necessary to create a hazardous mixture. Generally, the flammable limits are indicated in percent by volume, which is abbreviated % vol. Note that the explosion of a mixture in the middle the UEL and the LEL is much more violent than if the mixture were closer to either limit.
4. *Maximum surface temperature*—the maximum temperature generated by a piece of electric equipment under normal or fault conditions. This temperature must be below the minimum ignition temperature of the potentially explosive surrounding atmosphere. Equipment used in hazardous locations must be clearly marked to indicate class, group, and maximum surface temperature or range referenced to 40 °C (104 °F) ambient temperature. Table 94.1 shows that an apparatus with a specific T class can be used in the presence of all gases having an ignition temperature higher than the T temperature class of the device. For added safety, it is recommended that the maximum surface temperature be not more than 80% of the minimum ignition temperature of the surrounding gas. The reader is cautioned not to confuse maximum working (operating) temperature with maximum surface temperature, which is measured under worst-case conditions of the electric apparatus. An electric apparatus designed to operate with a maximum ambient temperature of 70 °C—even in the worst conditions of the expected temperature range—must not have a temperature rise greater than a safety margin of 10 °C to be classified as T6 or 5 °C for classes T3, T4, and T5 (Table 94.1).
5. *Vapor density*—the weight of a volume of pure vapor gas compared with the weight of an equal volume of dry air under the same normal atmospheric pressure and temperature. It is calculated as the ratio of molecular weight of the gas to the average molecular weight of air (28.96). Methane gas (CH_4) with molecular weight of 16 and vapor density of 0.6 tends to rise, while acetone (C_3H_6O) with molecular weight of 58 and vapor density of 2 tends to settle closer to ground levels.

In the United States, the National Electrical Code (NEC) defines a hazardous area as "an area where a potential hazard may exist under normal or abnormal conditions because of the presence of flammable, combustible,

TABLE 94.1 Maximum Surface Temperature under All Operating Conditions

Maximum Temperature		
°C	°F	Identification Number
450	842	T1
300	572	T2
280	536	T2A
260	500	T2B
230	446	T2C
215	419	T2D
200	392	T3
180	356	T3A
165	329	T2B
160	320	T3C
135	275	T4
120	248	T4A
100	212	T5
85	185	T6

Source: Reprinted with permission from NFPA 70-1996, the *National Electrical Code®*, Copyright 1995, NFPA, Quincy, MA 02269. This reprinted material is not the complete and official position of the NFPA, on the referenced subject that is represented only by the standard in its entirety.

Note: Surface temperature of electric apparatus during operation must not exceed limitations of the hazard present.

or ignitable materials" [1]. This general description is divided into different classes, divisions, and groups to assess the extent of the hazard properly and to design and specify safe operating electric systems.

The need for classification is important not only for safety but for economic reasons as well. Proper application, good engineering, and experience can reduce the extent of the most volatile areas (Class I, Division 1) within reasonably safe distances of potential leaks and ignition sources. Under Class I, Division 1, equipment and installation costs can become an economic burden because the equipment is considerably more expensive and must pass stringent tests to ensure proper and safe operation under normal or abnormal conditions. The National Fire Protection Association (NFPA 497 A and B) [2] and the American Petroleum Institute *Recommended Practice for Classification of Locations for Electrical Installations at Petroleum Facilities* (ANSI/API RP 500) [3] are excellent resources for defining hazardous area boundaries.

94.3 Classification of Hazardous Areas

Classification of a hazardous area within a facility is usually determined by highly qualified personnel including chemical engineers, process engineers, and safety officers. Their primary objective is to determine where a potentially hazardous atmosphere exists, under what conditions it exists, and how long it exists. Careful study and design of electric installations, especially in hazardous areas, are crucial for the safe operation of electric equipment and prevention of an accidental ignition of flammable materials. The NEC, which has been adopted by many states, agencies, and companies as the basis for inspections, describes the requirements and procedures for electric installations in hazardous areas. Articles 500–504 contain the requirements of electric equipment and wiring for all voltages in locations where fire or explosion hazards may exist due to flammable gases or vapors, flammable liquids, combustible dust, or ignitable fibers or flyings.

TABLE 94.2 Area Classification Based on NEC

Division 1—Hazard Is Present under Normal Operating Conditions	
Class I	Gases and Vapor
Group A	Acetylene
Group B	Hydrogen
Group C	e.g., ethylene
Group D	e.g., methane
Class II	Combustible Dusts
Group E	Metal dust
Group F	Coal dust
Group G	Grain dust
Class III	Fibers
Division 2—Hazard Is Present Only under Abnormal Operating Conditions	

Table 94.2 describes hazardous locations by class, division, and group. The class defines the physical form of combustible material mixed with oxygen molecules. The division defines the probability of an explosive fuel to air mixture being present. The group indicates the type of vapor or dust present. The NEC gives the following definitions [1]:*

Class I, Division 1 Locations

1. Where ignitable concentrations of flammable gases or vapors can exist under normal operation conditions, may exist frequently because of repair or maintenance operations or because of leakage, and where breakdown or faulty operation of equipment or processes might release ignitable concentrations of flammable gases or vapors and cause simultaneous failure of electric equipment.

Class I, Division 2 Locations

1. Where volatile flammable liquids or flammable gases are handled, processed, or used but where the liquids, vapors, or gases will normally be confined within closed containers or closed systems from which they can escape only in case of accidental rupture or breakdown of such containers or systems or in case of abnormal operation of equipment.
2. Where ignitable concentrations of gases or vapors are normally prevented by positive mechanical ventilation and where they might become hazardous through failure or abnormal operation of the ventilating equipment.
3. Adjacent to Class I, Division 1 locations, and where ignitable concentrations of gases or vapors might occasionally be communicated; unless such communication is prevented by positive-pressure ventilation from a source of clean air, and effective safeguards against ventilation failure are provided.

Class II, Division 1 Locations

1. Where combustible dust is in the air under normal operation conditions in quantities sufficient to produce explosive or ignitable mixtures.
2. Where mechanical failure or abnormal operation of machinery or equipment might cause such explosive or ignitable mixtures to be produced and also might provide a source of ignition through simultaneous failure of electric equipment, operation of protective devices, or from other causes.
3. Where combustible dusts of an electrically conductive nature may be present in hazardous quantities.

* Reprinted with permission from NFPA 70-1996, the *National Electrical Code*®,© 1995, NFPA, Quincy, MA. This reprinted material is not the complete and official position of the NFPA, on the referenced subject that is represented only by the standard in its entirety. *National Electrical Code*® and *NEC*® are registered trademarks of the NFPA, Inc., Quincy, MA.

Class II, Division 2 Locations
Where combustible dust normally is not in the air in quantities sufficient to produce explosive or ignitable mixtures, and dust accumulations normally are insufficient to interfere with the safe dissipation of heat from electric equipment or may be ignitable by abnormal operation or failure of electric equipment.

Class III, Division 1 Locations
Where easily ignitable fibers or materials producing combustible flyings are handled, manufactured, or used.

Class III, Division 2 Locations
Where easily ignitable fibers are stored or handled. Quantities and properties of hazardous materials are the basis upon which the NEC classifies hazardous locations. Each hazardous location must be evaluated carefully to determine the appropriate classification to facilitate the design process and to help specify the correct equipment.

94.4 Enclosure Types and Requirements

Choosing the proper type of enclosure for electric equipment is important for two reasons:

1. Personnel protection against accidental contact with enclosed electric equipment
2. Protection of internal equipment against outside harm

Enclosures are designated by a type number indicating the degree of protection and the condition for which they are suitable. In some applications, enclosures have a dual purpose and therefore are designated by a two-part-type number shown with the smaller number first (i.e., 7/9). The following enclosure types, with their enclosed equipment, have been evaluated in accordance with Underwriters Laboratories, Inc. UL 698, *Industrial Control Equipment for Use in Hazardous Locations*, and are marked to show the class and group letter designations.

Type 7 Enclosures. Type 7 enclosures are nonventilated, intended for indoor applications, and classified for Class I, Group A, B, C, or D, as defined in Table 94.2. The letters A to D sometimes appear as a suffix to the designation Type 7 to give the complete designation. According to UL 698, Type 7 enclosures must be designed to withstand an internal explosion pressure of specific gases and to prevent such an explosion from igniting a hazardous mixture outside the enclosure (explosion test). In addition, Type 7 enclosures fabricated from sheet steel are designed to withstand two times the internal explosion pressure for 1 min without permanent deformation and three times the explosion pressure without rupture. If constructed of cast iron, the enclosure must be capable of withstanding four times the explosion pressure without rupture or deformation. This test may be waived if calculations show a safety factor of five to one for cast metal or four to one for fabricated steel. The enclosed heat-generating devices are specifically designed to prevent external surfaces from reaching temperatures capable of igniting explosive vapor–air mixture outside the enclosure (temperature test).

Type 8 Enclosures. Type 8 enclosures are nonventilated, intended for indoor applications, and intended for Class I, Group A, B, C, or D, as outlined in Table 94.2. The letters A to D appear as a suffix to the designation Type 8 to give the complete designation. According to UL 698, the oil-immersed equipment must be able to operate at rated voltage and most severe current conditions in the presence of flammable gas–air mixtures without igniting these mixtures.

Type 9 Enclosures. Type 9 enclosures are nonventilated, intended for indoor applications, and classified for Class II, Group E, F, or G as outlined in Table 94.2. The letters E, F, or G appear as a suffix to the designation Type 9 to give the complete designation. According to UL 698, the enclosure with its enclosed equipment is evaluated in accordance with UL 698 in effect at the time of manufacture. This evaluation includes a review of dimensional requirements for shaft opening and joints, gaskets material,

and temperature rise under a blanket of dust. The device is operated at full rated load until equilibrium temperatures are reached and then allowed to cool to ambient temperature over a period of at least 30 h while continuously subjected to circulating dust of specified properties. No dust shall enter the enclosure (dust penetration test). Furthermore, Type 9 enclosures must also pass the "temperature test with dust blanket," which is similar to the temperature rise test except the circulating dust is not aimed directly at the device during testing. The dust in contact with the enclosure shall not ignite or discolor from heat, and the exterior surface temperature based on 40 °C (104 °F) shall not exceed specific temperatures under normal or abnormal conditions. Where gasketed enclosures are used, gaskets shall be of a noncombustible, nondeteriorating, vermin-proof material and shall be mechanically attached. Type 9 ventilated enclosures are the same as nonventilated enclosures, except that ventilation is provided by forced air from a source outside the hazardous area to produce positive pressure within the enclosure. The enclosure must also meet the requirements of a temperature design test.

Type 10 Enclosures. Type 10 enclosures are nonventilated and designed to meet the requirements of the U.S. Bureau of Mines that relate to atmospheres containing mixtures of methane and air, with or without coal dust present.

It is important to note that enclosures for hazardous applications are designed for specific applications and must be installed and maintained as recommended by the enclosure manufacturer, since any misapplication or alteration to the enclosure may jeopardize its integrity and may eventually cause catastrophic failure of the system. All enclosures should be solidly grounded and properly labeled with a warning sign reminding the operator of the importance of de-energizing the incoming power to the enclosure prior to its servicing.

94.5 Protection Methodologies

Choosing a protection technique that suits each application can appear complicated because safety, reliability, cost, and maintenance factors must all be considered. Over the years, few hazardous area safety protection methodologies have been used. Although methodologies differ in application and principles of operation, they all have one thing in common: to eliminate one or more components necessary for combustion. Three of the most widely used methodologies are

1. Intrinsic safety (IS)
2. Explosion proof
3. Purging and pressurization

94.5.1 Intrinsic Safety

Simply stated, IS is all about preventing explosions. IS is based on the principle of limiting the thermal and electric energy levels in the hazardous area to levels that cannot cause an ignition of a specific hazardous mixture in its most ignitable concentration. IS pertains to the minimum ignition temperature and the minimum ignition electric energy required to cause a specific group to ignite. The energy level provided by an IS circuit is low (≈1 W) and is used only to power up instruments with a low energy demand. An IS circuit incorporates an intrinsically safe apparatus (field device), an associated apparatus, and an interconnecting wiring system. Designing intrinsically safe systems begins with studying the field device. This will help determine the type of associated apparatus that can be used so that the circuit functions properly under normal operating conditions but still safe under fault conditions. Field devices can be simple, such as *resistance temperature devices* (*RTDs*), thermocouples, mechanical switches, proximity switches, and light-emitting diodes (LEDs), or they can be nonsimple, such as transmitters, solenoid valves, and relays. A field device is considered and recognized as a "simple device" if its energy storing or generating values do not exceed 1.2 V, 0.1 A, and 25 mW (or 20 μJ) in an intrinsically safe system under normal or abnormal conditions.

TABLE 94.3 Comparison of Entity Values of a Field Device and a Safety Barrier

Field Device	Safety Barrier
$V_{max} \geq V_{oc}$	
$I_{max} \geq I_{sc}$	
$C_i \leq C_a$	(Maximum allowed capacitance)
$L_i \leq L_a$	(Maximum allowed inductance)

Source: Reprinted with permission from NFPA 70-1996, the *National Electrical Code*®, Copyright 1995, NFPA, Quincy, MA 02269. This reprinted material is not the complete and official position of the NFPA, on the referenced subject that is represented only by the standard in its entirety. *National Electrical Code*® and *NEC*® are registered trademarks of the National Fire Protection Association, Inc., Quincy, MA 02269.

The simple device may be connected to an intrinsically safe circuit without further certification or approval. The fact that these devices do not have the ability to store or generate high levels of energy does not mean they can be installed in a hazardous area without modification. They must always be used with an associated apparatus to limit the amount of energy in the hazardous area, since a fault outside the hazardous area can cause sufficient high levels of energy to leak into the hazardous area. A nonsimple device (i.e., relay, transmitter) is capable of generating and storing energy levels exceeding the aforementioned values. Such devices require evaluation and approval under the entity concept (described later) to be used in conjunction with an intrinsically safe circuit. Under the entity concept, these devices have the following entity parameters: V_{max} (maximum voltage allowed), I_{max} (maximum current allowed), C_i (internal capacitance), and L_i (internal inductance). Under fault conditions, voltage and current must be kept below the V_{max} and I_{max} of the apparatus to prevent any excess heat or spark, which can be disastrous in hazardous areas. C_i and L_i indicate the ability of a device to store energy in the form of internal capacitance and internal inductance, and their value must be less than C_a and L_a of the associated apparatus (Table 94.3).

An associated apparatus (Figure 94.1), also known as a safety barrier, is an energy-limiting device needed to protect a field device located in a hazardous area from receiving excessive voltage or current. An associated apparatus is normally installed in a dust- and moisture-free enclosure (NEMA 4) located in a nonhazardous area, as close as possible to the hazardous area to minimize the capacitance effect of

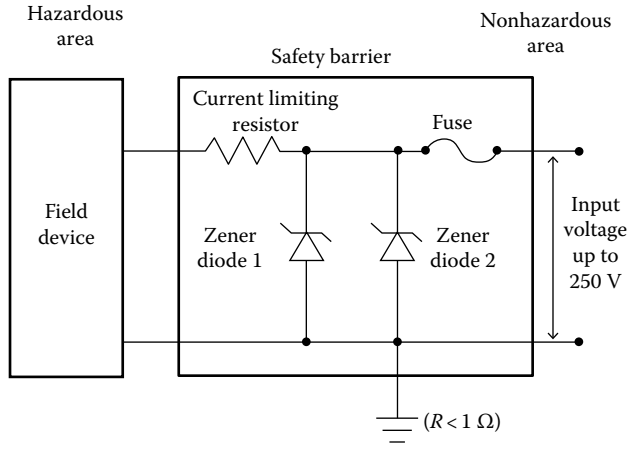


FIGURE 94.1 Major components of barrier circuit.

the cable. If installed in a hazardous area, the associated apparatus must be installed in an explosion-proof enclosure (i.e., NEMA 7D).

Figure 94.1 shows the three major components of a Zener safety barrier. (Note that there are other types of barriers such as isolation and repeater types.) The components are

1. The resistor, which limits the current to a specific value known as short circuit current (I_{sc}).
2. The fuse, which acts as an interrupter or protective device in case of a diode failure (fuse will blow if diode conducts).
3. The *Zener diode*, which limits the voltage to a specific value known as open-circuit voltage (V_{oc}). Zener diodes are unique in their ability to conduct current under reverse bias conditions. When voltage is applied to the Zener diode in the reverse direction, a small amount of current known as leakage current is passed through. This current remains small until the bias voltage exceeds the Zener breakdown voltage. Exceeding the breakdown voltage causes the inherently high resistance of the Zener diode to drop to a very low value, thus allowing the current to increase abruptly. This sudden current increase forces the Zener diode to become a conductor, thereby diverting the excess voltage to ground. If the current continues to flow above and beyond the fuse rating, the fuse will open and the circuit will be completely interrupted. Most safety barriers incorporate at least two diodes in parallel to provide maximum protection in case one diode fails (redundant safety).

In 1988, ANSI/UL 913 [5] allowed the use of IS barriers with replaceable fuses as follows: "if it is accessible for replacement, and the fuse on a fuse protected shunt diode barrier shall not be replaceable by one of higher rating." The fuses are housed in tamper-proof assemblies to prevent confusion or misapplication. The diodes have specific power ratings that must not be exceeded. The Zener diodes and fuses are governed by a very specific set of parameters that allow the fuse to operate at one-third the power rating of the Zener diode and to avoid irreversible damage to the Zener diode. The power rating for the Zener diode can be determined as follows:

$$Z_w = 1.5 \times V_{oc} \times 2 \times I_f$$

where

Z_w is the minimum power rating of the Zener diode
V_{oc} is the maximum Zener diode open-circuit voltage
I_f is the fuse current rating

Selecting the best barrier for the application depends on the field device and requires analysis to ensure proper operation of the intrinsically safe circuit under normal or abnormal conditions. Three of the more important characteristics requiring examination are (1) internal resistance, (2) rated voltage, and (3) circuit polarity. Regardless of the selected barrier, each has an internal resistance (R_i) that limits the short-circuit current under fault conditions. As current passes through R_i, it creates a voltage drop across the barrier that must be accounted for ($V = IR$). The rated voltage of the safety barrier must be equal to or reasonably greater than the supply voltage. The word *reasonably* is significant because excessive supply voltage can cause the diode to conduct, rushing high current through the fuse and blowing it. The use of a regulated power supply can significantly reduce problems associated with excessive supply voltage. To complete an analysis, the circuit polarity must be established. While AC barriers can be connected with either positive or negative power supply, DC barriers can be rated to either positive or negative.

94.5.2 Making Field Devices Intrinsically Safe

RTDs and thermocouples can be made intrinsically safe by using isolated temperature converters (ITCs) that convert a low DC signal from the field device into a proportional 4–20 mA signal. These ITCs require no ground connection for the safe and proper operation of the IS circuit. Because of their ability to store

energy, transmitters are considered nonsimple devices and must be approved as intrinsically safe. If they are third-party approved, their entity parameter must be carefully considered. Transmitters (4–20 mA) convert physical measurements in hazardous areas, such as pressure and flow, into electric signals that can be transmitted to a controller in a safe area. Depending upon the conditions, 4–20 mA signals can be made intrinsically safe by using a repeater barrier that duplicates the output signal to match the input signal. Repeaters can be grounded or ungrounded. Ungrounded repeater barriers are known as "transformer-isolated barriers," since the incoming voltage or signal is completely isolated from the outgoing voltage or signal via a transformer. Digital inputs, such as mechanical and proximity switches, which are simple devices, can be made intrinsically safe by using a switch amplifier. A switch amplifier is simply a relay or an optocoupler (a high-speed relay that uses optical isolation between the input and the output) that transfers a discrete signal (i.e., on/off) from the hazardous area to a safe area. Grounded safety barriers are passive devices designed specifically to prevent excessive energy in a nonhazardous area from reaching a hazardous area. These barriers can be used with most field devices. In order for such barriers to function properly, there is emphatic need for a solid, low-impedance (<1 Ω) connection to ground to prevent ground loops and induced voltages that can hinder operation of the system.

94.5.3 Ignition Curves

All electric circuits possess certain electric characteristics that can be classified under three categories: resistance, inductance, and capacitance. To some extent, all circuits possess these three characteristics. However, some of these characteristics may be so small that their effects are negligible compared with that of the others, thus the terms resistive, inductive, and capacitive circuit. Since the concept of IS is based on the principle that a large electric current can cause an explosion in a hazardous area and the lack of it cannot, it is necessary to identify the ranges of currents that are safe and those that are dangerous.

What is a dangerous amount of electric energy? The answer lies in the ignition curves. Ignition curves are published in most IS standards, such as ANSI/UL 913 [5]. Three of the most referenced curves are shown in Figure 94.2. The curves show the amount of energy required to ignite various hazardous atmospheric mixtures in their most easily ignitable concentration. The most easily ignitable concentration is determined by calculating the percentage of volume to air between UELs and LELs of a specific hazardous atmospheric mixture. In the three referenced curves, the energy levels (voltage and current) below the group curve are not sufficient to cause an ignition of the referenced group.

Since specific ignition temperature is directly related to the amount of voltage and current consumed, both V_{oc} and I_{sc} of the safety barrier must be less than V_{max} and I_{max}. When designing an intrinsically safe system, the cable resistance R (Ω/m), the inductance L (μH/m), and the capacitance C (pF/m), which are inherently distributed over the length of the cable, must be considered. The capacitance and inductance can be readily obtained from the cable manufacturer's literature. If these parameters are not available, certain default values can be used based on NFPA 493/78 (A-4-2). They are $C_c = 200$ pF/m (60 pF/ft), $L_c = 0.66$ μH/m, and (0.2 μH/ft). To determine the maximum wiring distance required to ensure proper operation, the capacitance and inductance must be calculated. One common approach uses "lumped parameters," in which the voltage and current of both the intrinsically safe apparatus and the associated intrinsically safe apparatus are compared and matched according to Equations 94.1 and 94.2. Any deviation from either Equations 94.1 or 94.2 can compromise the integrity of the system and introduce hazardous conditions. The reactive parts of the system must also be considered and verified to demonstrate that C_a and L_a values of the associated apparatus are not be exceeded by the field device and the field wiring values as shown in Equations 94.3 and 94.4. This method, although simple and effective, tends to exaggerate the wiring capacitance and inductance effect, which can be limiting in some applications. Another method takes advantage of the relation between the cable resistance and inductance. This method can be used if the L/R ratio of the associated apparatus is higher than the calculated L/R ratio of the cable. Under these conditions, the lesser D_a (maximum allowed distance) value can be ignored and the cable length can be extended to the

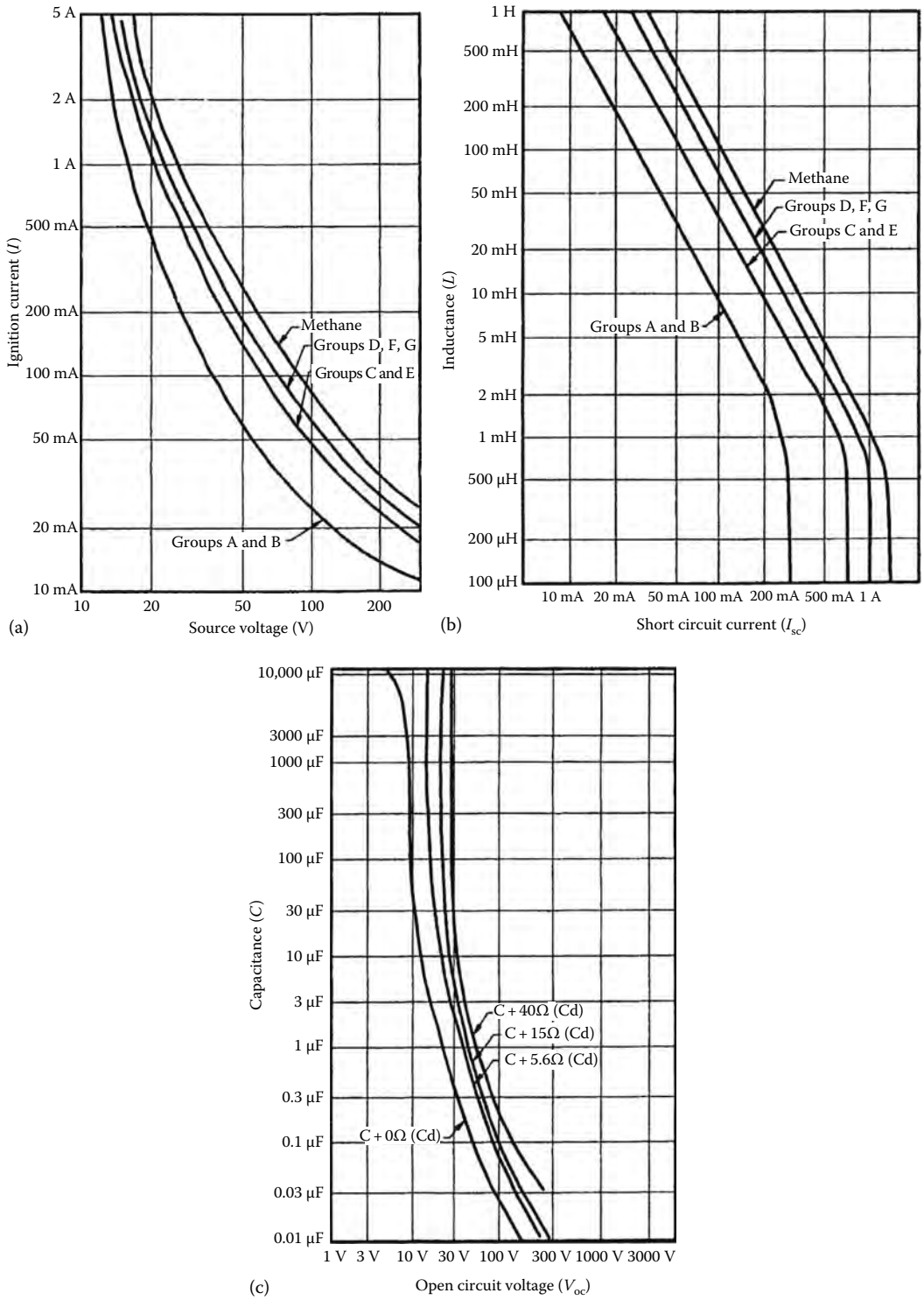


FIGURE 94.2 (a) Resistance circuit ignition curves for all circuit metals, (b) inductance circuit ignition curves at 24 V for all circuit metals, and (c) capacitance circuit ignition curves for Groups A and B for all circuit metals.

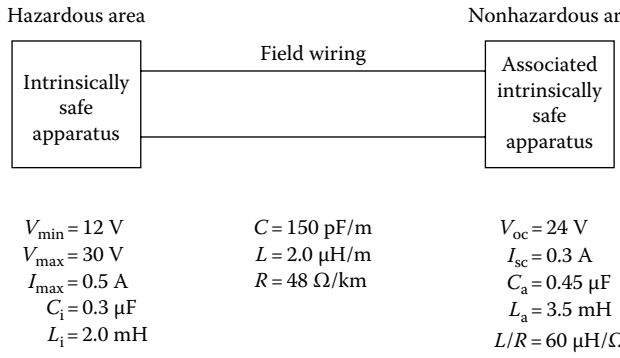


FIGURE 94.3 Analysis of an intrinsically safe system.

higher D_a value. This method is more flexible where cable length is an issue. Figure 94.3 and the following example illustrate these methods.

94.5.3.1 Lumped Parameters Method

$$V_{oc} \leq V_{max} \tag{94.1}$$

$$I_{sc} \leq I_{max} \tag{94.2}$$

$$C_c \leq C_a - C_i \quad 0.45 - 0.30\,\mu F = 0.15\,\mu F \tag{94.3}$$

$$L_c \leq L_a - L_i \quad 3.5 - 2.0\,mH = 1.5\,mH \tag{94.4}$$

The maximum length of the field wiring, referred to its capacitance and inductance, is the lesser value of D_a:

$$D_a = 0.15\,\mu F/150\,pF/m = 1000\,m$$

$$D_a = 1.5\,mH/2.0\,\mu H/m = 750\,m\ (\text{maximum distance of field wiring})$$

94.5.3.2 *L/R* Ratio Method

Since the cable L/R ratio of 41.6 μH/Ω (2 μH/m/48 Ω/km) is less than the given associated apparatus L/R ratio, the inductive effect can be ignored and the maximum distance can be increased to 1000 m.

94.5.4 Certification and Approval

Although approval and certification processes help to provide safety, careful planning, designing, and engineering are still necessary. IS standards, procedures, and tests are recognized worldwide. Testing authorities include Underwriters Laboratories, Inc. (UL), and Factory Mutual Research Corp. (FM) in the United States, Canadian Standards Association (CSA) in Canada, and Physikalisch-Technische Bundesanstalt (PTB) in Europe. Intrinsically safe products are suitable

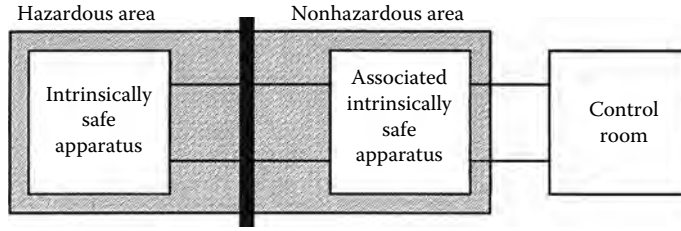


FIGURE 94.4 Loop approval. Intrinsically safe apparatus and associated apparatus are evaluated together. Shaded area indicates evaluated for loop approval.

for all classes, divisions, and groups outlined in Table 94.2. It is necessary to emphasize that the intrinsically safe product must be rated and classified for each specific application (class, division, and group). In the United States, FM adopted two methods for testing and approving equipment to be used in hazardous areas:

1. *Loop* (*system*) *approval*: where an intrinsically safe apparatus is evaluated in combination with a specific associated apparatus and is approved to be installed in this manner. Any changes to the circuit require reevaluation and certification (Figure 94.4).
2. *Entity approval*: where an intrinsically safe apparatus and the associated apparatus are separately evaluated and given their own electric entity parameters (Figure 94.5). The correct application matches the entity parameters shown in Table 94.3. When examining the safety of a circuit, it is crucial to compare the entity values of an intrinsically safe apparatus with an associated apparatus.

Most safety barriers are entity approved for all hazardous locations. Since most field devices have the ability to store energy, they must have loop approval or entity approval for the proper construction and operation of an intrinsically safe system.

IS engineers often advocate the use of intrinsically safe equipment for the following reasons:

1. *Safety*. No explosion can occur in an intrinsically safe system under any operating condition. IS equipment operates on lower power levels and prevents shocks, excess thermal energy, and arcing. In different systems and under various scenarios, shocks, thermal energy, and arcing may cause a hazard.
2. *Reliability*. The components and assemblies of intrinsically safe circuits are tested for reliability before they are labeled and certified. Most intrinsically safe equipment is designed with special circuitry to provide surge suppression and to prevent spikes and transients.
3. *Ease of handling and installation*. Intrinsically safe systems tend to be small and do not require expensive, bulky accessories such as enclosures, seals, and rigid metallic conduits that increase the initial investment.

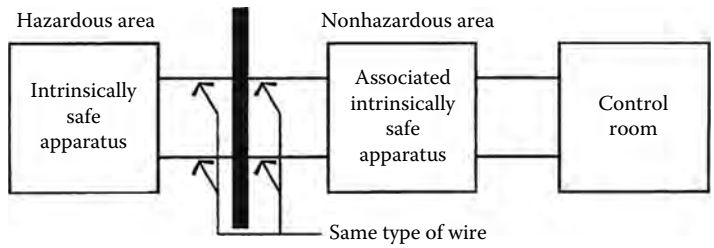


FIGURE 94.5 Entity approval. Intrinsically safe apparatus and associated apparatus are evaluated separately.

4. *Economy*. In some geographic locations, facilities containing hazardous conditions must carry special liability insurance. With the proper installation of intrinsically safe circuits and equipment, the probability of an explosion is 10^{-18} [8], or nearly nonexistent. As a result, insurance rates tend to be lower.
5. *Maintenance*. Equipment may be calibrated and maintained without disconnecting power, thereby resulting in less downtime.

The wiring of intrinsically safe systems is similar to any other application, but to ensure a proper operating system, certain guidelines regarding identification and separation must be strictly followed. All intrinsically safe components including terminal blocks, conductors, and intrinsically safe apparatus must be explicitly marked and labeled. The conventional color used to identify intrinsically safe equipment is blue. In an open wiring installation, intrinsically safe conductors must be physically separated from nonintrinsically safe conductors by at least 50 mm (2 in.), so an induced voltage does not defeat the purpose of IS. Where intrinsically safe conductors occupy a raceway, the raceway should be labeled, "intrinsically safe circuits." Intrinsically safe conductors should not be placed with nonintrinsically safe conductors. Where a cable tray is used, a grounded sheet metal partition may be used as an acceptable means of separation. Where intrinsically safe and nonintrinsically safe conductors occupy the same enclosure, a 50 mm (2 in.) separation must be maintained. In addition, a grounded metal partition shall be in place to prevent contact of any conductors that may come loose. Insulation deterioration of intrinsically safe conductors of different circuits occupying the same raceway or enclosure can be detrimental to the operation of the system. Intrinsically safe conductors must have an insulation grade capable of withstanding an AC test voltage of 550 V root mean square (rms) or twice the operating voltage of the intrinsically safe circuit. Nonintrinsically safe conductors in the same enclosure with intrinsically safe conductors must have an insulation grade capable of withstanding an AC test voltage of $2U + 1000$ V, with a minimum of 1500 V rms, where U is the sum of rms values of the voltages of the intrinsically safe conductors. A commonly used and highly recommended practice utilizes separate compartments for intrinsically safe and nonintrinsically safe conductors. In addition to physical separation of intrinsically safe conductors and nonintrinsically safe conductors, sealing of conduits and raceways housing intrinsically safe conductors is essential to prevent the passage of gases, vapors, and dusts from hazardous to nonhazardous areas. According to the NEC, seal-offs are not required to be explosion proof. Where an associated apparatus is installed in an explosion-proof enclosure in a hazardous area, seal-offs must be explosion proof. Although it is not required by code, it is a good engineering practice to install explosion-proof seal-offs on conduits housing intrinsically safe conductors, as shown in Figure 94.6.



FIGURE 94.6 Explosion-proof seal-off fitting. (Courtesy of Crouse-Hinds Division of Cooper Industries, Inc., Houston, TX.)

94.5.5 Explosion-Proof Fundamentals

Explosion-proof design is a mechanical concept that relies heavily on the mechanical construction of an enclosure and the narrow tolerances between its joints, threads, and flanges to safely contain, cool, and vent any internal explosion that may occur. By definition, explosion-proof enclosures must prevent the ignition of explosive gases or vapors that may surround it (Types 7 and 10 enclosures only). In hazardous areas, Class I, Divisions 1 and 2, arcing devices, such as switches, contactors, and motor starters, must be enclosed in an explosion-proof enclosure specifically rated for that area. Contrary to popular belief, explosion-proof enclosures are not and should not be vapor tight. Succinctly stated, an explosion inside an enclosure must be prevented from starting a larger explosion outside the enclosure. Unlike IS, explosion-proof enclosures address the maximum internal pressure (see NEMA Type 7 enclosures). Figure 94.7 illustrates the rugged construction of a standard explosion-proof panelboard.

In addition to their durability and strength, explosion-proof enclosures must also be "flame tight." The joints or flanges must be held within narrow tolerances to allow cooling of hot gases resulting from internal explosions. In this way, if any gases are released into the outside hazardous atmosphere, they are cool enough not to cause ignition.

Explosion-proof enclosures tend to be bulky (making them easy to identify) and heavy, requiring conduit seals and careful handling. Unlike intrinsically safe equipment, explosion-proof equipment operates on normal power levels that are necessary due to the high power requirements of some circuits and equipment. With the proper equipment, installation, and maintenance, explosion-proof enclosures can safely and effectively distribute high levels of voltage and power into hazardous areas.

Where ignitable amounts of dust are present, enclosures housing electric equipment must be dust-ignition proof. These enclosures must exclude combustible dusts from entering while preventing arcs, sparks, or heat generated internally from igniting dust surrounding the exterior of the enclosure. These enclosures must also efficiently dissipate the heat generated internally, since many types of dust will ignite at relatively low temperatures. Unlike Class I, Division 1 explosion-proof enclosures (Type 7), Class II,

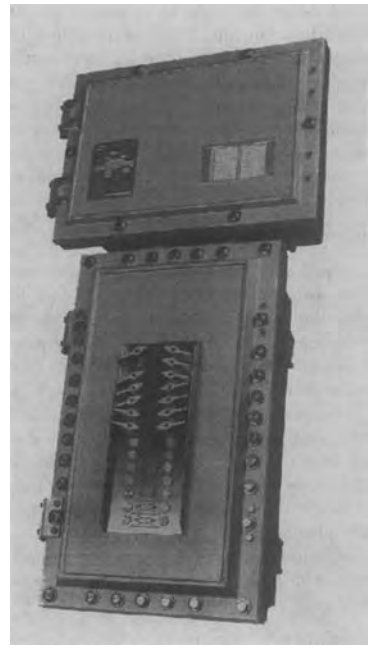


FIGURE 94.7 Explosion-proof lighting panelboard. (Courtesy of Crouse-Hinds Division of Cooper Industries, Inc., Houston, TX.)

Division 1 dust-ignition-proof enclosures (Type 9) are designed to prevent an explosion. Subsequently, dust-ignition-proof enclosures need not be as strong or have walls as thick as explosion-proof enclosures since there will be no internal explosion.

94.5.6 Purging and Pressurization

This methodology allows the safe operation of electric equipment where hazardous conditions exist and where no other methodology is applicable because of the imperative high-energy demands and actual physical dimensions. This is true for large-sized motors and switchgear units where they are not commercially available for Class I, Groups A and B. In addition, this methodology is used where control panels that house the instruments and electric equipment must be located in hazardous areas. Purging and pressurization is a protection method that relies on clean air or inert gas (i.e., nitrogen) to be continuously supplied to the enclosure at sufficient flow to keep the equipment adequately cooled and to provide adequate internal pressure to prevent the influx of combustible atmospheres into the enclosure. Although the enclosures are not explosion proof, they must be relatively vapor tight and must have adequate strength to perform safely and satisfactorily. The system consists of the following:

1. *Clean Air* (*or Inert Gas*) *Supply*: Careful study and analysis is of crucial importance to this process since the air supplied must be reasonably free of contaminants. Finding a safe location for an air intake requires skill and ingenuity. Consulting with an HVAC specialist is recommended. Other factors such as vapor density, location, wind pattern, and surrounding environment should also be considered. Where compressors and blowers are used to supply compressed air, caution must be exercised when selecting the proper compressor or blower size and location in order to meet airflow requirements without compromising the main objective of safety and reliability.
2. *Purging*: A pressurized enclosure that has been out of service for some time tends to collect a combustible mixture. Before energizing, inert gas and positive pressure must provide a sufficient initial clean air volume to minimize the concentration of any combustible mixture that may be present. For typical applications, a flow of four times the internal volume of the enclosure is usually sufficient to minimize the concentration of combustible mixture that may exist. For unusual applications, the flow volume must be carefully calculated to ensure the success of the purging process.
3. *Pressurization*: This process uses the concept of pressure differential between the outside and the inside of the enclosure to keep flammable materials from entering. This is accomplished by maintaining a higher pressure on the inside of the enclosure. For safe operation, the protected enclosure must be constantly maintained at a positive pressure of at least 25 Pa (0.1 in. water) [2] above the surrounding atmosphere during the operation of the protected equipment.
4. *Signals and Alarms:* When positive pressure is lost, warnings and alarms are essential. Three types of pressurization and alarms can be used, depending on the nature of the controls of the enclosure and the degree of hazard outside. In addition, door interlocks are required to prevent opening of the protected enclosure while the circuit is energized.

According to NFPA 496, there are three types of pressurization. They are

Type X—reduces the classification within the protected enclosure from Division 1 to nonclassified. This usually involves a source of ignition housed in a tight enclosure located in a potentially hazardous atmosphere. Type X requires a disconnecting means (flow or pressure switch) to de-energize power to the protected enclosure completely and automatically immediately upon failure of the protective gas supply (loss of either pressure or flow). The disconnecting means must be explosion proof or intrinsically safe, as it is usually located in the hazardous area.

Type Y—reduces the classification within the protected enclosure from Division 1 to Division 2. The protected enclosure houses equipment rated for Division 2 and does not provide a source of ignition. Therefore, no immediate hazard is created. Type Y requires a visual or audible alarm in case of system failure. *Word of caution*: Safeguards must be established to ensure that any malfunction in the system does not raise the external surface temperature of the enclosure to over 80% of the ignition temperature of the combustible mixture present.

Type Z—reduces the classification within the protected enclosure from Division 2 to nonclassified. Type Z requires a visual or audible alarm to be activated if failure to maintain positive pressure or flow within the protected enclosure has been detected.

Recognizing and understanding the potential danger associated with the use of electricity in hazardous areas is a crucial part of selecting the best protection. Techniques for most applications, where the need for different energy demands is required, will likely involve a combination of various methodologies and specialized technologies. Properly performed analysis and investigation may appear to be time consuming and expensive, but for those who are willing to adhere to the established guidelines and solid engineering practices, the process will help ensure the highest level of compliance while yielding tremendous savings and preventing property damage and injuries.

Defining Terms

Inductance: The ability of an electric apparatus to store an electric charge (energy). An inductor will release this energy when the circuit is opened (broken).

Oxidation: The process where negatively charged ions (anions) lose electrons at the anode during electrochemical process. For anions to become neutral, they must lose electrons.

Resistance temperature device (RTD): A device that measures temperature based on change of resistance.

Zener diode: A nonlinear solid-state device that does not conduct current in reverse bias mode until a critical voltage is reached. It is then able to conduct current in reverse bias without damage to the diode. Zener diodes have almost constant voltage characteristics in the reverse bias region (usual operation).

References

1. NFPA 70 (ANCI C1—1996), *National Electrical Code 1996*, Quincy, MA.
2. National Fire Protection Association (NFPA), Articles 493, 496, 497, Quincy, MA.
3. American Petroleum Institute 500 (RP 500), 1st edn., Washington, DC, June 1991. *Recommended Practice for Classification of Location for Electrical Installation at Petroleum Facilities*.
4. Elcon Instruments, Inc., *Introduction to Intrinsic Safety*, 3rd printing, Norcross, GA, 1990.
5. Underwriters Laboratories, Inc., Pub. UL 913, 4th edn., Northbrook, IL, 1988, *Standard for Intrinsically Safe Apparatus & Associated Apparatus for Use in Class I, II, III, Division 1 Hazardous Locations*.
6. UL's Standard Transition for UL 508 (Industrial Control Equipment and UL 508C (Power Conversion Equipment), Publications, Underwriters Laboratories, Inc., http://www.ul.com/global/eng/pages/offerings/industries/powerandcontrols/industrialcontrolequipment/ (accessed on March 15, 2013).
7. Clarence W de Silva, *Sensors and Actuators: Control System Instrumentation*, CRC Press, Boca Raton, FL, 2007.
8. R. Stahl, Inc., RST 49, *Comprehensive Applications and Data Catalog*, Woburn, MA, 1992.
9. Intrinsic Safety, PDF SB, Online, http://pdfsb.com/intrinsic+safety (accessed on March 15, 2013).

Further Readings

Explosion Proof Equipment, North American Industries. http://www.naicranes.com/cranes/explosion_proof.htm (accessed on August 1, 2012).

Explosion Proof motors and Applications Hazardous Location. http://www.reliance.com/mtr/eplprfmn.htm (accessed on August 2, 2012).

Electronic Code of Federal Regulations, Title 30—Mineral resources. http://ecfr.gpoaccess.gov/cgi/t/text/text-idx?c=ecfr&tpl=/ecfrbrowse/Title30/30cfr18_main_02.tpl (accessed on August 9, 2012).

Partial List of Manufacturers and Suppliers

Ashcroft, Inc., 250 E. Main St., Stratford, CT 06614-5145, Phone: 203-378-8281, 800-328-8258 (toll free), Fax: 203-385-0408, www.ashcroft.com

Purge Solutions, 12454 Galveston Rd, Suite C, Webster, TX 77598, Phone: 832-368-7166, Fax: 281-992-5726, www.purgesolutions.com

R. Stahl, Inc., 9001 Knight Rd., Houston, TX 77054, Phone: 713-792-9300, 800-782-4357 (toll free), Fax: 713-797-0162, www.rstahl.com

Stancor, Inc., 515 Fan Hill Rd, Monroe, CT 06468, Phone: 203-268-7513, Fax: 203-268-7958, www.stancorpumps.com/index.html

Unico, Inc., 3725 Nicholson Rd., Franksville, WI 53126, Phone: 262-886-5678, Fax: 262-504-7396, www.unicous.com

Ware, Inc., 4005 Produce Rd., Louisville, KY 40218-3007, Phone: 502-968-2211, 800-228-8861 (toll free), Fax: 502-969-5777, www.wareinc.com/



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

95

Measurement and Identification of DC Brush and Brushless Stepping Motors

Stuart Schweid
Xerox Corporation

Robert Lofthus
Xerox Corporation

John McInroy
University of Wyoming

95.1 Introduction **95**-1
95.2 Hybrid Stepping Motors **95**-2
Chopping Current Amplifiers • Microstepping • Closed-Loop Control • Position Measurement
95.3 DC Brush and Brushless Motors **95**-5
Pulse-Width-Modulated Power Amplifier • Measuring the System Dynamics
References **95**-9
Further Readings **95**-10
Partial List of Manufacturers and Suppliers **95**-10

95.1 Introduction

There are many systems, such as robots, whose ability to move themselves or other objects is their primary purpose. There is a myriad of other systems, such as xerographic printers, where the motion is not the desired outcome but is required in the performance of the mainline objective. All of these systems require one or more "prime movers," which directly or indirectly create all motion in the system.

All of these applications have differing requirements for both the type and the precision of the motion produced. There are applications, such as children's toys, that can perform well with imprecise motion requirements. Conversely, there are applications, such as printing, that require very precise motion control.

Depending on the type of motion required, a prime mover can be chosen from one of several candidates. The most ubiquitous motor types are DC brush–type motors, DC brushless motors, and hybrid stepping motors. The system, which includes the prime mover, can be controlled simply as a function of time (i.e., open loop) or as a function of both time and system state (i.e., closed loop).

A closed-loop system requires some mechanism of measuring the "state" of the system. For motion systems, this is typically one or more sensors that measure position and/or velocity. In addition, for many of the control techniques, specifically, linear controllers, it is also advantageous to have a model of the system being controlled. The model can be used in conjunction with a plethora of design techniques to improve the dynamic response of the system significantly.

This chapter will describe the major modes of operations for the prime movers listed and will detail a method for measuring the open-loop response of a candidate motion system.

95.2 Hybrid Stepping Motors

Hybrid stepping motors are useful devices that provide fairly accurate positioning in the open loop: their most common mode of operation. They are successful in open-loop operation because they will remain within a commanded position increment or step as long as the motor has sufficient torque to resist any external torque applied by the system.

Since the motor operates in the open loop, it is chosen a priori to ensure it provides the torque necessary to remain within step. The maximum torque needed to operate the system is a function of the inertia of the system, the acceleration profile of the system, the external torque on the system, etc. Once the maximum torque required at every speed is determined, a motor can be chosen from its torque–speed curve. The torque–speed curve describes the maximum torque that can be tolerated at each speed without failing to follow a commanded step (referred to as a "losing step"). A motor whose torque–speed curve exceeds the torque requirements of the system at all speeds will be sufficient for the application. Figure 95.1 shows a typical torque–speed curve.

Hybrid stepping motors have a plurality of coils (almost always two) that, when commutated in a predefined sequence, cause the motor to turn. When the motor is operated in a mode referred to as "full stepping," the current command to each winding is constant in magnitude but varying in polarity (positive or negative). Note that the definition of which direction of current flow is "positive" is purely arbitrary. In this mode, the motor has only four separate energizations of the winding to complete a full revolution, although clever mechanical arrangements allow the number of steps per revolution to be greatly increased: typically 200 but as large as 800.

With two distinct windings (referred to as A and B), the four possible arrangements of winding energizations are A+B+, A+B–, A–B–, and A–B+. If these combinations of winding currents are commanded to the motor in the sequence listed in the previous sentence, the motor will turn.

95.2.1 Chopping Current Amplifiers

As previously stated, the motion of the stepper requires a current of a particular amplitude to be applied. For economic reasons, most systems have a voltage power supply available (V_{ss})—not a current source. A "current chopper" is a low-cost closed-loop switching amplifier that regulates the current to the

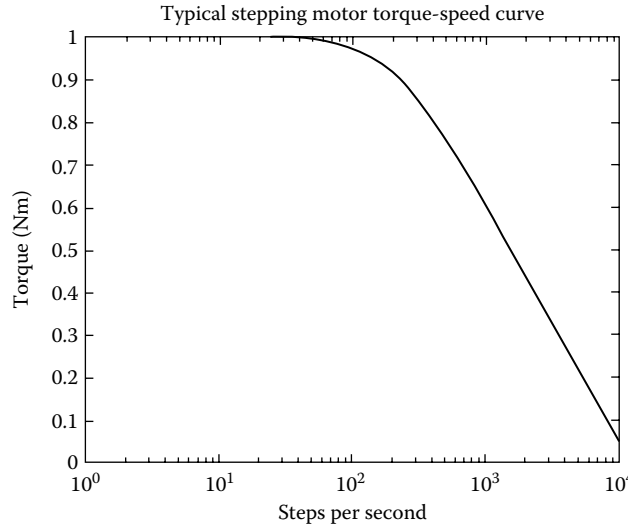


FIGURE 95.1 Typical stepping motor torque–speed curve.

motor winding. It takes advantage of the fact that the motor winding has a significant inductance. The two motor windings have dynamic equations [1]:

$$\frac{dI_A}{dt} = \frac{(V_A - I_A R_A + K_b \omega \sin\theta)}{L_A}$$
$$\frac{dI_B}{dt} = \frac{(V_B - I_B R_B + K_b \omega \cos\theta)}{L_B} \tag{95.1}$$

where

I is the current winding (A)
R is the current resistance (Ω)
L is the current inductance (H)
V is the applied voltage to the winding (V)
K_b is the torque constant of the motor (Nm/A)
ω is the motor velocity (rad/s)
θ is the motor position (electrical rad)

In every t_s (where t_s is very small, i.e., 10 μs), the current chopper applies either V_{ss} or $-V_{ss}$ to the motor winding, depending on whether the desired current (I_M) is greater than or less than the measured (i.e., actual) winding current:

$$\text{if } (I_A > I_M), \quad V_A = -V_{ss} \quad \text{else } V_A = V_{ss}$$
$$\text{if } (I_B > I_M), \quad V_B = -V_{ss} \quad \text{else } V_B = V_{ss} \tag{95.2}$$

The usual requirement is that ($V_{ss}\, t_s/L$) is small in comparison with I_M, resulting in a chopping range that is small when compared with the desired current. Furthermore, the mechanical system that includes the motor is typically low pass in nature, so it will have minimal response to the small-amplitude, very-high-frequency components present in the motor winding current.

95.2.2 Microstepping

Another common open-loop mode for stepper operation is microstepping. In microstepping, the commanded reference positions can be between the "full-step" positions of the motor. The commanded currents for the reference position θ_{ref} are

$$I_A = I_M \cos(\theta_{ref})$$
$$I_B = I_M \sin(\theta_{ref}) \tag{95.3}$$

where θ_{ref} is the desired electrical position of the motor (in four full steps, θ_{ref} traverses one full electrical revolution). For constant-velocity applications,

$$\theta_{ref} = N\omega_d t \tag{95.4}$$

where

N is the number of electrical revolutions per mechanical revolution (one-quarter the number of steps per revolution) of the motor
ω_d is the desired velocity (rad/s)

95.2.3 Closed-Loop Control

In addition to the open-loop uses of hybrid stepping motors, several researchers have recently invented methods of operating these motors in a closed-loop fashion. The principal difficulty arises because hybrid stepping motors are nonlinear devices, so the plethora of design techniques available in linear control theory cannot be directly applied. Hamman [1] uses a fourth-order model to linearize the system about each step. This is useful in altering the undamped response of the stepper motor to a commanded change in position. Schweid [2] suggests a scheme that allows good transient behavior during constant-velocity applications but uses the complete fourth-order model. In systems that use current choppers, a second-order model is sufficient. The following control law can then be applied [3–5]:

$$\begin{aligned} I_A &= I_M \cos(\theta_{ref}) - I_C \sin\theta \\ I_B &= I_M \sin(\theta_{ref}) + I_C \sin\theta \end{aligned} \tag{95.5}$$

The currents of Equation 95.6 linearize the model of the stepper motor, allowing common control techniques such as PID control to be applied [6]. The value I_M is used as an open-loop component to microstep the motor, while I_C is a feedback term that permits an improved dynamic response.

95.2.4 Position Measurement

The implementation of closed-loop control requires a method of measuring the position of the motor. Furthermore, there is the added constraint that this position measurement must be aligned with the motor construction and not have a positional offset with respect to it. This is because the control scheme of Equation 95.5 requires a command current I_C be multiplied by the sine and cosine of the motor position. One way to implement this is to have an incremental encoder with an index pulse attached to the motor. A calibration can then determine the relative position of the index pulse to the motor shaft and use this relationship in all future positional measurement.

Another scheme uses the back emf of the two coil windings to estimate the position. Some motors are constructed with additional sensing coils that allow easy measurement of these values. From Equation 95.2, the back emf of both coils can be found using

$$\begin{aligned} V_{emfA} &= -K_b \omega \sin\theta = V_A - I_A R_A - \frac{L_A dI_A}{dt} \\ V_{emfB} &= -K_b \omega \cos\theta = V_B - I_B R_B - \frac{L_B dI_B}{dt} \end{aligned} \tag{95.6}$$

and the position could be found using

$$\theta = \text{atan}\left(\frac{-V_{emfA}}{V_{emfB}}\right) \tag{95.7}$$

In most motors, however, V_{emfA} and V_{emfB} are not directly measurable. However, it is possible to obtain an excellent estimate of position in the constant-velocity system by approximating the back emf signals through low-pass-filtered measurements of the coil voltages. One method developed at Xerox that has a U.S. patent (U.S. 5378975) is described here. Taking the Laplace transform of Equation 95.6 yields

$$\begin{aligned} V_{emfA}(s) &= V_A(s) - (R_A + sL_A) I_A(s) \\ V_{emfB}(s) &= V_B(s) - (R_B + sL_B) I_B(s) \end{aligned} \tag{95.8}$$

Filtering both sides with low-pass analog filters matched to the dynamics of the coil produces

$$V_{\text{filtA}}(s) = \frac{V_{\text{A}}(s)}{s(L_{\text{A}}/R_{\text{A}})+1} - R_{\text{A}}I_{\text{A}}$$
$$V_{\text{filtB}}(s) = \frac{V_{\text{B}}(s)}{s(\text{L}_{\text{B}}/R_{\text{B}})+1} - R_{\text{B}}I_{\text{B}} \tag{95.9}$$

The quantities V_{filtA} and V_{filtB} are easily obtained by low-pass filtering the coil voltage with a simple *RC* circuit and subtracting the *IR* term. The subtraction is easily implemented because the current is a known input (in fact, it is commanded), and the resistance is easily measured.

Assuming that the coils are matched: $L = L_{\text{A}} = L_{\text{B}}$, $R = R_{\text{A}} = R_{\text{B}}$, and the motor is moving at constant velocity, $\theta = N\omega_{\text{d}}t$, and the back EMF voltages are sinusoidal. The low-pass-filtered versions of these voltages are given by

$$V_{\text{filtA}}(s) = |H(N\omega_{\text{d}})|K_{\text{B}}\sin(\theta+\phi)$$
$$V_{\text{filtB}}(s) = |H(N\omega_{\text{d}})|K_{\text{B}}\cos(\theta+\phi) \tag{95.10}$$

where

$|H(N\omega_{\text{d}})|$ is the magnitude response of the low-pass filter at the input frequency, $N\omega_{\text{d}}$

$\phi = -\text{atan}(N\omega_{\text{d}}\ L/R)$ is the phase shift due to the filtering

Dividing the two Equations 95.10 and taking the inverse tangent yields

$$\theta = \text{atan}\left(\frac{-V_{\text{filtA}}(t)}{V_{\text{filtB}}(t)}\right) + \text{atan}\left(\frac{N\omega_{\text{d}}L}{R}\right) \tag{95.11}$$

The full derivation and analysis are found in Ref. [7]. In addition, Ref. [7] includes a Kalman filtering technique to improve the velocity measurement estimate from the position estimate. This is needed because the derivative or difference of the position measurement is highly sensitive to noise.

95.3 DC Brush and Brushless Motors

The DC servomotor is the oldest and probably the most commonly used actuator for servo systems. The name arises because a constant (or DC) voltage applied to the motor will produce motor movement. Over a linear range, these motors exhibit a torque that is proportional to the current flowing in the winding. The chief disadvantage of the brush DC motor is the brush, which performs the mechanical commutation. These brushes are the first part of the motor to fail, severely limiting the reliability of DC brush motors when compared with other actuators.

Another motor that operates similarly to a DC brush motor is the DC brushless motor (a DC voltage will cause it to rotate). Its method of commutation, however, is much different. Whereas the DC motor accomplishes switching mechanically with brushes, the DC brushless motor contains Hall effect sensors that electronically control the switching. As a result, the DC brushless motors have much greater life and reliability.

95.3.1 Pulse-Width-Modulated Power Amplifier

In many systems where they operate, DC motors are required to have a varying voltage at their input in order to create the position or velocity profiles desired. Many of these systems, however, have only

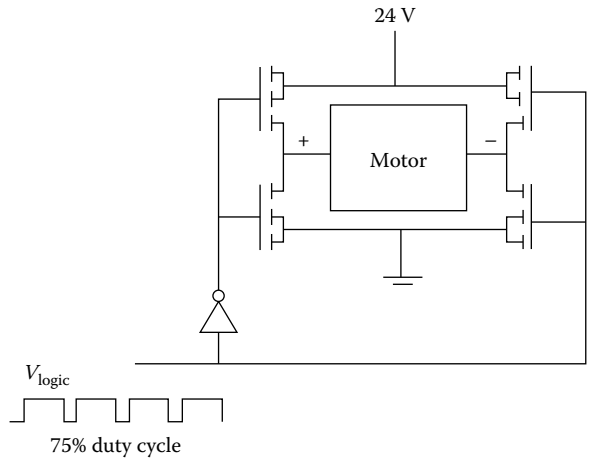


FIGURE 95.2 PWM amplifier set to deliver 12 V to the motor.

a single power supply V_{ss}. Rather than using analog power amplifiers that are both expensive and inefficient, a pulse-width-modulated (PWM) amplifier is used. A pulse-width power amplifier can deliver only V_{ss} or $-V_{ss}$ (or zero in some designs) V to the motor. Different voltages are delivered to the motor by varying the relative percentage that V_{ss} and $-V_{ss}$ are applied to the motor at a high fixed frequency. The mean or average voltage to the motor is a linear function of this relative percentage or duty cycle. The motor also receives the higher harmonics of the switched input that are integer multiples of the base frequency of the square wave. Most motors, however, have at least two poles (an electric and a mechanical) that are significantly slower than the switching frequency and thus greatly attenuate any possible effect they might have on motion. Figure 95.2 shows a PWM amplifier set to deliver 12 V to the motor.

95.3.2 Measuring the System Dynamics

In systems with moderate or stringent motion requirements, the DC brush and brushless motors are operated in a closed-loop fashion. In order to perform in closed loop, some measure of either position or velocity is needed. A tachometer is commonly used to measure rotational speed. It outputs a voltage that is proportional to the speed of a rotating member. If a position measurement is required, an incremental encoder can be used. The incremental encoder produces a digital pulse at equal points in position. For example, a 1000 line/revolution encoder would produce a digital wave with 1000 rising (and falling) edges during one revolution.

Once the full system (including feedback sensors) is determined, a feedback control algorithm can be developed. As part of the design of many compensation algorithms, especially linear compensators such as state feedback or PID, one of the most important tasks is to have a good model of the dynamics of the system.

There are several ways to generate the model. One of the most common methods is to create the state equations of the system from the physical equations. This requires knowledge of all the parameters of the system. Sometimes these parameters are hard to measure and are not provided by the manufacturer. One example of this might be the torsional constant of a hard rubber roll in the system.

Even if all of the parameters are known, it is still desirable to have a method that measures the response of the system. This provides a way to either validate the model or inform the designer that the model is not an accurate representation of the system.

A method is presented that permits the collection of input/output data of a system. These data can then be used to fit a frequency response curve or a transfer function model of the system. Signal analyzers,

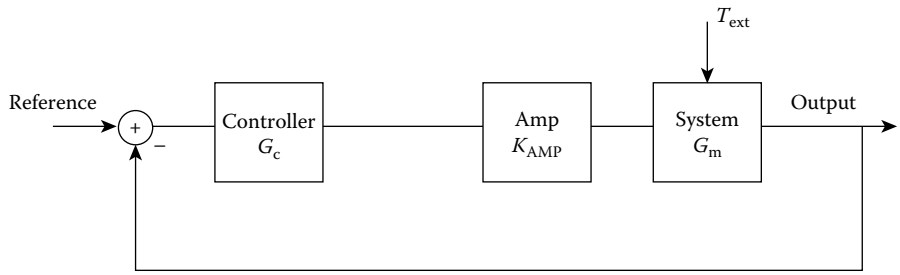


FIGURE 95.3 Typical closed-loop system.

such as the HP3562A, have the ability to perform such curve fitting accurately. In addition, if the data are collected in a discrete fashion, there are commercially available software tools, such as MATLAB®, that can provide model estimation in either a parametric (e.g., state space/transfer function) or a non-parametric (e.g., frequency response) form. Although this procedure is described for motor systems, it is easily implemented in most linear systems.

Assume a typical closed-loop system as shown in Figure 95.3. There are several obstacles that make it difficult to obtain the open- or closed-loop transfer function of this system. These obstacles include the inability to measure some of the outputs and the inability to control/modify certain inputs.

To acquire the frequency response of a system, it is useful to perturb the system with a known input and measure the elicited response. A controllable input—one that can be externally manipulated—is required to accomplish this task. The only controllable inputs of the system in Figure 95.3 are the torque acting upon the motor shaft and the reference command.

Changing the external torque would require mechanically attaching a calibrated brake or another motor. This is not only cumbersome, but it also can alter the system dynamics, as it introduces inertia to the true system under measurement. Within many systems, the reference signal is a digital square wave (since the feedback element is an incremental encoder and produces a similar wave) and therefore requires a voltage-to-frequency converter if the reference is to be directly manipulated. This converter may include dynamics that can corrupt the measured data.

In addition to a controllable input, a measurable output must be available. In many servo systems, the only measurable outputs are the motor armature voltage and the position/velocity. The armature voltage may be a PWM signal that needs to be filtered to create an analog voltage. In the case of an incremental encoder, the output consists of a digital square wave generated by an incremental encoder: a frequency-to-voltage converter is required to produce an analog output. This converter may include dynamics that can corrupt the measured data.

Consider the system described in Figure 95.4. The summing junction placed in the system has the advantage that can be added at any point that is convenient for measurement purposes—it is no longer limited to unaugmented system inputs or outputs.

In the case of analog controllers, the inputs and the outputs of the summing junction are analog voltages, making them easily controlled and measured. The summing node can be easily implemented using a simple op-amp adder circuit. This introduces no appreciable dynamics into the original system and will not, therefore, skew the measured data.

When the system response is to be measured, the time-varying contributions of the other system inputs, the velocity reference and external torque, T_d, acting upon the system must be eliminated. The designer must set the velocity reference as a constant value. There is usually little control of the external torque acting on the system, and this will introduce noise into the measurement technique. To ensure a large signal-to-noise ratio (SNR), the alternative reference input (signal $z(t)$) should be made as large as possible.

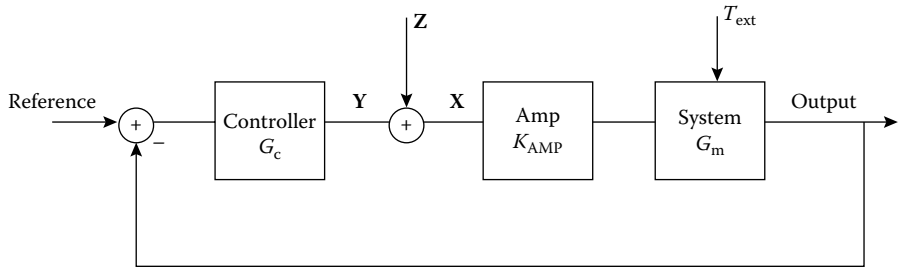


FIGURE 95.4 Instrument-integrated closed-loop system.

The alternative reference input, $z(t)$, can stimulate the system with a swept sine input, a random input, or any other desired signal. During the stimulation, a time record is collected of the other nodes of the summing junction, $y(t)$ and $x(t)$. The transfer function between **Y** and **X** is

$$\frac{\mathbf{Y}(s)}{\mathbf{X}(s)} = -G_c(s)K_{amp}(s)G_m(s), \quad -\frac{\mathbf{Y}(s)}{\mathbf{X}(s)} = G_c(s)K_{amp}(s)G_m(s) \tag{95.12}$$

The second term is, by definition, the open-loop transfer function of the system. As earlier, the MATLAB or a signal analyzer can provide either a parametric or a nonparametric fit of this transfer function from the data record.

Note that this technique requires the system be functioning in a closed-loop mode during data collection. Typically, G_c is set as a proportional controller at a value that will ensure stability of the overall system. For a good data collection, the proportional gain should be made as large as possible.

The case of a digital or microprocessor-controlled system can be handled similarly, with some slight modification. Consider the system of Figure 95.5, where a microprocessor acts as the system controller. In this instance, the summing junction can be implemented internal to the microprocessor by executing an add instruction. The problem here becomes getting the data into and out of the microprocessor. Getting the data into the microprocessor requires the addition of an analog-to-digital converter (ADC). In the block diagram, this ADC process is represented by the block A_f, A_f that allows representation of any dynamics inherent in the ADC design (especially gain). The output of the microprocessor may be a digital PWM signal. This signal may need to be low-pass-filtered to create an analog voltage for the signal analyzer taking the data. The low-pass filter is represented by P_f in Figure 95.5. It is typically a simple *RC* circuit whose cutoff frequency is much lower than the PWM signal frequency.

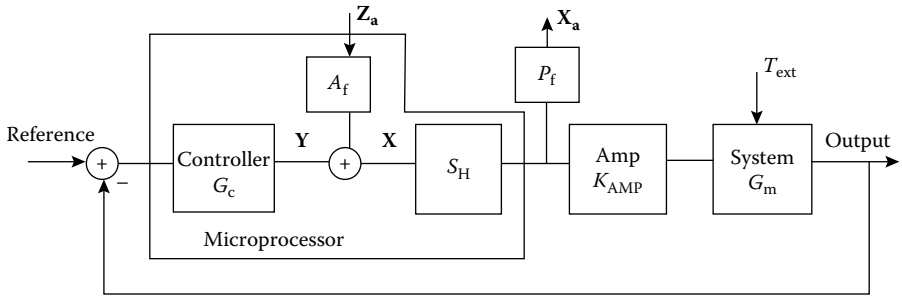


FIGURE 95.5 Microprocessor-controlled closed-loop system.

Given the condition that ω_{ref} is static and external torque, T_d, is minimal, the transfer function between $\mathbf{Z_a}(s)$ and $\mathbf{X_a}(s)$ in Figure 95.5 is

$$\frac{\mathbf{Z}_a(s)}{\mathbf{X}_a(s)} = \frac{1 + G_c(s)S_H(s)K_{amp}(s)G_m(s)}{A_f(s)S_H(s)P_f(s)} \equiv \mathbf{O}(s) \tag{95.13}$$

when $G_c = 0$, $\mathbf{Z_a}(s)/\mathbf{X_a}(s) = 1/(A_f(s)S_H(s)P_f(s)) \equiv \mathbf{C}(s)$, and $(\mathbf{O}(s) - \mathbf{C}(s))/\mathbf{C}(s) = \boldsymbol{G}_c(s)\boldsymbol{S}_H(s)\boldsymbol{K}_{amp}(s)\boldsymbol{G}_m(s)$; this is the system open-loop transfer function.

A more direct approach is possible if a logic analyzer is not used, but a software-based system identification package such as MATLAB is used to fit the data. The stimulus, z, can either be stored as a set of data in a lookup table in the microprocessor memory or can be generated in real time using any random number generator algorithm or other algorithm. The data $\mathbf{Y}$ and $\mathbf{X}$ at the internal summing junction can be captured via a logic analyzer or emulator if the values of $y(k)$ (the value of y at the kth sample period) and $x(k)$ are written to an external bus every sample period. Under this configuration, neither a DAC nor an ADC is necessary, nor is any mathematical combination of two different transfer functions required to produce the desired open-loop transfer function. The transfer function between $\mathbf{Y}(z)$ and $\mathbf{X}(z)$ is

$$\begin{aligned} \frac{\mathbf{Y}(\mathbf{z})}{\mathbf{X}(\mathbf{z})} &= -G_c(z)S_H(z)K_{amp}(z)G_m(z) \\ -\frac{\mathbf{Y}(\mathbf{z})}{\mathbf{X}(\mathbf{z})} &= G_c(z)S_H(z)K_{amp}(z)G_m(z) \end{aligned} \tag{95.14}$$

Again, this is the open-loop transfer function of the system, acquired using only minimal additional software and a logic analyzer. As earlier, the input/output set of data can be sent to any system identification algorithm to fit the transfer function.

References

1. E. Hamman, Closed-loop control of a DC stepping motor using state space techniques, Master's thesis, Rensselaer Polytechnic Institute, Troy, NY, 1983.
2. S. Schweid, Velocity control of DC stepping motors utilizing state space theory, quasilinearization and reduced order estimation/control, Masters thesis, Rensselaer Polytechnic Institute, Troy, NY, 1985.
3. J. Tal, L. Antignini, P. Gandel, and N. Veignat, Damping a two-phase step motor with velocity coils, *Incremental Motion Control Systems and Devices Symposium*, Champaign, IL, pp. 305–309, 1985.
4. D. Reignier, Very accurate positioning system for high speed, high acceleration motion control, *Power Conversion and Intelligent Motion*, 13, 52–62, June 1987.
5. D. Reignier, Damping circuit and rotor encoder cut disc magnet step motor overshoot, settling time and resonances, *Power Conversion and Intelligent Motion*, 14, 64–69, April 1988.
6. S. Schweid, J. McInroy, and R. Lofthus, Closed loop low-velocity regulation of hybrid stepping motors amidst torque disturbances, *IEEE Transactions on Industrial Electronics*, 42, 316–324, 1995.
7. R. Lofthus, S. Schweid, J. Mclnroy, and Y. Ota, Processing back EMF signals of hybrid step motors, *Control Engineering Practice*, 3(10), 1–10, 1995.

Further Readings

Hanselman, D., *Brushless Permanent Magnet Motor Design*, 2nd edn., McGraw-Hill, New York, 2003.

Hughes, A., *Electric Motors and Drives: Fundamental Types and Applications*, 3rd edn., Elsevier, Oxford, U.K., 2006.

Gieras, J.F. and Wing, M., *Permanent Magnet motor Technology: Design and Applications*, 2nd edn., Marcel Dekker, Inc., New York, 2002.

Partial List of Manufacturers and Suppliers

Acme Controls, 6140 W. Higgins Ave., Chicago, IL 60630-1845, Phone: 877-314-2371 (toll free), Fax: 773-774-3737, acmecontrols.rtrk.com/?scid=2216829&rl_a…

Aerotech, Inc., 101 Zeta Dr., Pittsburgh, PA 15238-2811, Phone: 412-963-7470, Fax: 412-963-7459, www.aerotech.com

Dart Controls, 5000 W. 106th St., Zionsville, IN 46077, Phone: 317-873-5211, Fax: 317-873-1105, www.dartcontrols.com

Electric Motor Centre, 3222 Bernice Ave., Russellville, AR 72802, Phone: 479-968-2532, 800-871-3697 (toll free), Fax: 479-968-2052, www.electricmotorcenterinc.com

Unico, Inc., 3725 Nicholson Rd., Franksville, WI 53126, Phone: 262-886-5678, Fax: 262-504-7396, www.unicous.com

Trola Industries, 2360 N. George St., York, PA 17406-3202, Phone: 717-848-3700, Fax: 717-848-6993, www.trolaindustries.com

96

Human Factors in Displays

96.1 Introduction **96**-1
96.2 Fundamentals of Light Measurement **96**-2
96.3 Fundamentals of Vision **96**-2
96.4 Visual Performance **96**-3
96.5 Display Performance Considerations **96**-4
Contrast • Gray Scale • Pixels • Resolution and Aspect Ratio • Sharpness and Legibility • Gamut and Color Spaces • Flicker • Refresh Rate
96.6 Display Content Considerations **96**-10
Alphanumeric Displays • Quantitative and Status Displays • Summary
96.7 Cathode-Ray Tube Displays **96**-11
CRT Performance
96.8 Liquid Crystal Displays **96**-12
LCD Performance • TN-Effect LCDs • IPS LCDs • VA LCDs
96.9 Plasma Displays **96**-13
Plasma Display Performance
96.10 Electroluminescent Displays **96**-14
EL Display Performance
96.11 Light-Emitting Diode Displays **96**-15
LED Display Performance
96.12 Electronic Paper/Electronic Ink **96**-15
Electronic-Paper Display Performance
Defining Terms **96**-16
References **96**-16
Further Information **96**-17

Jeffrey D. Onken
Purdue University

Barrett S. Caldwell
Purdue University

Steven A. Murray
University of San Diego

96.1 Introduction

The display system is the final information flow link between the electronic system (task automation, measuring process, or other electronic information technology) and the user. If the display is not easy to access, perceive, and understand, then that flow is degraded. The user's sensory capabilities and cognitive characteristics, therefore, must be addressed in defining and improving the system display requirements.

Since the 1990s, there has been a substantial proliferation in the nature and form factors of video- and visually based displays. The range of devices that can be considered "computer displays" has expanded into new classes of mobile and flat-screen device options. Furthermore, display technologies and performance capabilities are easier to evaluate in the context of their intended application. Consideration of

the following issues can narrow the search for candidate systems and can prevent needless frustration during system use:

Environment: Will the display be operated in sunlight or at night or in confined conditions?
Application: Will the display present alphanumeric data, video images, graphics, or some combination?
Task scenario: Is portability, handheld operation, or group viewing required?
System characteristics: Are there critical operational constraints such as weight, volume, power, maintenance, or cost?

This chapter begins with basic treatments of light and vision. It then proceeds to discussions of visual capabilities and the display characteristics that must be matched to them. Rather than elaborating characteristics of specific visual display examples, this chapter will focus on video and visual information presentation technologies and their interactions with human performance.

96.2 Fundamentals of Light Measurement

The foundation metric of light is *luminous flux*, which is the rate at which light energy is emitted from a source and is expressed in lumens (lm). *Luminous intensity* is luminous flux per unit solid angle, and its unit of measurement is the candela (cd). This is distinguished from *illuminance*, or illumination, which is simply luminous flux per unit area, expressed as lux (lx). *Luminance* is a measure of the brightness, that is, the amount of light, per unit area, either emitted by or reflected from a surface. Units of luminance measurement are candelas per square meter (cd/m^2) or nits. Finally, *reflectance* is a unitless ratio of the amount of light striking a surface to the amount of light leaving it:

$$R = \pi \times \frac{\text{Luminance}}{\text{Illuminance}} \tag{96.1}$$

High reflectance can create glare, dramatically reducing visual performance.

96.3 Fundamentals of Vision

The eye functions very much like a conventional camera. Light enters the eye through a transparent *cornea* and is modulated by the *pupil*, a variable aperture opening controlled by muscles of the *iris*. The pupil grows larger in dark surroundings and smaller in bright surroundings to control the range of light intensity to the eye. Light rays are then refracted by an adjustable *lens* and brought into focus on the *retina*, where neural imaging begins. The retina contains both *cones* and *rods*, two distinctly different types of photoreceptors. Cones are concentrated near the *fovea*, or the central 2° of the visual field, and decrease rapidly with distance from this region. In contrast, rods are essentially absent in the fovea and increase in density with distance.

The eye is sensitive to three characteristics of electromagnetic radiation: (1) *brightness* (the intensity of ambient or incident light, measured in lux), (2) *hue* (the wavelength of light, measured in nm), and (3) *saturation* (relative concentration of specific hues in light, measured as a dimensionless ratio from 0 to 1). Cones are differentially sensitive to wavelength, that is, hue, and have greater resolving power than rods because of their one-to-one mapping onto visual nerves. Cones can be further divided into three types, each maximally sensitive to a different portion of the visible light spectrum: (1) red (670 nm peak), (2) green (470 nm peak), and (3) blue (400 nm peak). Rods are more sensitive to light than cones and have many-to-one connections with the nerves that exit the eye, a feature that permits neural summation of low light signals. Human ability to discriminate differences in levels of brightness, saturation, or hue is governed by a psychophysical function known as *Weber's law*:

$$K = \frac{\Delta I}{I} \tag{96.2}$$

where ΔI is the difference, or change, in intensity, I is the initial intensity, and K is the Weber fraction. Values of K have been experimentally determined for brightness (0.079), saturation (0.019 for red), and hue (≈0.02–0.04, depending on the region of the visible spectrum).

Photopic vision occurs at light levels where both rods and cones are sensitive. The minimum light intensity required for photopic vision is approximately 2 lx; colors are seen in this region. As brightness decreases, a transition from photopic to scotopic vision takes place and color perception drops out gradually, a phenomenon that can impact the interpretation of color-coded information in poor light. The perception of blues and reds is lost first, then cyan and yellow orange, and finally green, that is, the wavelengths where the eye is most sensitive. The eye becomes most sensitive to wavelengths of about 550 nm (green) near the limit of photopic vision. *Scotopic vision* occurs at low light levels (2×10^{-7} lx to 2 lx) and primarily involves the rods; only achromatic shades of gray are seen. The transition from photopic to scotopic vision occurs slowly, requiring approximately 30 min for complete adjustment from photopic to scotopic visual sensitivity.

96.4 Visual Performance

Visual acuity is the ability to discriminate detail. The action of the lens, to change focus for objects at different distances, is called *accommodation*. *Minimum separable acuity*, the most common measure of discrimination, is determined by the smallest feature that the eye can detect and is measured in terms of the reciprocal of the visual angle subtended at the eye by that feature. *Visual angle*, in minutes of arc, is calculated as

$$\mathrm{VA} = \frac{3438H}{D} \tag{96.3}$$

where
H is the height of the object
D (in the same units) is the distance from the observer

The ability to distinguish an object from its surroundings is known as *visibility*. The term is related to visual acuity but implicitly combines factors of object size, contrast (i.e., including differences in hue and saturation), and brightness that all interact to determine true visual detection performance. On a more functional level, *readability* or *legibility* describes the ability to distinguish meaningful groups of objects (e.g., words extracted from groups of letters on a display).

Other parameters affecting visual performance include viewing angle and viewing distance. *Viewing angle* at the eye is measured from a line through the visual axis to the point being viewed and determines where an object will register on the retina. The best image resolution occurs at the fovea, directly on the line of gaze, and visual acuity degrades with increasing angle away from this axis. *Viewing angle* at the display is the angle, in degrees, between a line normal to the display surface and the user's visual axis. The best viewing angle is, of course, on the visual axis and normal to the display surface, as luminance falls off for most displays as the angle from normal increases. Luminance reduction with viewing angle can be calculated as

$$E = E_{\mathrm{m}} \cos^4 \theta \tag{96.4}$$

where
E_{m} is the illuminance at the center of the display
θ is the viewing angle

Note that two viewing angles—at the eye and at the display—have been defined. *Viewing distance* is determined primarily by the minimum size requirements (i.e., visual angle) for objects that the user must see. A conventional reading distance is about 71 cm, although VDTs are frequently read at 46 cm. Most design criteria assume a viewing distance of between 50 and 70 cm.

Visual fatigue is an imprecise term but one in common use, referring to the annoyance, irritation, or discomfort associated with visual tasks performed under poor conditions or for extended periods of time. A common cause of visual fatigue is *glare*, which can be due to a combination of high-brightness, high-reflectance, and specular (mirrorlike) reflections causing source light to reflect directly into the eye. Minimizing or eliminating glare is essential for effective display performance and usually involves a thoughtful placement of the display, proper positioning of room lights, control of ambient light (e.g., window shades), or the use of display hoods.

96.5 Display Performance Considerations

96.5.1 Contrast

Contrast is the measure of the luminance difference between an object and its background. While different definitions exist in the literature, luminance contrast as adopted by the International Lighting Commission (CIE) is given as

$$C_R = \frac{\text{Luminance of brighter object} - \text{Luminance of darker object}}{\text{Luminance of brighter object}} \tag{96.5}$$

Lower luminance displays require greater contrast to achieve the same visibility of objects. The *contrast*, or *luminance*, *ratio* between two surfaces in the central field of vision (e.g., a display and the desk on which it rests) should be around 3:1, while the ratio between the central field and the surfaces farther away (e.g., around a room) can be as high as 10:1. Ratios greater than 10:1 can induce glare. The simplest methods for contrast enhancement are the use of hooded shades or displays that can be tilted away from the offending light. Contrast-enhancing *filters*, however, can be more effective. All filters involve reducing the amount of ambient light reflected back to the user while leaving the emitted light from the display content as unchanged as possible. Several strategies for filtering exist, including *etching* or *frosting* the display surface to break up and scatter specular reflections. *Neutral density filters* increase contrast by reducing the amount of light passing through them; ambient light must pass through twice before reaching the user's eye, while display content must only pass through once. Micromesh filters placed on the display surface limit light penetration, so only rays falling perpendicular to the mesh can penetrate; this stops both specular and diffuse reflections and increases contrast. *Circular polarizers* are neutral density filters that polarize incident light, which is then prevented from returning through the filter. *Quarter-wave thin-film coatings* interfere with both specular and diffuse reflections.

In recent years, advertising has distorted the meaning of the term contrast ratio by presenting numbers that display-technology companies calculate using unnatural conditions. In its most basic form, the *contrast ratio of a display* is the ratio of the luminance of a black part to the luminance of a white part, as perceived by an average human eye. The luminance of any light source is the power it emits weighted by the luminosity function, the approximate sensitivity of the human eye to each of the wavelengths of the visible spectrum. The SI unit of measurement of luminance is candela per square meter, cd/m^2. For comparison, a normal candle emits a luminous intensity of roughly one candela.

The problem with the term *contrast ratio* arises when the display specification presents a measurement taken in nonstandard—or even unfair—methods or conditions. In backlit display devices, like liquid-crystal displays (LCDs), the hardware can vary the intensity of the backlight depending on the signal to achieve brighter bright colors and darker dark colors. Thus, the first nonstandard method of measuring contrast ratio is to measure the ratio between a completely black video signals, such that the

device dims the backlight to its lower limit, and a completely white video signal, such that the device increases the backlight to its maximum output. Since normal viewing conditions involve some combination of dark and light parts of the display, the contrast ratio measured in this way will never be seen by viewers because the backlight can only have one output level. An even more unfair measurement technique involves turning the backlight off completely—allowing a ratio of infinity. A ratio measured this way clearly says nothing about normal operating conditions of the display. Some specifications advertise this as a *dynamic contrast ratio* and omit the proper standard—the *static contrast ratio*—since the latter number is so much smaller. Since different display manufacturers can have completely different methods to measuring their dynamic contrast ratio, these numbers are not comparable.

A method for measuring a proper contrast ratio in normal operating conditions involves the following. The display shows a pure-black and pure-white checkerboard test pattern set in a room with typical wall reflectivity. The light from the display that reflects off the walls and back onto the display will reduce the contrast of the display. However, measuring the ratio of black and white in this method provides a useful specification for the display's contrast ratio. Unfortunately, few computer-monitor and television (TV) manufacturers readily provide such specifications.

Despite this misinformation, it is possible to understand the mechanics of displays that can permit higher contrast ratios. For the most part, displays with similar technology will have similar contrast ratios. Display devices that can selectively darken portions of the backlighting or illumination while leaving other portions at full brightness will achieve higher contrast ratios than display devices that have a single backlight. Devices that can achieve this have mechanisms like independent display elements (LED displays), phosphor illumination (CRT displays, plasma displays), or matrix backlighting (like full-array LED-backlit LCDs). Other devices, such as CCFL-backlit or LED-edge-lit LCDs, are unable to darken only portions of the screen and will have a smaller contrast ratio.

96.5.2 Gray Scale

Gray scale refers to the number of luminance levels, or shades of gray, available in a display. The common definition is a luminance ratio of 1.4 between levels, although the eye can discriminate changes as small as 1.03 (a Weber's K value of 0.03). The number of gray shades is useful for evaluating the capability of a display to render pictorial information or the range of luminance levels that can be used for coding. The highest luminance level is determined by display capabilities, but the lowest level is determined by the luminance of the display surface when no signal is present. Bright light incident on the display can elevate this minimum level and reduce the number of usable gray shades.

96.5.3 Pixels

Many display parameters are stated in terms of the *pixel* (the name comes from "picture element"). The pixel is the smallest resolvable information element seen by the user. Pixels in trichromatic displays (composed of red, green, and blue [RGB] components) are themselves composed of three *subpixels*, sometimes referred to as dots that are designed to be so small that the eye only perceives their combined color, the color of the pixel.

96.5.4 Resolution and Aspect Ratio

Resolution is a measure of the size of the smallest resolvable object on a display and is expressed as display lines per centimeter or pixels per square centimeter. Another measurement is *pixel pitch*, which is the distance between adjacent pixels (or between the same points in the recurring pattern of component-color subpixels). The specification of a display's resolution allows the differentiation between, for example, standard-definition (SD) displays and high-definition (HD) displays. HD devices are desirable for their refined display of detailed images and video. In comparison to HD displays, an

TABLE 96.1 Chart Comparing a Few Video Standards

Video Standard	Familiar Ratio	Comparable Resolution	Normalized Ratio
SDTV and standard computer monitors (VGA)	4:3	640 × 480	1.33:1
HDTV	16:9	1920 × 1080	1.78:1
Wide-screen computer monitors (WUXGA)	8:5 or 16:10	1920 × 1200	1.6:1
Wide-screen cinema	2.39:1	2870 × 1200	2.39:1

There is a one-to-one mapping between a *GA standard (like VGA or XGA) and a resolution. There are theoretically infinitely many-to-one mappings between resolutions and aspect ratios; however, manufacturers tend to make large enough gaps in product resolutions to warrant separate marketing, design, advertising, etc. These examples are all very close in size except for VGA, which has been the norm for decades.

image on an SD display of comparable size can appear blurry and lacking, especially when comparing highly detailed imagery such as a dense foliage or a person's head of hair.

Resolution is also useful for describing the size of the display, when considering computer monitors and TVs, which have microscopic, square pixels. This is because resolutions for these and similar devices come in standardized *aspect ratios*. This is to facilitate comparison of similar devices by consumers but mostly due to standardization of data encoding and decoding of computer-graphics and TV media. The aspect ratio is the relationship between the width and the height of the display and is presented in a format similar to "16:9," which is the format of standardized wide-screen high-definition television (HDTV). For comparison, modern digital projection shown on wide screens in cinemas has an aspect ratio of 2.4:1, which is 135% the width of HDTV. Table 96.1 lists a few aspect-ratio standards to demonstrate the differences between the aspect ratios and some associated resolutions.

Traditionally, computer monitors have had aspect ratios of 4:3. Although, recent consumer trends have seen wide-screen versions that match the 16:9 size of HDTV and a 8:5 (also written as 16:10 to be more easily compared with the 16:9 ratio of HDTV), which is slightly taller, allowing for more real estate with which to work on a display of a comparable resolution. The tiers of resolution appear where nontrivial technology and manufacturing differences, or distinct market segment needs, are required to achieve a greater resolution, which translates into major differences in cost to the end user. For example, when selecting a display, devices of roughly comparable prices will have formats of HD 1080, UXGA, WUXGA, or 2 K. The next tier in price will contain displays with formats of QXGA or WQXGA. The next lower tier will contain displays with formats of HD 720, XGA, SXGA, or WXGA. The diagonal sizes are all very similar within these groups, but shifting between the aspect ratios may be important, depending on the intended use of the display. Figure 96.1 compares the sizes of a number of these resolution standards; the tiers are revealed by the gaps around horizontal sizes of 900, 1100, 1500, 1700, and 2300.

These "resolutions" are somewhat misleading since these standards are independent of screen size. An 81.3 cm (32 in.) display at 480i would have fewer phosphors illuminated per inch than a 50.8 cm (20 in.) display, simply because the same number of horizontal lines must cover more space in the larger display. However, current product performance has not generated a need for an intermediate-definition display for larger screen sizes or a different measuring standard.

96.5.5 Sharpness and Legibility

Although *sharpness* and its converse *blur* are normally defined by the subjective reaction of the display user, sharpness has been formally measured as the ratio of the blurred border zone of letters to their stroke width [3]. *Legibility* is related to character quality, or readability, and depends on the sharpness of characters.

The recent proliferation of LCD devices has introduced a large number of display devices with discrete, sharp pixels—in contrast to other display technologies that have some natural blending

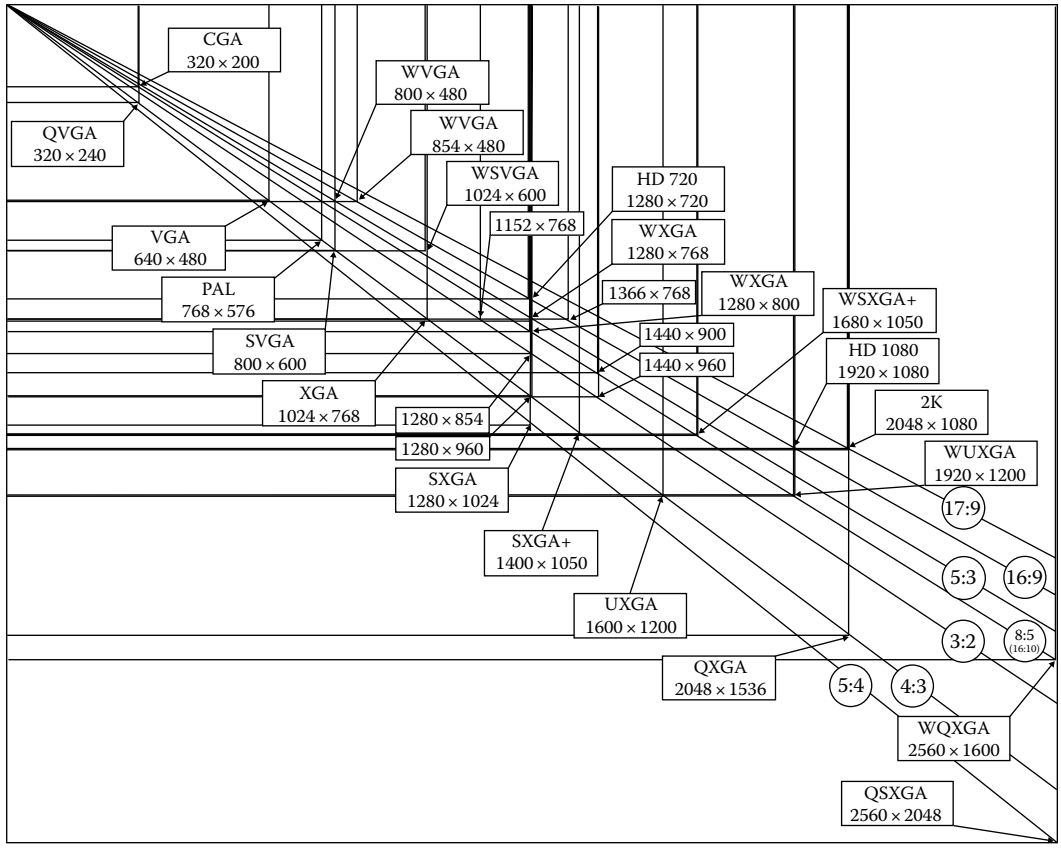


FIGURE 96.1 Variety of aspect ratios and resolutions, showing how they compare in size. The 4:3 aspect ratio has been the historical norm for computer monitors and TVs; however, the popularization of wide-screen formats, especially 16:9 and 8:5, is slowly becoming the new norm. Many dimensions in the variety are used in more than one aspect ratio, such as the vertical dimension of 768—used in 3:2, 4:3, 5:3, and 16:9. This makes it convenient for users when comparing screen resolutions. (Adapted from Steve Pomeroy's similar diagram, retrieved from http://en.wikipedia.org/wiki/File:Vector_Video_Standards2.svg)

between adjacent pixels. LCD devices that display small text (~12 pt fonts) naturally suffer from degraded readability of text, due to the pixels interfering with the shapes of the small letters. The effect causes jagged blocks in the letters where straight lines are not possible due to the pixels in the display. Although this effect is slight, it can degrade the readability of text on the display. One solution is to increase the resolution of the display, although this is usually not cost-effective for improving readability of text alone. Some computer software mitigates this effect by *antialiasing* the text that is displayed. Antialiasing blends pixels together (in effect, creating gray where previously there was only pure black or white) to suggest angled lines and curves that black-and-white pixels cannot display smoothly. Going further, Microsoft's ClearType, Adobe's CoolType, and Apple's Quartz employ *subpixel rendering* to improve type rendering beyond the capabilities of regular antialiasing. Subpixel rendering is not constrained to working with complete pixels and can render any of RGB, GBR, or BRG as white. This allows text to be refined even further than regular antialiasing by using subpixels; the increased resolution is intended to make the readability of text on LCDs approach the readability of printed text on paper [4]. When a discrete-pixel display technology like an LCD is used for extended periods of reading text, subpixel rendering will usually improve readability and reduce fatigue.

96.5.6 Gamut and Color Spaces

Display devices are often discussed in terms of their color reproduction. *Color reproduction* is the similarity between the source color (real-life or high-quality film photograph) and the display color. The *color space* to which humans are sensitive is describable with a convex, smooth mathematical function, called the CIE 1931 XYZ color space. The CIE created this color space in 1931 by combining the visible spectrum (435.8–700 nm) with the photopic luminosity function. The luminosity function is the approximating curve of the human eye's sensitivity to the visible spectrum in normal lighting conditions, with the familiar peak in the green region; the photopic version is for vision in well-lit conditions, as opposed to when the eyes adjust for seeing in the dark (scotopic). The smooth mathematical function is filled to create the color space by mixing adjacent colors, all the way from the curve of the color space (monochromatic light) to the center, white point where all colors mix to form white. In essence, the CIE 1931 color space can describe all combinations of all colors in the visible spectrum.

Decomposing color into a finite (usually small) number of components limits the mixing of colors that can occur. The color *gamut* of a regular trichromatic display device fills a triangle (3 vertices, 1 for each color) that covers some fraction of the CIE 1931 color space. Any given RGB color space completely describes the colors within the triangle but is unable to describe any color outside of the triangle because it requires the mixing of another component of color. The gamuts for SDTV (Rec. 601), HDTV (Rec. 709), and computer graphics (sRGB) are derived from the gamut of CRT displays and cover 35% [5] of the CIE 1931 color space. Color spaces designed especially for professional artists and photographers such as Adobe RGB and Kodak's ProPhoto RGB cover 50% [5] and roughly 80%, respectively, of the CIE 1931 color space. Figure 96.2 illustrates sRGB, Adobe RGB, and ProPhoto as three example color spaces. Most computer monitors follow the sRGB norm, but a few computer monitors are available that display the Adobe RGB color space.

When considering a display device and a color space, it is important to remember that the display device can only display the data given to it. In other words, if you have a digital photograph recorded in the sRGB color space, displaying it on a professional Adobe RGB monitor will not show anything different from displaying the photograph on a household sRGB monitor because its colors have already been clipped—to borrow terminology from audio recording. Similarly, HDTV data are encoded in the same primary colors as sRGB. Adding a yellow component introduces an extra point with which to mix colors and extends the space outside of the HDTV triangle. But if the yellow does not exist in the HDTV specification, then the RYGB display device must extrapolate or not display at all the yellows outside of the sRGB color triangle. Despite sRGB's small coverage of the CIE 1931 color space, it still incorporates most of the colors that exist in the real world. Since natural, real-life colors are not fully saturated and monochromatic, a display gamut that extends to the edges of the CIE 1931 color space is only further capable of displaying unnatural colors. Therefore, given how small the space is between the HDTV/sRGB color space and the CIE 1931 color space, adding yellows (even if extrapolating from HDTV data) does not make colors more lifelike, just oversaturated [6].

96.5.7 Flicker

Flicker is the term for detectable changes in display luminance and occurs when the frequency of those changes is below the integrating capability of the eye. The minimum frequency at which this occurs is the *critical flicker fusion frequency*, or CFF, which depends on the luminance level of the image; that is, displays that do not flicker at high luminance levels may still flicker at low levels. The CFF is calculated as

$$\mathrm{CFF} = a \log L_a + b \tag{96.6}$$

where $a = 12.5$ for high (photopic) ambient light levels and 1.5 for low (scotopic) levels; L_a is the average luminance of the image in cd/m^2, or nits; and $b = 37$. This is an empirical formula, and the values for a

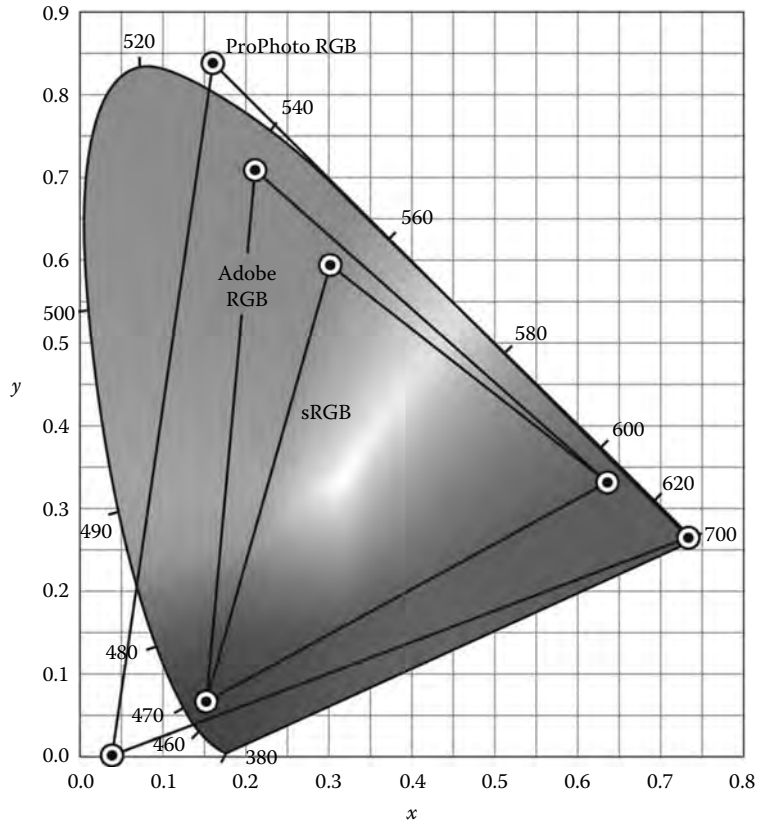


FIGURE 96.2 The ProPhoto, Adobe RGB, and sRGB gamuts overlaid on the CIE 1931 color space. The colors shown in the color space are for representation only. The vertices of the triangles of each of the gamuts represent the locations of the component RGB colors in the CIE 1931 color space. sRGB covers 35% [5], Adobe RGB 50% [5], and ProPhoto roughly 80%, of the CIE 1931 color space. Adobe RGB and sRGB share their red and green components. (Adapted from user "Spigget's" sRGB diagram, user "Juanjfb's" Adobe RGB diagram, and user "Entirety's" ProPhoto diagram, available from http://en.wikipedia.org/wiki/File:Cie_Chart_with_sRGB_gamut_by_spigget.png, http://en.wikipedia.org/wiki/File:CIExy1931_AdobeRGB.png, and http://en.wikipedia.org/wiki/File:CIExy1931_ProPhoto.png)

and *b* are only approximate. Because the eye cannot adapt to flicker fast enough to control the light on the retina, visual irritation usually occurs where flicker is present.

96.5.8 Refresh Rate

Refresh rate is the frequency with which display pixels are re-illuminated. Refresh rates below 50–60 Hz may induce perceptible flicker. The *update rate* is the frequency with which the information content of display is changed and is usually dependent only on the medium through which the content is transmitted to the display and any industry standards for the transmission or medium.

CRT displays are mostly interlaced SD, which for NTSC is 480i (actually 486 scan lines interlaced by two passes of 243 lines each) and for PAL is 576i (two interlaced passes of 288 lines each). HD displays have resolutions of either 720 or 1080 lines, interlaced or progressive. When a display is interlaced, its refresh rate still measures the redrawing time of the entire screen.

Refresh rates become extremely important when one display's content is to be video recorded (and displayed) on a second display. In these situations, the shutter speed (and thus frame rate) of the

recording device may run at a different speed than the refresh rate of the first display. On the second display, the beat frequency may be visible, manifesting somewhere between a brighter band that slowly cycles across the recorded display content and a dramatic, rapid alternating flickering of the bottom and top portions of the screen that makes the recorded display's content difficult and discomforting to view. Since many display refresh rates are not user adjustable, the video recording device needs to have an adjustable shutter speed in this situation.

96.6 Display Content Considerations

Configurable software packages for scientific measurement (e.g., LabView™) allow great flexibility in the design of display formats. Human factors principles for display content, therefore, are as important to effective measurement as the electronic characteristics of the display itself. The following principles are an introduction to the kinds of issues that must be considered when designing display content. The interested reader is referred to Helander (1987) and Grandjean [3] for information on human factors and design guidelines beyond those presented here.

96.6.1 Alphanumeric Displays

The size of a letter is its *pitch*. A general recommendation is that characters should subtend a minimum of about 12 min of arc at common reading distances. Alphanumeric displays are usually constructed of pixel arrays or segmented bars. A 5×7 pixel array is considered the minimum necessary to represent the ASCII character set.

Stroke width is the ratio of the thickness of the stroke to the height of the letter. Recommendations for legibility are 1:6 to 8 or 1:8 to 10 [2]. As illumination is reduced, thick letters are easier to read than thin ones. With low illumination and low contrast, letters should be boldface type with low stroke width–height ratios (e.g., 1:5).

96.6.2 Quantitative and Status Displays

While numeric readout displays are easy to implement in software, analog pointer displays show an advantage when it is important to observe the direction or rate of change of the values presented [7]. If the measurement application involves "more or less" or "up or down" interpretations, a straight-line or thermometer scale is preferred because it shows the measurement in relation to zero. *Moving pointers* are better able to convey minor changes in readings than fixed pointers or numeric readouts. *Scale markings* should reflect the length of the smallest unit to be read. It is desirable to have a marker for each unit, so that no interpolation is required.

Check reading indicators are used to determine whether a condition is "normal." The normal criterion point, therefore, should be clearly coded. If several indicators are grouped together, they should be arranged so that the deviant reading stands out (e.g., by indicating a different column height or dial angle).

Color is an excellent method for organizing information on a display and for locating objects rapidly. Although display users can distinguish between many different colors, they usually cannot remember more than seven to ten of them, so the number should be limited if color is going to be used as a coding dimension.

96.6.3 Summary

The next sections address different display technologies in light of the principles discussed here. While display guidelines are available for essentially any parameter, it is important to remember that visual perception is an integrative process. No single guideline functions alone, and display quality is usually a product of interacting needs and trade-offs.

96.7 Cathode-Ray Tube Displays

The cathode-ray tube (CRT, see Part X of *Electromagnetic, Optical, Radiation, Chemical, and Biomedical Measurement*) in the last decade has seen its ubiquity in electronics stores replaced by the newer solid-state active-matrix LCD and LED display devices and the revisited plasma display. The primary reason for this is size and weight: the traditional CRT displays that were used for every TV and computer monitor require a deep box, often deeper than the screen's diagonal dimension. "Flat" CRT displays have only recently hit the shop floor to catch up to real flat-panel displays: LED, LCD, and plasma displays. However, even "flat" CRT displays are several inches deeper than flat-panel displays and are limited in their diagonal dimension by production costs. Commercial advertising has a major role in shifting consumer preference from CRT to flat-panel TVs and computer monitors. Although a majority of these devices still in use are CRT, device depth and *perceived* technology advancements (some of which are not validated from a technical perspective on human perceptual capabilities) are rapidly influencing the proliferation of flat-panel display technologies.

CRT displays have good contrast ratio, brightness, refresh rate, and product lifetime. The sRGB color space (made ubiquitous by the World Wide Web) was defined based on CRT capabilities, and other display technologies like LCD and plasma follow the sRGB standard, although there are specialized devices that have better gamuts. What sets CRT displays apart from other display technologies is their black levels, which are darker than LCDs and plasma. This is because the CRT can selectively not illuminate the black areas of the screen's content, unlike LCDs, which have a backlight, and plasma displays, which require a small activation charge. For most users, plasma displays will have comparable black levels, although OLED displays that have just recently been made available to consumers promise black levels just as good as CRT displays. On the other hand, CRTs have generally lower luminance compared to other technologies. A typical CRT display may not have a luminance as high as 200 cd/m^2, while LCD and plasma displays easily achieve over that number.

Other than the aforementioned disadvantage of size and weight, CRT screens have also been, in the past, very susceptible to being washed out by ambient light. The ubiquity of smooth glass in CRT displays drove the development of attachable glare filters and the application of antireflective coatings to CRT displays. The former especially can also block a lot of illumination from the display, reducing the contrast ratio and the number of colors that the user can perceive from the display. Finally, the use of CRT displays with continuous exposure to the same screen or user-interface elements (such as on-screen menus, scoreboards, and row or column labels) causes a noticeable "burn-in" of those elements. The phosphors that have displayed those elements for extended time respond poorly to the CRT mechanism, causing colors to be misrepresented in those areas of the screen permanently. Screen burn-in can make a CRT display unusable after hundreds of hours of presentation of a single static image. However, the use of screen savers and multifunction displays has greatly reduced the likelihood of burn-in in modern CRTs compared to early monochrome monitors.

96.7.1 CRT Performance

CRT resolution has plateaued and been standardized to such consistency that the specific measurements are sometimes not even mentioned in product descriptions. CRT displays are mostly interlaced, meaning they require two passes of phosphor illumination to draw a complete screen. This is done at a high enough frequency that the redrawing is invisible to the user, although he or she may perceive a flicker. CRTs used for TVs are SD and interlaced. CRTs used for computer monitors can have resolutions of up to 1200 lines and are interlaced. Historically, CRTs have been manufactured in a very wide spectrum of sizes and resolutions for many different purposes.

A very important parameter of a CRT display is the refresh rate. Because of the mechanics of the phosphors and cathode-ray gun, the refreshing of an interlaced CRT display produces a flicker. On modern screens, the refresh rate is usually one of the standard 60 or 75 Hz. However, because this

measurement refers to the redrawing of the entire screen, due to interlacing, this means alternating halves of the screen is updated 30 or 37 times/s. To many users, CRT displays produce a noticeable flicker that can be fatiguing on the eyes (especially in cases of presenting rapid motion or extremely fine detail). If screen content is to be recorded with a video camera, CRT flicker may make the screen content very difficult to see due to the beat frequency between the video camera's frame rate and the CRT refresh rate.

Contrast ratio is diminished by high ambient light levels, and it is often necessary to compromise between light requirements for work tasks and light levels for optimum CRT visibility. In low ambient lighting conditions, a CRT contrast ratio of 10:1 is usually attainable. Higher ratios of 25:1 or more can be achieved with contrast-enhancing filters but at the expense of brightness and increased power requirements.

96.8 Liquid Crystal Displays

LCDs (see Part X of *Electromagnetic, Optical, Radiation, Chemical, and Biomedical Measurement*) are nonemissive: they do not generate light of their own but control the transmission or reflection of another light source—ambient or built into the same device as the display itself. Many modern LCDs include backlighting, such as cold-cathode fluorescent lamps (CCFLs) or light-emitting diodes (LEDs), to provide ample brightness to the display without the need for often-glare-inducing ambient light. Backlighting also allows the LCD to display color effectively. The simple addition of high-brightness backlighting and an active addressing matrix (rather than passive) allowed LCD technology to go from use in monochrome 7-segment displays to large-screen HDTV capable of displaying one billion colors.

Internally, the liquid-crystal material within an LCD alters the optical path of light when an electric field is placed across it. In a passive-matrix LCD, the material must be stable enough to not require an electric charge to maintain its state. Because of this, passive-matrix LCDs suffer from slow response times. In an active-matrix LCD (AMLCD), an active potential is used to maintain the state of the pixel. An unstable material is desirable so that response times can be as quick as possible.

The principal advantages of LCDs are (1) low power consumption (which is important for battery-operated systems), (2) a flat-panel display package, (3) low cost of the basic materials, and (4) good contrast in high ambient illumination. As mentioned, passive-matrix LCDs can suffer from slow dynamic response for switching display elements on and off: 100–500 ms response times, for example, are visually noticeable and such systems may be unacceptable for many applications. However, for many TVs and computer monitors, AMLCDs have been engineered with response rates sufficient to allow the device to operate with a refresh rate of 120 Hz. Without a backlight, LCDs can be difficult or impossible to read in low light conditions—consider reading a digital watch at night before pressing its light button. Finally, the viewing angle is limited by inherent characteristics with some LCD technologies. Many LCD features, such as switching thresholds and response times, are temperature dependent. Devices with backlights and other electronic components that generate heat (this includes TVs and computer monitors) usually will not suffer from temperature problems in everyday, inside use.

96.8.1 LCD Performance

A full range of resolution capabilities is available, from simple alphanumeric displays to systems with over 5 million pixels and resolutions of 128 pixels per cm (326 ppi) [8]. LCDs have been used in small-display applications, like calculators and watches, all the way through a 274 cm (108 in.) diagonal, full-color HDTV [9]. The difficulty in fabricating the sheet of glass and the TFT matrix useful for an LCD (i.e., with few defects) makes displays of this size remarkable. LCD displays that have been made specifically for signage have a brightness of 700 cd/m^2 [10] for making them clearly visible in locations with bright ambient lighting.

Color in LCDs is achieved by placing a color mosaic over the LCD and switching the cells behind the proper combinations (in RGB) to display color. Modern color displays have colored cells that are small

enough that they cannot actually be resolved. Modern color LCDs are AMLCDs, which employ the use of thin-film transistors (TFTs) as the switching technology, where each red, green, or blue cell is accompanied by a transistor controlling the liquid-crystal material.

96.8.2 TN-Effect LCDs

The matrixes of TFTs that control the LCDs come in several variations. The most common right now utilizes the twisted-nematic (TN) effect, which has been around for many decades for use in calculators, clocks, digital watches, and full-color, high-resolution displays more recently. Backlit displays of this type are easily capable of achieving contrast ratios up to 1000:1 (and luminance commonly between 250 and 300 cd/m^2). In nonbacklit passive-matrix devices, much lower contrast ratios (less than 80:1) are common, but improved capabilities with contrast ratios exceeding 500:1 have been prototyped by late 2010. TN-effect displays suffer from viewing angle limitations, especially in the vertical direction. A large positive viewing angle or a large negative viewing angle results in loss of contrast, distortion of colors, and thus reduced legibility. This makes it very important to install or use TN-effect LCDs at an appropriate viewing angle. The standard ergonomic solution of a downward 15°–20° viewing angle from the user (i.e., measured from the floor horizontal down to the center of the screen), and screen normal tilted to match, works well for TN-effect displays for both user comfort and proper color and contrast display. Aside from viewing angle color shortcomings, TN-effect displays are only capable of 6 bits per color channel, which means they are unable to reproduce the 24 bit color that computer graphics are capable of generating. To accommodate for this shortcoming, TN-effect displays employ dithering (blending colors by using adjacent pixels) to simulate the missing colors. Professional artists and photographers usually avoid TN-effect displays for this reason. TN-effect displays are also insufficient for multiple-user settings because of the aforementioned viewing angle limitations.

96.8.3 IPS LCDs

Designers intended a newer LCD technology, called in-plane switching (IPS), to improve upon the limitations in viewing angle and the poor color reproduction of TN-effect displays. The state of the art in IPS displays achieves 10 bit per color channels, which is capable of over one billion colors. IPS does not suffer from the same color distortion from large horizontal viewing angles that TN-effect displays have. Their superior color reproduction makes IPS displays much more attractive to professionals, artists, and photographers. Originally, IPS displays had poor pixel response times, but recent engineering improvements have made them comparable (around 5 ms) to TN-effect displays. Altogether, IPS displays are superior to TN-effect displays in everything except for their price to the end user. Users who need the excellent color reproduction or viewing angles (e.g., in multiple-user settings) should find the increased price well worth the superior product.

96.8.4 VA LCDs

A third TFT technology is vertical alignment (VA), which comes in multidomain VA (MVA) and patterned VA (PVA). For the most part, VA displays fall in between TN-effect and IPS displays in terms of viewing angles and color reproduction.

96.9 Plasma Displays

Currently, plasma technology (see Part X of *Electromagnetic, Optical, Radiation, Chemical, and Biomedical Measurement*) is used almost exclusively in large-format displays. In the past, monochrome plasma displays were available for portable, laptop, and handheld computers, but advancements in full-color technologies like CRT displays and LCDs dominated the use of plasma displays in these sizes.

Now, plasma displays share the market of TVs of the largest sizes, 100 cm (39 in.) and larger, with LCDs. Cost-effective manufacturing of full-color plasma displays limits the minimum size of the display, so laptop displays and displays in mobile phones—roughly 46.7 cm (18.4 in.) diagonally and smaller—all have LCDs or OLED displays.

The advantages of plasma displays include brightness, luminous efficiency, and dark display of black (as opposed to a washed-out gray). This advantage also allows for excellent contrast ratios in high ambient illumination. However, as discussed earlier, the manufacturer reports of very high contrast ratios should be considered with care, due to the limited human perceptual ability to meaningfully distinguish these ratios in real-world settings. Plasma displays also exhibit long display life and tolerate high shock and vibration environments.

Some disadvantages of plasma displays include higher cost, relative to CRTs, and higher power consumption, relative to LCDs. Similarly sized (107 cm, or 42 in., diagonal) CCFL-backlit LCD and plasma displays can consume 199 W [11] and 327 W [12], respectively.

96.9.1 Plasma Display Performance

The largest plasma display built through 2010 is 386 cm (152 in.) diagonal with a 4 K resolution, 4096 × 2160 [13], which is four times the HDTV standard, almost nine million pixels. Plasma TVs are commercially available in sizes generally greater than 76 cm (30 in.) and up to 275 cm (108 in.). Displays of the largest sizes, advertised for home-theater use, have the best contrast ratios among plasma displays. Advancements in plasma technology have moved plasma into the first place among common display technologies for luminance, with some displays achieving 1300 cd/m^2 [14].

Plasma technology involves similar refresh rates to LCDs, between 60 and 120 Hz. Some advertisements promote plasma displays with "subfield" refresh rates of 480 or even 600 Hz, but this is a misrepresentation [15]. Due to the brief illumination time of subpixels in plasma displays (about 2 ms), maintaining a frame for 1/60s (i.e., displaying a single frame for a full cycle at 60 Hz) requires pulsing the subpixels eight or ten times during the frame, resulting in a misleading 480 or 600 Hz refresh rate, respectively. In fact, the frame is not fully redrawn that quickly—most of the time; the video signal is not even presented to the display this quickly. However, the very short 2 ms duration of the excited states means that plasma displays have nearly no ghosting or blurring when pixels change colors, which is desirable for displaying video containing high movement speeds.

96.10 Electroluminescent Displays

With the exception of LEDs, electroluminescent (EL, see Part X of *Electromagnetic, Optical, Radiation, Chemical, and Biomedical Measurement*) technologies are not as prominent in the commercial arena as other types of display systems. EL materials are complex (i.e., driven and controlled by processes related to solid-state semiconductor physics) and are more difficult to work with than other display materials. Nevertheless, they offer great potential for high brightness and low cost that deserves consideration, especially as new designs become available. Matrix addressing is used for control of information display applications. EL materials are applied in two forms—powders (PEL) and thin-films (TFEL)—and are controlled by both AC and DC voltages, generating four basic design approaches. Some advantages of EL displays are (1) high luminous efficiency (except ACPEL designs); (2) readability in sunlight, (3) color capability, (4) compact, flat-panel designs, and (5) significant potential for low-cost manufacture. A disadvantage of EL displays is that ACPEL systems have low luminance and contrast ratio. In addition, phosphors in powder designs scatter and reflect ambient light, reducing their contrast.

96.10.1 EL Display Performance

Contrast ratios of 50 to 150:1 have been demonstrated with monochrome ACTFEL systems, while 15 to 20:1 have been achieved with DC designs. ACTFEL designs also show excellent brightness, with

demonstrated luminance levels of over 157 cd/m^2 (monochrome) and 26 cd/m^2 (color). DCPEL designs, representing newer technologies, have achieved over 100 cd/m^2 with monochrome designs but may soon meet or exceed AC-based values. The resolution of EL displays is limited by the duty factor of matrix addressing; a finite amount of time is needed to energize each row, and a minimum luminance level is needed for adequate display performance, so the remaining variable becomes the number of addressable lines. A demonstrated ACTFEL display with 640 × 400 elements, with six colors and a resolution of 27 lines/cm, is typical of this technology.

96.11 Light-Emitting Diode Displays

LED displays (see Part X of *Electromagnetic, Optical, Radiation, Chemical, and Biomedical Measurement*) involve single-crystal phosphor materials, which distinguishes them from the polycrystal EL materials discussed in the previous section. The basic physics behind their operation is, however, quite similar. LED displays are highly versatile and well suited to a variety of measurement applications. The advantages of LED displays include high reliability and graceful degrades; individual LED elements can fail without affecting overall display performance. LEDs are rugged, for operation in harsh environments, and they are more tolerant of temperature extremes than other technologies. LEDs demonstrate better viewing angles than LCDs and excellent brightness for visibility in sunlight. Unfortunately, LED displays also have high power consumption when packaged in large, flat-panel displays, and the cost is high for the complex assembly. Optical cross talk between array elements can occur if transparent substrates are used. LEDs are the most-restricted display in terms of color range (e.g., no blue device is commercially available).

96.11.1 LED Display Performance

LED devices have excellent brightness, but because display brightness is also a function of the filters or magnification lenses used over the LED elements, device luminance is not, by itself, a reliable measure of overall system performance. LED displays also show very good luminance contrast. *Chrominance contrast*—the color difference between the LED and its background— however, is a factor in evaluating LED performance that is not found in other technologies. Chrominance contrast is significant because of the high saturation of most LED phosphors. It is affected by display filters and can have significantly more influence on display performance than luminance contrast.

LED displays with resolutions of 20–25 lines/cm have been constructed. Flat-panel displays of 38,400 discrete elements have also been demonstrated with luminance levels of around 137 cd/m^2 (increasing to 240 cd/m^2 with reduced resolution), and at least one aircraft display with 49,000 elements has been built. CRT-equivalent displays with 600 × 400 elements have also been realized with engineering development models.

96.12 Electronic Paper/Electronic Ink

The newest technology to emerge in the past few decades is electronic paper. The researchers currently improving electronic-paper technologies are working to provide a dynamic display that rivals paper in readability and portability (i.e., paper requires no power source). Electronic paper can display content forever with no power required to sustain the image, only change it. Thus, it is possible with electronic paper to provide a single sheet (currently more like a thin tablet) that can display every page of every book, with the proper data connection providing digital versions of the books.

The primary advantage of electronic paper is its readability. The intention is to mimic the way paper is read: a reflective white surface and nonreflective content that provides contrast in almost any lighting (darkness makes electronic paper, like regular paper, unreadable). Currently, electronic paper cannot handle fast-moving images, due to its slow refresh rate.

The technology behind electronic paper is, appropriately enough, electronic ink. There are several different implementations of electronic paper. Either pigment particles are moved to the surface to replace the white background or two-sided particles (one side white and the other black) are rotated in a fluid to display the two colors. Color electronic paper employs either colored inks instead of black or some sort of filtering system. Note that colored inks mean the primaries are subtractive: cyan, magenta, and yellow instead of the additive primaries discussed in the rest of this chapter.

There are other technologies instead of using pigment particles that are also gaining attention, including using microelectromechanical systems to control interference patterns—like the iridescent wings of a butterfly—to display different wavelengths to the viewer, permitting a full-color display [16].

96.12.1 Electronic-Paper Display Performance

Electronic-paper research is diverse, in an attempt to find the best mechanisms for different applications of the technology: tablet computers, electronic books, mobile phones, signage, flexible displays, and others. Currently, the largest set of electronic-paper products is in electronic books, which mostly use the black-and-white mechanism described earlier. A major manufacturer of black-and-white and color electronic paper claims a minimum of 10:1 contrast ratio and viewing angles near 180°, two performance characteristics that make electronic paper so popular. One area in which electronic paper suffers is refresh rate. The mechanism to switch the pigments requires, according to the same manufacturer, between an eighth and a full second [17].

Defining Terms

Decay time: The time required for the peak brightness of a phosphor to drop to a defined fraction of peak luminance; a measure of how long the phosphor remains illuminated after being energized by the electron beam

Duty factor: The time spent addressing each pixel during a refresh cycle; inversely proportional to the number of pixels

Font: Refers to the graphic in which alphanumeric characters and symbols are presented by computers

Spot size: The sizes of the illuminated spot from an electron beam in a CRT; limits the size of the raster line

Wash out: The loss of contrast (i.e., reduction in dynamic range) in an LED as the ambient light reflected off the background of the display surface approaches the light level of the active area

References

1. S. Sherr, *Electronic Displays*, 2nd edn. New York: John Wiley & Sons, 1993.
2. M.S. Sanders and E.J. McCormick, *Human Factors in Engineering and Design*, 6th edn. New York: McGraw-Hill, 1987.
3. E. Grandjean, Design of VDT workstations. In: G. Salvendy (Ed.), *Handbook of Human Factors*, New York: John Wiley & Sons, pp. 1297–1359. 1987.
4. Microsoft Typography. (2002). What is clear type? Retrieved from http://www.microsoft.com/typography/whatiscleartype.mspx (accessed on March 14, 2013).
5. McHugh, S. (n.d.). sRGB vs. Adobe RGB 1998. Retrieved from http://www.cambridgeincolour.com/tutorials/sRGB-AdobeRGB1998.htm (accessed on March 14, 2013).
6. Soneira, R. (2010, July). Display Myths: Shattered. *Maximum PC,* 51. Retrieved from http://dl.maximumpc.com/Archives/MPC0710-web.pdf (accessed on March 14, 2013).
7. D.G. Payne, V.A. Lang, and J.M. Blackwell, Mixed vs. pure display format in integration and nonintegration visual display monitoring tasks, *Human Factors*, 37(3), 507–527, 1995.

8. Apple Inc. (2010). Learn about the high-resolution retina display. Retrieved from http://www.apple.com/iphone/features/retina-display.html (accessed on March 14, 2013).
9. Hanlon, M. (2007). Sharp showcases world's largest LCD TV—108 inches. Retrieved from http://www.gizmag.com/go/6704 (accessed on March 14, 2013).
10. Hewlett-Packard Development Company, L.P. (2011). HP LD4710 47-inch LCD digital signage display specifications. Retrieved from http://h10010.www1.hp.com/wwpc/us/en/sm/WF06a/382087-382087-64283-72270-3915216-5054199.html (accessed on March 14, 2013).
11. Panasonic Corporation. (2010). 2010 Viera U22 Series LCD. Retrieved from http://www2.panasonic.com/match-maker/assets/pdf/viera/2010/VieraU22SeriesLCD.pdf (accessed on March 14, 2013).
12. Panasonic Corporation. (2010). 2010 Viera G25 Series Plasma. Retrieved from http://www2.panasonic.com/match-maker/assets/pdf/viera/2010/VieraG25SeriesPlasmaSpec.pdf (accessed on March 14, 2013).
13. Panasonic Corporation. (2010). Panasonic develops world's largest 1520 inch full-HD 3D plasma display. http://www2.panasonic.com/webapp/wcs/stores/servlet/prModelDetail?storeId=11301&catalogId=13251&itemId=389530&modelNo=Content01062010103949121&surfModel=Content01062010103949121 (accessed on March 14, 2013).
14. SAMSUNG. (2011). 50″ Plasma display specifications. Retrieved from http://www.samsung.com/us/business/commercial-display-solutions/PH50KLPLBF/ZA (accessed on March 14, 2013).
15. Best-3DTVs.com. (2011). What does 600Hz sub-field drive mean? Retrieved from http://www.best-3dtvs.com/guides/what-does-600hz-sub-field-drive-mean/ (accessed on March 14, 2013).
16. Epaper Central. (2010). Spotlight on Mirasol: Qualcomm's full-color e-paper. Retrieved from http://www.epapercentral.com/spotlight-on-mirasol-qualcomms-full-color-epaper.htm (accessed on March 14, 2013).
17. E Ink Corporation. (2010). E Ink Triton Imaging Film. Retrieved from http://www.eink.com/sell_sheets/triton%20sell%20sheet.pdf (accessed on March 14, 2013).

Further Information

Grandjean, E., *Ergonomics in Computerized Offices*, London, U.K.: Taylor & Francis Group, 1987—is an excellent all-around treatise on the principles of effective VDT selection and use. Summarizes a wide range of research literature. If this volume is difficult to obtain, a chapter by the same author is also included in the *Handbook of Human Factors* (Reference 2).

Helander, M.G., Design of visual displays. In: G. Salvendy, (Ed.), *Handbook of Human Factors*, New York: John Wiley & Sons, 1987—is an excellent and concise review of major human factors principles for display design and use. Includes a critical review of the foundation literature in this area.

Sherr, S., *Electronic Displays*, 2nd edn., New York: John Wiley & Sons, 1993—offers clear presentations of all important display technologies, together with a good summary of performance measurement methods for display systems. Well illustrated with a variety of commercial products.

Tannas, L.E., Jr. (Ed.), *Flat-Panel Di splays and CRTs*, New York: Van Nostrand Reinhold, 1985—provides a thorough, yet highly readable examination of the physical principles behind essentially every major display technology. Although the technology capabilities have become dated since publication, this is well worth review.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Appendix: Units and Conversions

B.W. Petley
National Physical Laboratory

This appendix contains several tables that list the SI base units (Table A.1), define the SI base units (Table A.2), list their derived units (Table A.3), list their prefixes (Table A.4), and list their conversion units (Table A.5).

TABLE A.1 SI Base Units

Base Quantity	Name of Base Unit	Symbol
Length	Meter	m
Mass	Kilogram	kg
Time	Second	s
Electric current	Ampere	A
Thermodynamic temperature	Kelvin	K
Amount of substance	Mole	mol
Luminous intensity	Candela	cd

TABLE A.2 International Definitions of the SI Base Units[a]

Unit of length (m)

Meter is the length of the path traveled by light in vacuum during a time interval of 1/299 792 458 of a second (17th CGPM,[b] 1983, Resolution 1).

Note: The original international prototype, made of platinum–iridium, is kept at the BIPM[c] under conditions specified by the 1st CGPM in 1889.

Unit of mass (kg)

Kilogram is the unit of mass: it is equal to the mass of the international prototype of the kilogram (3rd CGPM, 1901).

Unit of time (s)

Second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom (13th CGPM, 1967, Resolution 1).

Unit of electric current (A)

Ampere is that constant current that, if maintained in two straight parallel conductors of infinite length, of negligible cross section, and placed 1 m apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} N/m of length (CIPM, 1946, Resolution 2 approved by the 9th CGPM, 1948).

Note: The expression "MKS unit of force" that occurs in the original text has been replaced here by "newton," a name adopted for this unit by the 9th CGPM (1948), Resolution 7.

Unit of thermodynamic temperature (K)

Kelvin, unit of thermodynamic temperature, is the fraction 1/273.16 of the thermodynamic temperature of the triple point of water (13th CGPM, 1967, Resolution 4).

The 13th CGPM (1967, Resolution 3) also decided that the unit kelvin and its symbol K should be used to express an interval or a difference in temperature.

Note: In addition to the thermodynamic temperature (symbol T), expressed in kelvin, use is also made of Celsius temperature (symbol t) defined by the equation.

$t = T - T_0$

where $T_0 = 273.15$ K by definition. To express Celsius temperature, the unit "degree Celsius" that is equal to the unit "kelvin" is used; in this case, "degree Celsius" is a special name used in place of "kelvin." An interval or difference of Celsius temperature can, however, be expressed in kelvins as well as degrees Celsius.

Unit of amount of substance (mol)

1. Mole is the amount of substance of a system that contains as many elementary entities as there are atoms in 0.012 kg of carbon-12.
2. When the mole is used, the elementary entities must be specified and may be atoms, molecules, ions, electrons, other particles, or specified groups of such particles.

In the definition of the mole, it is understood that unbound atoms of carbon-12, at rest and in their ground state, are referred to.

Note: This definition specifies at the same time the nature of the quantity whose unit is the mole.

Unit of luminous intensity (cd)

Candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} Hz and that has a radiant intensity in that direction of (1/683) watt per steradian (16th CGPM, 1979, Resolution 3).

[a] The United States denotes the unit of length by "meter" in place of the international usage of "meter."

[b] CGPM, Conférence Général de Poids et Mesures; CIPM, Comité International des Poids et Mesures.

[c] BIPM: Bureau International des Poids et Mesures.

TABLE A.3 SI-Derived Units with Special Names[a]

Derived Quantity	Name	Symbol	Expressed in Terms of Other Units	Expressed in Terms of SI Base Units
Plane angle	Radian	rad		$m\ m^{-1}$
Solid angle	Steradian	sr		$m^2\ m^{-2}$
Frequency	hertz	Hz		s^{-1}
Force	newton	N		$m\ kg\ s^{-2}$
Pressure, stress	pascal	Pa	$N\ m^{-2}$	$m^{-1}\ kg\ s^{-2}$
Energy, work, quantity of heat	joule	J		$m^2\ kg\ s^{-2}$
Power, radiant flux	watt	W		$m^2\ kg\ s^{-3}$
Electric charge, quantity of electricity	coulomb	C		s A
Electric potential, potential difference, electromotive force	volt	V	W/A	$m^2\ kg\ s^{-3}\ A^{-1}$
Capacitance	farad	F	C/V	$m^{-2}\ kg^{-1}\ s^4\ A^2$
Electric resistance	ohm	Ω	V/A	$m^2\ kg\ s^{-3}\ A^{-2}$
Electric conductance	siemens	S	A/V	$m^{-2}\ kg^{-1}\ s^3\ A^2$
Magnetic flux	weber	Wb	V s	$m^2\ kg\ s^{-2}\ A^{-1}$
Magnetic flux density	tesla	T	Wb/m^2	$kg\ s^{-2}\ A^{-1}$
Inductance	henry	H	Wb/A	$m^2\ kg\ s^{-2}\ A^{-2}$
Celsius temperature	Degree Celsius	°C		K
Luminous flux	Lumen	lm	cd sr	$cd\ m^2\ m^{-2} = cd$
Illuminance	Lux	lx	m^{-2} cd sr	m^{-2} cd
Activity (referred to a radio nuclide)	becquerel	Bq		s^{-1}
Absorbed dose, specific energy imparted, kerma	gray	Gy	J/kg	$m^2\ s^{-2}$
Dose equivalent, ambient dose equivalent, organ equivalent dose	sievert	Sr	J/kg	$m^2\ s^{-2}$

[a] Note that when a unit is named after a person, the *symbol* takes a capital letter and the name takes a lowercase letter.

TABLE A.4 SI Prefixes[a]

Factor	Prefix	Symbol	Factor	Prefix	Symbol
10^{24}	Yotta	Y	10^{-1}	Deci	d
10^{21}	Zetta	Z	10^{-2}	Centi	c
10^{18}	Exa	E	10^{-3}	Milli	m
10^{15}	Peta	P	10^{-6}	Micro	μ
10^{12}	Tera	T	10^{-9}	Nano	n
10^{9}	Giga	G	10^{-12}	Pico	p
10^{6}	Mega	M	10^{-15}	Femto	f
10^{3}	Kilo	k	10^{-18}	Atto	a
10^{2}	Hecto	h	10^{-21}	Zepto	z
10	Deca	da	10^{-24}	Yocto	y

[a] The 11th CGPM (1960, Resolution 12) adopted a first series of prefixes and symbols of prefixes to form the names and symbols of the decimal multiples and submultiples of SI units. Prefixes for 10^{-15} and 10^{-18} were added by the 12th CGPM (1964, Resolution 8), those for 10^{15} and 10^{18} were proposed by the 15th CGPM (1975, Resolution 10), and those for 10^{21}, 10^{24}, 10^{-21}, and 10^{-24} were proposed by the CIPM (1990) for approval by the 19th CGPM (1991).

TABLE A.5 Conversion Factors from English Measures to SI Units[a]

Unit	Equivalent
1. *Acceleration*	
Acceleration of free fall, standard gravity	9.806 65 m s^{-2}
1 ft/s^2	0.304 8 m s^{-2}
1 gal	0.01 m s^{-2}
2. *Angle*	
1 second (″)	$4.484\ 81 \times 10^{-6}$ rad
1 minute (′)	$2.908\ 9 \times 10^{4}$ rad
1 degree (°)	0.0174 532 rad
1 rad	206 264.8″
3. *Area*	
1 barn (b)	10^{-28} m^2
1 $in.^2$	$6.451\ 6 \times 10^{-4}$ m
1 ft^2	0.092 903 04 m^2
1 yd^2	0.836 127 36 m^2
1 are	100 m^2
1 acre [43560 (statute $ft)^2$]	4046.86 m^2
1 hectare	10,000 m^2
1 mi^2	$2.590\ 0 \times 10^{6}$ m^2
1 square mile (based on U.S. survey foot)	2.589 998 km^2
4. *Concentration, density, mass density*	
1 grain/gal (United States)	0.017 118 kg/m^3
1 lb/ft^3	16.018 46 kg/m^3
1 lb/gal (United States)	119.826 4 kg/m^3
1 short ton/yd^3	1186.6 kg/m^3
1 long ton/yd^3	1328.9 kg/m^3
1 oz (avdp)/$in.^3$	1730.0 kg/m^3
1 oz (avd)/gal (United States)	7.489 152 kg/m^3
1 $lb/in.^3$	27,680 kg/m^3
5. *Energy*	
1 ft lbf	1.355 818 J
1 cal_{th} (thermochemical calorie)	4.184 J
1 cal_{15} (15 °C calorie)	4.185 5 J
1 cal_{IT}[b]	4.186 8 J
1 kilocalorie (nutrition)[c]	4.186.8 J
1 watt second (W s)	1 J
1 watt hour (W h)	3600 J
1 therm (EC)	$1.055\ 06 \times 10^{8}$ J
1 therm (United States)	$1.054\ 804 \times 10^{8}$ J
1 ton TNT (equivalent)	4.184×10^{9} J
1 BT_{th}	1 054.350 J
1 Btu_{15}	1 054.728 J
1 Btu_{ST}	1 055.055 852 62 J
1 quad ($=10^{15}$ Btu)	$\approx 10^{18}$ J = 1 EJ

TABLE A.5 (continued) Conversion Factors from English Measures to SI Units[a]

Unit	Equivalent
6. *Force*	
1 dyne	10^{-5} N
1 ounce-force	0.278 013 9 N
1 pound-force	4.448 222 N
1 kilogram-force	9.806 65 N
1 kip (1000 lbf)	4448.222 N
1 ton-force (2000 lbf)	8.896 443 N
7. *Fuel consumption*	
1 gallon (United States) per horsepower hour	$1.410\,089 \times 10^{-9}$ m^3/J
1 gallon (United States)/mile	2.352 15 1/km
1 gallon (United Kingdom)/mile	2.824 81 1/km
1 mile/gallon (United States), mpg	0.425 144 km/L
1 mile/gallon (United Kingdom)	0.354 006 km/L
1 pound per horsepower	$1.689\,659 \times 10^{-7}$ kg/J
1 1/(100 km)	235.215/(mpg) (United States)
8. *Length*	
1 fermi	10^{-15} m = 1 fm
1 angstrom (Å)	10^{-10} m
1 microinch	2.54×10^{-8} m
1 mil	2.54×10^{-5} m
1 point (pt) [0.013837 in.][d]	—
1 pica (12 pt)	4.217 5 mm
1 inch (in.)	0.025 4 m
1 hand (4 in.)	0.101 6 m
1 foot (12 in.) (0.999998 statute ft.)	0.304 8 m
1 foot (U.S. survey)	0.304 800 6 m
1 statute foot [(1200/3937) m]	0.304 800 6 m
1 yard (yd)	0.914 4 m
1 fathom (6 ft, U.S. survey)	1.828 8 m
1 rod (16.5 statute ft)	5.029 2 m
1 chain (4 rod)	20.116 8 m
1 furlong (10 chain)	201.168 m
1 mile (8 furlong, 5280 ft)	1609.344 m
1 statute mile (8 furlong, 5280 statute ft)	1609.347 2 m
1 nautical mile (international)[e]	1852 m
1 light-year[f]	$9.640\,73 \times 10^{15}$ m
9. *Light*	
1 foot-candle	10.763 91 lx
1 phot	10,000 lx
1 cd/in.^{2}	1550.003 cd/m^2
1 footlambert	3.426 259 cd/m^2
1 lambert	3183.099 cd/m^2
1 stilb	10,000 cd/m^2

(*continued*)

TABLE A.5 (continued) Conversion Factors from English Measures to SI Units[a]

Unit	Equivalent
10. *Mass*	
1 pound (avdp.) (lb) (7000 g)	0.453 592 37 kg
1 pound (troy) (5760 g)	0.373 241 721 6 kg
1 grain (g)	64.798 91 mg
1 scruple (20 g)	1.296 0 g
1 pennyweight (24 g)	1.555 174 g
1 dram (60 g)	3.887 9 g
1 ounce (avdp) (437.5 g)	28.349 52 g
1 ounce (troy) (480 g)	31.103 48 g
1 carat (metric)	0.2 g
1 stone (14 lb)	6.350 29 kg
1 slug	14.593 9 kg
1 hundredweight (long)	50.802 35 kg
1 ton (short) (2000 lb)	907.184 7 kg
1 ton (long) (2240 lb)	1016.047 kg
	1.016 047 t
Mass per unit length	
1 tex	10^{-6} kg/m
1 denier	$1.111\,111 \times 10^{-7}$ kg/m
1 pound per foot	1.488 164 kg/m
1 pound per inch	17.857 97 kg/m
1 ton/mile	0.631 342 mg/km
1 ton/1000 yd	1.111 6 kg/m
1 lb/ft	1.488 16 kg/m
Mass per unit area	
1 ton/mile^2	3.922 98 kg/ha
1 ton/acre	2510.71 kg/ha
1 oz/yd^2	33.905 7 g/m^2
Mass carried × distance (traffic factor)	
1 ton mile	1635.17 kg km
Mass carried × distance/volume (traffic factor)	
1 ton mile/gal (United States)	431.967 6 mg km/m^3
11. *Power*	
1 erg/s	10^{-7} W
1 ft lbf/h	$3.766\,161 \times 10^{-4}$ W
(1 Btu_{ST})	1.000 669 Btu_{th}
1 metric horsepower (force de cheval)	735.498 8 W
1 horsepower (550 ft lbf/s)	745.70 W
1 electric horsepower	746 W
12. *Pressure, stress*	
1 standard atmosphere	101 325 Pa
1 dyne/cm^2	0.1 Pa
1 torr [(101th 325/760) Pa]	133.322 4 Pa
1 N/cm^2	10,000 Pa
1 bar	100,000 Pa
1 lbf/ft^2	47.880 26 Pa
1 lbf/in.^{2} (psi)	6894.8 Pa

TABLE A.5 (continued) Conversion Factors from English Measures to SI Units[a]

Unit	Equivalent
1 kgf/cm^2	98 066.5 Pa
1 cm water (4 °C)	98.063 7 Pa
1 mm of mercury (0 °C)	133.322 4 Pa
1 in of water (39.2 °F)	249.082 Pa
1 in of mercury (60 °F)	3376.85 Pa
1 ft water (39.2 °F)	2988.98 Pa
13. *Thermal quantities*	
Fixed points	
Triple point of natural water T_{tp}	273.16 K
Zero Celsius (= $T_0 = t_{F,0}$)	273.15 K = 32 °F
Temperature conversions	
Kelvin to Rankine (T_R)	$T = (5/9)\ T_R$
Kelvin to Celsius	$t = T - T_0$
Kelvin to Fahrenheit	$t_F = (9/5)(T - T_0) + t_{F,0}$
Celsius to Fahrenheit	$t_F = (9/5)\ t + t_{F,0}$
[Numerically, $5(\{t_F\} + 40) = 9(\{t\} + 40)$, where $\{t\}$ and $\{t_F\}$ are the numerical values of the Celsius and Fahrenheit temperatures, respectively.]	
Temperature interval conversions	
1 degree centigrade	1 degree Celsius, denoted 1 °C
1 °C	1 K
1 °F	(1/1.8) K
1 °R	(1/1.8) K
Other thermal quantities	
1 Btu$_{th}$/h	0.292 875 W
1 Btu$_{IT}$/h	0.293 071th 1 W
1 cal$_{IT}$/s	4.186 8 W
1 cal$_{th}$/s	4.184 W
1 cal$_{IT}$/(g°C)	4186.8 J/(kg K)
1 Btu ft/(ft^2 h°F)	1.730 735 W m^{-1} K^{-1}
1 Btu in./(ft^2 s°F)	519.220 4 W m^{-1} K^{-1}
1 clo	0.155 m^2 K/kW
1 °F h ft^2/Btu	0.176 110 2 K m^2/W
1 °F h ft^2/Btu in	6.933 472 K m/W
1 Btu/lb°F ≡ 1 cal$_{ST}$/g°C	4186.8 J/kg.K
14. *Torque, moment of force*	
1 dyne·cm	10^{-7} N m
1 kgf·m	9.806 65 N m
1 ozf·in.	0.007 061 552 N m
1 lbf·in.	0.112 984 8 N m
1 lbf·ft	1.355 818 N m
15. *Velocity (includes speed)*	
1 foot per hour	8.466 667 × 10^{-5} m/s
1 foot per minute	0.005 08 m/s
1 knot (nautical mile per hour)	0.514 444 m/s
1 mile per hour (mi/h)	0.447 04 m/s

(*continued*)

TABLE A.5 (continued) Conversion Factors from English Measures to SI Units[a]

Unit	Equivalent
16. *Viscosity*	
1 poise	0.1 Pa s
1 ft^2/s	0.092 903 04 m^2/s
1 lb/(ft s)	1.488 164 Pa s
1 lb/(ft h)	$4.133\,789 \times 10^{-4}$ Pa s
1 lbf s/ft^2	47.880 26 Pa s
1 lbf-s/$in.^2$	6894.757 Pa s
1 rhe	10 $Pa^{-1}\ s^{-1}$
1 slug/ft s	47.880 26 Pa s
1 stokes, St	$1.0 \times 10^{-4}\ m^2/s$
17. *Volume (includes capacity)*	
1 stere, st	1 m^3
1 liter[g]	0.001 m^3
1 ft^3	0.028 316 8 m^3
1 $in.^3$	$1.638\,7 \times 10^{-5}\ m^3$
1 board foot	$2.359\,7 \times 10^{-3}\ m^3$
1 acre-foot	1233.48 m^3
1 dram (U.S. fluid)	$3.696\,7 \times 10^{-6}\ m^3$
1 gill (United States)	$1.182\,941 \times 10^{-4}\ m^3$
1 ounce (U.S. fluid)	$2.957\,353 \times 10^{-5}\ m^3$
1 teaspoon (tsp)[h]	$4.9288\,922 \times 10^{-6}\ m^3$
1 tablespoon (tbsp)	$1.4787\,676 \times 10^{-5}\ m^3$
1 pint (U.S. fluid)	$4.731\,765 \times 10^{-4}\ m^3$
1 quart (U.S. fluid)	$9.463\,529 \times 10^{-4}\ m^3$
1 gallon (U.S. liquid) [231 $in.^3$]	$3.785\,412 \times 10^{-3}\ m^3$
1 wine barrel (bbl) [31.5 gal (United States)]	0.119 240 m^3
1 barrel (petroleum, 42 gal, United States), bbl	0.158 987
1 ounce (U.K. fluid)	$2.841\,3 \times 10^{-5}\ m^3$
1 gill (Canada and United Kingdom)	$1.420\,6 \times 10^{-4}\ m^q$
1 gallon (Canada and United Kingdom)	$4.546\,09 \times 10^{-3}\ m^3$
	1.200 950 gal (United States)
1 pint (U.S. dry)	$5.506\,105 \times 10^{-4}\ m^3$
1 quart (U.S. dry)	$1.101\,221 \times 10^{-3}\ m^3$
1 gallon (U.S. dry)	$4.404\,884 \times 10^{-3}\ m^3$
1 peck	$8.809\,768 \times 10^{-3}\ m^3$
1 bushel (United States) [2150.42 $in.^3$]	$3.523\,907 \times 10^{-2}\ m^3$

[a] The conversion factor for a compound unit is usually not given here if it may easily be derived from simpler conversions; e.g., the conversion factors for "ft/s" to "m/s" or "ft/s^2" to "m/s^2" are not given, since they may be obtained from the conversion factor for "ft." Values are given to five or six significant digits except for exact values, which are usually indicated in bold type. A few former cgs measures are also included.

[b] The International Steam Table calorie of 1956.

[c] In practice, the prefix kilo is usually omitted. The kilogram calorie or large calorie is an obsolete term for the kilocalorie that is used to express the energy content of foods.

[d] Typographer's definition, 1886.

[e] Originally, in 1929, the international nautical mile.

[f] Based on 1 day = 86,400 s and 1 Julian century = 36,525 days.

[g] Post-1964 value, SI symbol l or L. Between 1901 and 1964, the liter was defined as 1.000th 028 dm^3.

[h] Although often given, it is doubtful whether normal usage justifies this accuracy. In Europe and elsewhere, teaspoon and tablespoon are usually exactly 5 and 15 mL, respectively.

Bibliography

Anderton P. and Bigg P. H., *Changing to the Metric System*, HMSO, London, U.K., 1980.

CIPM, Proces-Verbaux CIPM, 49th Session, 1960, pp. 71–72; *Comptes Rendues, 11th CGPM*, 1960, p. 85.

Cohen E. R., *The Physics Quick Reference Guide*, American Institute of Physics Press, New York, 1995.

Standard for Use of the International System of Units (*SI*): *The Modern Metric System, IEEE/ASTM SI 10–1997*, IEEE Standards Coordinating Committee 14 (Revision and redesignation of ANSI/IEEE Std 268–1992 and ASTM E380), IEEE, New York, 1997.

Taylor B. N., *Interpretation of the SI for the United States and Metric Conversion Policy for Federal Agencies*, Natl. Inst. Stand. Technol., Spec. Publ. 814, U.S. Government Printing Office, Washington, DC, 1991.

Taylor B. N., *Guide for the Use of the International System of Units*, 1995 ed., Natl. Inst. Stand. Technol., Spec. Publ. 811, U.S. Government Printing Office, Washington, DC, 1995.

The International System of Units, 7th edn., BIPM, Sèvres, France, 1998.

The International System of Units (*SI*), Natl. Inst. Stand. Technol., Spec. Publ. 330, 1991 ed., U.S. Government Printing Office, Washington, DC, 1991.

The NIST Guide for the Use of the International System of Units. http://www.nist.gov/pml/pubs/sp811/index.cfm (last accessed August 23, 2013).



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Index

A

Abbé error, 27-7
Absolute angular velocity sensor, 35-14–35-16
Absolute encoders
  direct binary, 30-2–30-3
  gray code, 30-4–30-5
  optical resolving, 30-6
  pseudorandom code, 30-4–30-7
  Vee-Scan, 30-3–30-4
Absorption dynamometers
  drag torque, 44-11–44-12
  eddy current dynamometer, 44-13
  friction brake, 44-11
  Froude brake, 44-12
  hydraulic dynamometers, 44-12–44-13
  prony brake, 44-11–44-12
  water brake, 44-12
Absorption-type densitometers, 45-16
Accelerometers, 35-8
  active-type accelerometer, 36-1–36-2
  application type, 36-28
  calibration methods, 36-3
  cantilever-type accelerometers, 36-22
  cross-axis sensitivity, 36-26–36-27
  deflection-type seismic accelerometer, 36-2
  differential-capacitance accelerometers, 36-14–36-16
  displacement and velocity sensors, 36-2
  dynamic, 36-2
  electromechanical force-balance accelerometers
    acceleration-sensitive mass, 36-8
    advantages, 36-8
    coil-and magnetic-type accelerometers, 36-8–36-11
    induction-type accelerometers, 36-11
  electrostatic force feedback accelerometers, 36-23–36-24
  environmental conditions, 36-29
  frequency range, 36-28
  full-scale range, 36-29
  kinematic, 36-2
  microaccelerometers, 36-24–36-26
  nonstationary random vibrations, 36-5–36-8
  overload capability, 36-29
  passive-type accelerometer, 36-1–36-2
  periodic vibrations, 36-3–36-5
  piezoelectric accelerometers, 36-11–36-13
  piezoresistive accelerometers, 36-14
  seismic accelerometers
    displacement, 36-19
    LVDT, 36-20
    Newton's second law of motion, 36-18
    potentiometer accelerometer, 36-20
    schematic diagram, 36-17
    strain-gage seismic instrument, 36-20–36-21
  signal conditioning, 36-29–36-32
  stationary random vibrations, 36-5
  strain-gage accelerometers, 36-17
  transient response, 36-28–36-29
  transients and shocks, 36-5
  vibrating string accelerometer, 36-21–36-22
Acoustic gas thermometer, 75-8, 75-10, 75-12
Acoustic measurement
  acoustic waves
    description, 48-1–48-2
    plane sound waves, 48-3
    spherical waves, 48-3–48-4
    wave equation, 48-2–48-3
  calibration
    electrostatic actuator method, 48-21
    pistonphone method, 48-20
    reciprocity calibration method, 48-20
    sound pressure calibrator, 48-20–48-21
  condenser microphone (*see* Condenser microphone)
  fluid viscosity measurement, 46-20–46-22
  frequency analyzers, 48-12
  instrumentation and manufacturers, 48-22
  near-field acoustic holography, 48-16–48-20
  pressure-based measurement, 48-12–48-14
  sound intensity measurement, 48-14–48-16
  sound pressure level measurement, 48-9–48-12

Acoustic microscopy, 49-2, 49-23
Acoustic plethysmography, 20-16
Active sensors, 2-3
AD522 amplifier, 50-12
Adhesives, 23-12
Adiabatic calorimeter, 79-5
Air gages, 28-9–28-11
Air/water plethysmography, 20-3–20-4, 20-15
Alphanumeric displays, 96-10
Alpha track dosimetry, 84-10
Altitude measurement
  ground-based height estimation (*see* Ground-based height estimation)
  height measurement, 15-2–15-3
  pressure altitude measurement, 15-1–15-2
  radar altimeter, 15-2
Aluminum capacitive humidity sensor, 25-11–25-12
Ambiguity interval, 14-7
American gage design (AGD), 28-4
American Society for Testing and Materials (ASTM), 74-3
A-mode display, 49-20
Amplitude-modulated (AM) proximity sensors
  calibration-per-surface strategy, 22-9
  configuration, 22-6, 22-9
  distance measurement, 22-8
  light intensities, 22-7–22-8
  symmetrical three-fiber constellation, 22-7
  Y-guide transducer, 22-6–22-7
Amplitude-shift keying (ASK), 85-8
Analog quadrature signal
  resistor network, 30-10–30-12
  sampling and numerical processing, 30-12–30-13
Analog sensors, 3-4
Analog-to-digital converter (ADC), 1-7, 23-6, 95-8
Angle measurement
  clinometer, 42-5
  definition, 42-1–42-2
  instruments and devices, 42-1, 42-3–42-4
  manufacturers and suppliers, 42-1–42-2
  optical comparator, 42-5
  protractor, 42-5–42-6
  sine bar, 42-5–42-6
  sine plate, 42-6–42-7
  taper, 42-7
Angle position indicator (API), 32-5–32-6
Angular acceleration, 44-2
Angular displacement, 44-1
Angular momentum flowmeters, 54-11
Angular velocity, 35-9–35-10, 44-2
Anisotropic stress effect, 41-10
Annihilation radiation, 80-9
Annular array transducer, 49-7
Anthropomorphic test device (ATD), 35-16
Application development environments (ADEs), 10-6–10-7
Application programming interface (API), 10-3
Application software
  ADEs, 10-6–10-7
  configuration-based programs, 10-7
  DAQ system, 10-5
  data analysis and visualization, 10-8–10-9
Area measurement
  digitizer, 19-4
    arm digitizer, 19-5–19-6, 19-9
    evaluation, 19-10–19-11
    manufacturers, 19-10
    sonic digitizer, 19-5
    tablet digitizer, 19-5, 19-9
  grid overlay, 19-6
  planimeter
    basic wheel and bar mechanism, 19-2
    battery-operated planimeters, 19-8
    electronic planimeters, 19-7–19-8
    evaluation, 19-10–19-11
    manufacturers, 19-10
    measuring wheel, 19-1–19-2
    mechanical planimeters, 19-7
    operation, 19-3
    polar planimeter, 19-1–19-2
Arm digitizer, 19-5–19-6, 19-9
Atomic force microscope (AFM), 13-9
Attitude measurement
  aircraft
    directional gyro analysis, 16-13–16-14
    IMU, 16-14
    vertical gyro analysis, 16-11–16-13
  crane lifters
    dynamic equation, 16-9–16-10
    high-resolution camera, 16-8
    illustrative diagram, 16-8–16-9
    Lagrange's equations of motion, 16-9
    on-line attitude measurement, 16-10
    sensor output, 16-10
    swing signals, 16-8–16-9
  gyroscopes, 16-1–16-2
  panoramic annular lens, 16-17–16-18
  servo-type accelerometer, 16-2–16-3
  servo-type inclinometers, 16-2–16-3
  ships
    heaving, rolling, and pitching signals, 16-6–16-7
    on-line attitude measurement, 16-7–16-8
    sensor location, 16-5–16-6
  spacecraft applications, 16-3–16-5, 16-15–16-17
Automotive Industry Action Group (AIAG), 28-14
Autonomous sensors, 1-10
  block diagram, 88-1–88-2
  components, 88-2
  energy harvesting, 88-2–88-3
  energy sources
    AC magnetic energy, 88-5
    biochemical energy, 88-5
    mechanical energy, 88-4–88-5
    radiant energy, 88-4

remote powering, 88-4
thermal energy, 88-5
hybrid storage, 88-9
primary batteries, 88-2
secondary batteries, 88-7–88-8
supercapacitors, 88-8–88-9
zero-order electric model, 88-6
Autonomous underwater vehicle (AUV), 17-12
Axial turbine flowmeters
aerodynamic lift, 54-3
applications, 54-1
calibration, 54-6
design and construction, 54-7–54-8
dual-rotor axial, 54-9–54-10
flat-bladed rotor, 54-3–54-4
fluid driving torque, 54-3–54-5
helical turbine meters, 54-8–54-9
insertion flowmeter, 54-10–54-11
installation, 54-7
maintenance, 54-7
momentum exchange approach, 54-3
performance characteristics, 54-2–54-3
propeller, 54-8
schematic longitudinal section, 54-1–54-2
swirl-type axial, 54-9

B

Balanced-flow vessel, 45-6
Balance-type densitometers, 45-6
Bandgap reference, 71-4
Battery-operated planimeters, 19-8
Bayard–Alpert gage, 40-12–40-14
Beam-type load cells, 41-6–41-7
Beer's law, 72-5
Bending temperature, 76-5
Bernoulli's equation, 62-1–62-2
Bethe–Bloch equation, 82-2
Bicell detector, 34-4–34-5
Bingham fluid, 46-4
Binomial theorem, 62-8
Birefringent film strain sensor, 37-14–37-15
Birotor type positive displacement meter, 53-3–53-4
B-mode display, 49-20
B-mode imaging, 49-4, 49-20, 49-23
Bore gaging
accuracy of, 28-2
classification, 28-2–28-3
form-measuring instruments, 28-11–28-13
repeatability and reproducibility (R&R) methodology
application, 28-15–28-16
probability density function, 28-13–28-15
tolerance, 28-2
types of
air gages, 28-9–28-11
Bowers metrology UK, 28-6, 28-8
cylinder bore gage, 28-5–28-6
fixed-limit, 28-3–28-5
iWave, 28-7–28-8
micrometer, 28-6–28-7
multipoint contact, 28-9
pistol-grip, 28-8–28-9
self-centering two-point bore gage, 28-6
TESA Veribor, 28-6–28-7
Borehole inclinometers, 21-4
Bore tolerance, 28-2
Bourdon gage, 40-4–40-5
Bourdon tube, 39-2–39-3
Bowers metrology, 28-6, 28-8
Bremsstrahlung
beam transmission equation, 80-10
mean range, 80-11–80-13
narrow-beam attenuation coefficient, 80-9–80-10
production, 80-9
stopping power, 80-11–80-12
Bubble dosimetry, 84-11
Built-in-test (BIT) flag, 32-11
Bullers ring, 76-5–76-8
Buoyancy-type densitometers, 45-4
Bushing mount, 23-8–23-9

C

Calibration, 1-5–1-6, 7-7
accelerometers, 36-3
acoustic measurement
electrostatic actuator method, 48-21
pistonphone method, 48-20
reciprocity calibration method, 48-20
sound pressure calibrator, 48-20–48-21
advantages, 8-3
axial turbine flowmeters, 54-6
calorimetry, 79-9–79-10
chemical mixtures, 8-10
cost of, 8-6–8-7
DAQ device driver software, 10-4
definition, 4-18, 8-1
dynamic calibration, 8-4
dynamic pressure, 8-11
electromagnetic flowmeters, 56-8
electronic calibration, 8-7–8-8
empirical temperature scale, 65-6
errors and uncertainties, 8-2–8-3
flow-calibration, 8-10, 59-11–59-13, 63-12–63-14
flowmeters, 4-18–4-20
heat flux, 67-12–67-13
impeller flowmeters, 55-4
important element, 8-2
instrumentation, 4-20
laboratories and institutions, 7-7, 8-4–8-5
procedure and personnel, 8-3–8-4
remote and e-calibration, 8-8–8-9
RTD, 68-17, 68-19

self-calibrations, 4-21, 8-8
sensor calibration, 8-10
soft calibrations, 8-8
software support, 8-5–8-6
S-shaped force sensor, 12-2–12-3
standards
classification, 43-12
design, 43-13
handling, storage, packing, and shipping, 43-13
materials, 43-13
maximum permissible error (MPE), 43-13
static calibration, 8-4
static pressure, 8-10
thermistor thermometers, 69-4–69-5
thermocouple thermometry, 70-31–70-34
turbine and vane flowmeters, 54-6
user calibration, 8-2
VNA, 8-9
Calibration management software (CMS), 8-6
Callendar–Van Dusen equations, 68-3
Calorimeter methods, 67-11
Calorimetric sensor, *see* Capillary-type thermal mass flow meter
Calorimetry
applications, 79-10
calibration, 79-9–79-10
calorimetric bomb/solution calorimeter, 79-13
classification
adiabatic calorimeter, 79-5
chip calorimeter, 79-3, 79-5–79-6
essential measures, 79-3–79-4
heat-flux, 79-5
isoperibol, 79-5
isothermal, 79-4
phase-transition compensation, 79-4
instrument control and data management, 79-14
scanning
DSC, 79-7–79-9
heat-conducting disk, 79-6–79-7
measurement curve, 79-9
signals, 79-3
thermal analysis of materials, 79-10–79-12
thermal energy, heat, and temperature, 79-2
vendors and specifications, 79-14–79-15
work and enthalpy, 79-2–79-3
Canon laser rotary encoder, 30-17
Cantilever-type accelerometers, 36-22
Capacitance-type level device, 18-13–18-14
Capacitive accelerometers, 25-8–25-9
Capacitive force transducer, 41-12
Capacitive gaging, 13-5
Capacitive humidity sensor
aluminum, 25-11–25-12
moisture, 25-13
polymer, 25-12–25-13
silicon, 25-12
tantalum, 25-12
Capacitive moisture sensor, 25-13
Capacitive pressure sensor, 25-7–25-8, 39-3–39-5
Capacitive sensors, 22-13–22-14
accelerometer and force transducer, 25-8–25-9
humidity sensor
aluminum, 25-11–25-12
polymer, 25-12–25-13
silicon, 25-12
tantalum, 25-12
liquid-level measurement, 25-9–25-10
moisture, 25-13
pressure, 25-7–25-8, 39-3–39-5
principles of, 25-2
signal processing
feedback, 25-17
operational amplifiers, 25-14–25-15
oscillator circuits, 25-17
pulse width modulation, 25-15–25-16
square-wave linearization, 25-15–25-16
variable displacement
area, 25-3
dielectric, 25-3–25-4
differential capacitive sensor, 25-4–25-5
distance, 25-2–25-3
smart position sensor, 25-5–25-7
Capacitive tactile sensor, 38-6–38-8
Capacitor displacement sensor
area, 25-3
dielectric, 25-3–25-4
differential capacitive sensor, 25-4–25-5
distance, 25-2–25-3
smart position sensor, 25-5–25-7
Capillary-type thermal mass flow meter
accuracy, 50-12–50-13
advantages, 50-1
electronic readout
analog-to-digital converter and display, 50-12
block diagram, 50-9
current source, 50-9–50-10
functions, 50-8
instrumentation amplifier, 50-12
power supply, 50-9
sensor assembly, 50-10–50-11
fluid flow, 50-2–50-4
fluid velocity *vs.* sensed temperature, 50-3
heat transfer, 50-2
incompressible fluids, 50-6
laminar flow element, 50-4, 50-5–50-7
principle of, 50-2
requirement, 50-1
resistance temperature detectors, 50-4
schematics, 50-3
temperature profile, 50-5, 50-7–50-8
Capillary viscometers, 46-11–46-12
Capillary waves, 47-5
Carbon composition elements, 23-3
Cathode ionization gages

cold cathode, 40-14–40-16
hot cathode
Bayard-Alpert gage, 40-12–40-14
triode, 40-11–40-12
Cathode-ray tube displays (CRT), 96-11–96-12
Ceramics, 26-4–26-7
Cermet element, 23-3
Certified reference material (CRM), 7-8
Chain-balanced float, 45-6–45-7
Chain code, 30-4–30-5
Chamber plethysmography, 20-3–20-4, 20-15
Charged-particle measurement
energy loss
electrons, 82-3–82-4
protons and ions, 82-2–82-3
gas-filled detectors (*see* Gas-filled detectors)
nuclear emulsions
advantage, 82-4
image tracking, 82-6
pellicles, 82-5
shrinkage, 82-6
thickness, 82-6
scintillation counters
components, 82-5
construction, 82-13–82-14
description, 82-12
light detection and readout, 82-13
light guides/Winston cone, 82-12–82-13
manufacturers lists, 82-15
operating information, 82-14
operation principle, 82-11
plastic scintillation counter hodoscope, 82-14
scintillation medium, 82-12
thin finger-shaped counters, 82-14
VLPC, 82-14
solid-state detectors
advantages, 82-5
categories, 82-16–82-17
front-end electronics, 82-20
manufacturers and prices, 82-17, 82-20
modern silicon processing technologies, 82-20
multichannel analyzer (MCA), 82-17–82-18
preamplifiers, 82-19
pulse height deficit, 82-19–82-20
signal generation, 82-15–82-16
silicon strip detector, 82-18–82-19
thin transmission-type (D*E*) detector, 82-18
Check reading indicators, 96-10
Chip calorimeter, 79-3, 79-5–79-6
Chopping current amplifiers, 95-2–95-3
Clinometer, 42-5
Closed conduits flow measurements
acoustic flow rate Q_{ADM}, 57-5
flow profile, 57-5
multipath circular cross sections, 57-8–57-9
single path circular cross sections, 57-6–57-8
single path rectangular cross sections, 57-6–57-8
sonic path arrangements, 57-6
Closed-loop servo model, 32-8–32-9
C-mode display, 49-20
Coaxial detectors
closed-end coaxial geometry, 81-13
^{60}Co spectrum, 81-14–81-15
Czochralski technique, 81-14
Fano factor, 81-15
relative absorption efficiencies, 81-14
Coil-and magnetic-type accelerometers
Ampere's law, 36-8
magnetic structure, 36-9
permanent magnet accelerometer, 36-8–36-9
rotational-type servo accelerometer, 36-9–36-11
Cold cathode ionization gages, 40-14–40-16
Color-change temperature indicator
characteristics, 76-12
crayons, 76-13
irreversible paints, 76-12–76-13
liquid crystals, 76-14–76-16
operating principle, 76-11
reversible paints, 76-14
Colossal magnetoresistance (CMR), 31-9
Column-type densitometers, 45-7–45-9
Combinatorial multiplexing switch (CoMux), 70-36–70-37
Commercial thermal anemometers, 59-2–59-4
Comparators, 13-5
Compliant couplings, 23-9, 23-11
Computer tomography (CT) scanning, 20-2
CoMux, *see* Combinatorial multiplexing switch (CoMux)
Concerning capacitive sensors, 13-7
Condenser microphone
basic elements, 48-5–48-6
free-field microphone, 48-8–48-9
frequency response, 48-6–48-7
mechanical analogy, 48-6
pressure microphone, 48-7–48-8
size, 48-6
Conduction, 67-2–67-3
Conductive plastic element, 23-3
Cone-and-plate viscometers, 46-9–46-10
Conformity, 23-4
Constant-bandwidth modulation (CBW), 90-9–90-10
Constant-temperature anemometer, 59-8–59-10
Constant-volume gas thermometer (CVGT), 75-5, 75-7–75-8, 75-11–75-12
Contact angle, 47-2
Contact line, 47-2
Contact sensing, 38-1
Contact thermometers, 65-14–65-16
Continuous emission monitoring system (CEMS), 62-13
Continuous position sensor, 34-6

Convection, 67-3
coefficient, 63-8
gage, 40-9–40-11
Cooled optically stimulated luminescence (COSL), 84-7
Coordinate-measuring machines (CMMs), 27-14
Coriolis densitometers, 45-15–45-16
Coriolis effect, 60-4
Coriolis flowmeters, 51-1
advantages, 60-1, 60-4–60-5
applications, 60-5
construction, 60-4
disadvantages, 60-5
theory of operation
central point, 60-2
Coriolis effect, 60-4
degree of tube, 60-3
flow tube responses, 60-1–60-3
geometric shapes, 60-1
phase shift, inlet and outlet side, 60-3
Correlation method, 29-5–29-6
Couette-type viscometer, 46-7
CVGT, *see* Constant-volume gas thermometer (CVGT)
Cyclic redundancy check (CRC), 11-3
Cylinder bore gage, 28-5–28-6
Cylindrical rotor flowmeters, 54-12

D

Data acquisition (DAQ) software, 10-2–10-3
Data age, 27-7
Data analysis, 1-2
DC brush and brushless motors
ADC process, 95-8
closed-loop system, 95-7
feedback control algorithm, 95-6
frequency response, 95-7
frequency-to-voltage converter, 95-7
instrument-integrated closed-loop system, 95-7–95-8
microprocessor-controlled closed-loop system, 95-8–95-9
position measurement, 95-6
PWM power amplifier, 95-5–95-6
rotational speed measurement, 95-6
voltage-to-frequency converter, 95-7
Dead-volume error, 75-5, 75-7
Defense Advanced Research Projects Agency (DARPA), 89-1, 89-3
Deflection-type seismic accelerometer, 36-2
Delta modulation (DM), 90-13
Density measurement
fluid density
absorption-type densitometers, 45-16
balanced-flow vessel, 45-6
balance-type densitometers, 45-6
buoyancy gas balance, 45-7–45-8
buoyancy-type densitometers, 45-4
chain-balanced float, 45-6–45-7
column-type densitometers, 45-7–45-9
Coriolis densitometers, 45-15–45-16
gas specific gravity balance, 45-7
hydrometers, 45-4–45-5
hydrostatic weighing densitometers, 45-5–45-6
magnetic drives, 45-10
piezoelectric drives, 45-10
pycnometric densitometers, 45-3–45-4
radioactive densitometers, 45-13–45-14
refraction densitometers, 45-14–45-15
static and dynamic measurements, 45-2
tuning fork densitometers, 45-12–45-13
vibrating cylinder densitometers, 45-11–45-12
vibrating element densitometers, 45-9–45-10
vibrating tube densitometers, 45-10–45-11
solid density, 45-2
specific gravity, 45-2
Device driver software
advanced driver capabilities, 10-4
API, 10-3
automatic data logging to disk, 10-5
calibration, 10-4
configuration, 10-3–10-4
continuous buffered acquisition, 10-4
data buffering and management, 10-4
finite buffered acquisition, 10-4
hardware simulation, 10-5
intelligent timing, triggering, and synchronization, 10-5
Device under test (DUT), 44-10, 71-12
Dial indicators, 13-3, 13-5
Diaphragm gage, 40-5–40-6
Diaphragm positive displacement meter, 53-6–53-7
Differential-capacitance accelerometers, 36-14–36-16
Differential pressure flowmeters
advantages, 51-7
Coriolis, 51-1
Dall tube, 51-7
disadvantage, 51-2
elbow, 51-7
fluid flow in pipes
Bernoulli's equation, 51-3–51-4
fluid velocity, 51-2
laminar flow, 51-2–51-3
Reynolds number, 51-2
transitional flow, 51-2
turbulent flow, 51-2–51-3
installation, 51-8–51-9
multiphase flow metering, 51-11
operation, 51-1
performance characteristics and applications, 51-7–51-8
producer, 51-2
segmental wedge, 51-7
square-edged orifice plate (*see* Orifice plate)
standards, 51-10–51-11

suppliers, 51-8–51-9
transducer, 51-2, 51-9–51-10
V-cone, 51-7
venturi tube, 51-6–51-7
Differential pressure transmitter, 51-11
Differential pulse code modulation (DPCM), 90-13
Differential scanning calorimeter (DSC)
applications, 79-10
calorimetric bomb/solution calorimeter, 79-13
Flash DSC 1, 79-6
FS-DSC, 79-7, 79-9
inaccuracy, 79-9
instrument control and data management, 79-14
principal types, 79-6–79-7
thermal analysis of materials, 79-11
TM-DSC, 79-7–79-8
Differential thermocouple thermometer, 70-8–70-9
Diffraction-based encoder, 30-16–30-17
Digital controllers, 91-7
Digital quadrature signal
commercial chips and suppliers, 30-10, 30-21
count signal, 30-8–30-9
divide-by-four counter circuit, 30-10, 30-21
flip-flops, 30-8
rotary incremental encoder, 30-7–30-8
Digital sensors, 3-4–3-5
Digital signal processing, 64-10–64-11
Digitizer, 19-4
arm digitizer, 19-5–19-6, 19-9
evaluation, 19-10–19-11
manufacturers, 19-10
sonic digitizer, 19-5
tablet digitizer, 19-5, 19-9
Dilatancy, 46-4
Directional gyro (DG) analysis, 16-13–16-14
Directivity, 49-17
Direct reading gages
liquid wall, 40-2–40-6
solid wall, 40-3–40-6
Disappearing filament pyrometer, 72-14–72-16
Discrimination, 4-3–4-4
Displacement sensor, 26-10–26-11, 36-2
Display system
alphanumeric displays, 96-10
aspect ratios, 96-6–96-7
contrast, 96-4–96-5
CRT, 96-11–96-12
electroluminescent displays, 96-14–96-15
electronic paper/electronic ink, 96-15–96-16
flicker, 96-8–96-9
gamut and color spaces, 96-8–96-9
gray scale, 96-5
LCD
active-matrix LCD, 96-12
advantages, 96-12
IPS, 96-13
passive-matrix LCD, 96-12
performance, 96-12–96-13
twisted-nematic (TN) effect, 96-13
VA LCD, 96-13
LED displays, 96-15
light measurement, 96-2
pixels, 96-5
plasma displays, 96-13–96-14
quantitative and status displays, 96-10
refresh rate, 96-9–96-10
resolution, 96-5–96-6
sharpness and legibility, 96-6–96-7
vision, 96-2–96-3
visual performance, 96-3–96-4
Distance measurement
active/passive techniques, 14-2–14-3
coherent/noncoherent detection, 14-7
contact/noncontact techniques, 14-2
field-based approaches, 14-5
GPS, 14-7
laser-based active triangulation (*see* Laser-based active triangulation)
LIDAR, 14-7
AM LIDAR, 14-12–14-13
resonating LIDAR, 14-13
performance characteristics
depth of field, 14-8
lateral resolution, 14-10
maximum range, 14-9–14-10
range accuracy, 14-8
rate of acquisition, 14-10
position tracking
active target approaches, 14-14
devices, 14-7, 14-15
magnetic position tracking, 14-15–14-16
OPTOTRAK position-tracking system, 14-14–14-15
radar, 14-6
range imaging devices, 14-7–14-8, 14-15
stationary magnetic fields, 14-6
TOF systems, 14-3, 14-5
triangulation techniques, 14-4–14-5
ultrasonic ranging techniques, 14-6
Distributed impedance sensor, 31-7–31-8
Distributed/quasi-distributed optical fiber temperature sensor
Bragg gratings, 77-9–77-10
Brillouin scatter, 77-8–77-9
core/cladding indices, 77-8
heat and cold alarm system, 77-8
Kerr effect, 77-9
mixed pump and probe signal, 77-9
principle, 77-7
Raman distributed temperature probe, 77-7
simple step index fiber, 77-8
spatial resolutions, 77-7
Doppler shift frequency, 12-8
Dot mote, 86-4–86-5

Draft international standard (DIS), 6-6
Drag force flowmeters
advantages, 61-1
drag element, 61-1
flow sensor design, 61-1–61-2
fluid flow measurement principle
drag coefficient *vs.* Reynolds number, 61-3
drag equation, 61-2
strain, 61-4
theory of elasticity, 61-3
velocity direction, 61-5
velocity magnitude, 61-4–61-5
optical fiber strain gage
intensity variation, 61-7
optical root strain magnitude, 61-7
optical strain direction, 61-8
principle, 61-7
with rubber beam, 61-8
wind velocity, 61-7
without rubber beam, 61-8–61-9
resistive strain gages, 61-5–61-7
Driver software, 10-3
Driving and universal dynamometers, 44-14
Drop weight/volume method, 47-10–47-11
Dry-sensor microprocessor-based system, 63-3, 63-11–63-12
DSC, *see* Differential scanning calorimeter (DSC)
Dual-beam system
components of, 64-2–64-3
Doppler shift, 64-4–64-6
fringe model, 64-2–64-3
Dual-reference junction circuit, 70-5–70-6, 70-13–70-15
Dual-rotor axial turbines, 54-9–54-10
Dual-synchronized autonomous monitoring system, 15-8
du Noüy ring method, 47-8
Dynamic characteristics
characteristic equation development, 4-13–4-14
dynamic states, 4-11–4-12
first-order block
amplitude and phase shift, 4-16
Laplace transformation, 4-16
step response, 4-15–4-16
time-dependent term, 4-14
time domain, 4-15
forcing functions, 4-12–4-13
linear systems, 4-14
second-order block
characteristic function, 4-17
critical damping, 4-17
natural frequency, 4-18–4-19
resonance, 4-18
step input response, 4-17–4-18
system performance, 4-17
zero-order block, 4-14–4-15
Dynamic light scattering, 46-24–46-25

E

Eddy current dynamometer, 44-13
Eddy-current sensing, 13-5–13-6
Eddy current transducer, 24-17–24-19
Eddy viscosity, 46-2
EHA, *see* Electrohydrostatic actuator (EHA)
Electrohydraulic actuator
block diagram, 93-3
EHA
maintainability and survivability, 93-4
motor model, 93-5
piston and flight control surface (*see* Piston and flight control surface)
pump and fluid, 93-5–93-6
electronic "fly-by-wire" mechanization, 93-4
feedback control system, 93-3
first-and fourth-order actuator models, 93-3–93-4
first-order transfer function, 93-3
irreversible control system, 93-2
modeling requirements, 93-2
QFT compensator, 93-9–93-11
time response, 93-3–93-4
Electrohydrostatic actuator (EHA)
maintainability and survivability, 93-4
motor model, 93-5
piston and flight control surface
aircraft short-period approximation, 93-9
flexible hinge joint model, 93-6–93-7
hinge moments, 93-8
load components, 93-6
stiff hinge joint model, 93-7–93-8
pump and fluid, 93-5–93-6
Electroluminescent displays, 96-14–96-15
Electromagnetic emission radioactivity
conversion electrons, 80-8
gamma rays, 80-7–80-8
x-rays, 80-7
Electromagnetic flowmeters
accuracy, 56-1
AC-type, 56-5–56-6
applications, 56-8–56-9
calibration, 56-8
conductor, 56-2
construction, 56-2–56-5
DC-type, 56-6–56-7
Faraday's law of induction, 56-1–56-2
fluid, electric conductivity, 56-9
induced voltage, 56-2–56-3
installations, 56-7–56-8
magnetohydrodynamic effect, 56-9
noise sources, 56-3
operation, 56-5
signal pickup and demodulation techniques, 56-10–56-11
volume of liquid flow, 56-2–56-3

Electromagnetic interference shielding, 24-20
Electromagnetic transducer, 29-4
Electromechanical force-balance accelerometers
  acceleration-sensitive mass, 36-8
  advantages, 36-8
  coil-and magnetic-type accelerometers, 36-8–36-11
  induction-type accelerometers, 36-11
Electron-capture (EC) process, 80-7
Electronic controllers, 91-6–91-7
Electronic gages, 13-5
Electronic personal dosimeter (EPD), 84-12
Electronic planimeters, 19-7–19-8
Electronics architecture, 78-9–78-10
Electro-optical sensors, 22-3
Electrostatic force feedback accelerometers, 36-23–36-24
Electrostatic gyro (ESG), 17-2–17-3
Electrostatic transducer, 29-4
Elementary thermocouple thermometry, 70-2
Ellipsometry, 13-7
Embedded vehicle area network (EVAN), 87-8
  advantages, 87-10
  "always-on" mode, 87-8
  design, 87-9–87-10
  electric system behavior tracking, 87-8
  microcontroller, 87-8
  peripherals, 87-10
  physical and electric footprint advantages, 87-11
  power consumption, 87-8
  power supply design considerations, 87-10–87-11
  system-wide implementation, 87-3
Emissivity, 72-4–72-6, 72-8, 72-15, 78-2, 78-3
Empirical temperature scale
  calibration, 65-6
  fixed point, 65-8
  interpolating instrument, 65-8
  nonuniqueness, 65-8
  semi-empirical scales, 65-6–65-7
  smoothness, 65-9
Epic mote, 86-11
Equal-arm balance scale, 3-1–3-2
Error of nonlinearity, 4-10–4-11
Error reduction, 1-5, 2-7–2-8
Error/uncertainty propagation, 5-5
European railway traffic management system (ERTMS)
  Eurobalises, 11-10
  high-level architecture, 11-10
  system requirement specifications, 11-9
  telegram, 11-10
  trackside subsystem, 11-10–11-11
EVAN, *see* Embedded vehicle area network (EVAN)
Expanded uncertainty
  degrees of freedom, 5-11
  engineering system, 5-10
  ISO, 5-10
  round-off errors, 5-12
  temperature measurement, 5-11
Explosion protection
  approval and certification processes, 94-11–94-13
  electric equipment
    enclosure types and requirements, 94-5–94-6
    flammable limits, 94-2
    flash point, 94-2
    ignition temperature, 94-2
    maximum surface temperature, 94-2–94-3
    vapor density, 94-2
  enclosures, 94-14
  exothermic reactions, 94-2
  hazardous area classification, 94-3–94-5
  ignition curves
    ANSI/UL 913 standard, 94-9
    capacitance circuit, 94-9–94-10
    inductance circuit, 94-9–94-10
    L/R ratio method, 94-11
    lumped parameters method, 94-11
    resistance circuit, 94-9–94-10
  intrinsic safety (*see* Intrinsic safety)
  lighting panelboard, 94-14
  purging and pressurization, 94-15–94-16
Extrinsic Fabry–Perot interferometric (EFPI) sensor, 33-2–33-3

F

Fabry–Pérot interferometer, 27-12–27-13
Fail-safe electronics
  Level 1 ERTMS
    Eurobalises, 11-10
    high-level architecture, 11-10
    system requirement specifications, 11-9
    telegram, 11-10
    trackside subsystem, 11-10–11-11
  realization
    auxiliary clock and timer block, 11-6
    central unit, 11-4
    communication security, 11-3
    computer structure, 11-3–11-4
    documentation, 11-2
    error minimization, 11-2
    I/O and safety board, 11-4–11-5
    I/O ports, 11-5–11-6
    programmable circuits, 11-3
    reactive, inherent and composite safety, 11-2
    signal generator, 11-5–11-7
    technical analysis, 11-2
    voltage generator, 11-7
  reconfigurability, 11-9
  relays, 11-8–11-9
  safety performance, 11-1
  tests and their timing, 11-7–11-8
Falling cylinder viscometer, 46-14–46-15
Falling sphere viscometer, 46-13–46-14
Faraday's law of induction, 56-1–56-2
Fast neutron track dosimetry, 84-10–84-11

Fast-scan differential scanning calorimeter (FS-DSC), 79-7, 79-9
Fiber Bragg-Grating (FBG) sensor
  fabrication techniques, 33-6–33-8
  limitation of, 33-8–33-9
  operating principle, 33-5–33-6
  wavelength, 33-8
Fiber optic/laser, 35-16
Fiber-optic strain gages, 37-14
Fiber-optic thermometers
  linear and nonlinear optical process, 77-1–77-2
  point and multiplexed systems, 77-3
  principal benefits, 77-3
  sensor system outputs, 77-2–77-3
  temperature sensors (*see* Temperature sensor)
Field-mountable transducers, 57-4
Field propagation velocity, 31-4
Film dosimetry, 84-8–84-9
Finite element analysis (FEA), 38-2–38-3, 38-14
Fixed-limit gages, 28-3–28-5
Flexible hinge joint model, 93-6–93-7
Floating element flowmeter, *see* Rotameter
Fluid viscosity measurement
  acoustic method, 46-20–46-22
  cone-and-plate viscometers, 46-9–46-10
  falling body method
    falling cylinder viscometer, 46-14–46-15
    falling sphere viscometer, 46-13–46-14
    in opaque liquid, 46-16
    rising bubble/droplet, 46-16–46-17
  high-pressure rheology
    application, 46-25–46-26
    high-pressure rheometry, 46-26–46-27
  microrheology
    Brownian motion, 46-22
    diffusion constant, 46-23
    dynamic light scattering, 46-24–46-25
    particle-tracking microrheology, 46-23–46-24
  Newtonian and non-Newtonian fluids, 46-3–46-5
  oscillation method
    oscillating cup method, 46-17–46-20
    vibrating-wire viscometer, 46-20
  parallel disks, 46-10
  pressure and gravity flow methods
    capillary viscometers, 46-11–46-12
    glass capillary viscometer, 46-12
    orifice/cup, short capillary viscometer, 46-12
  rheometer
    classification, 46-6
    description, 46-5
    errors, 46-6
    manufacturers/distributors, 46-7
  rotational viscometers/concentric cyclinders
    advantage, 46-5
    apparent fluid viscosity, 46-9
    Couette-type viscometer, 46-7
    fluid shear stress, 46-7
    representative radius, 46-8
    Reynolds number, 46-9
    Searle-type viscometer, 46-7
    shear rate, 46-7
  shear viscosity, 46-1–46-3
Flywheel gyroscopes, 17-5–17-7
Force measurement
  capacitive force transducer, 41-12
  end effector, 41-3
  FSR, 41-13
  Hooke's law, 41-2–41-3
  magnetoelastic force sensor, 41-14
  piezoelectric methods
    crystals, 41-9
    electrorestrictive elements, 41-7
    inductive, 41-9
    magnitude and polarity, 41-7–41-8
    materials, 41-8
    modes of operation, 41-8
    multicomponent dynamometers, quartz crystals, 41-10–41-12
    piezotransistor, 41-10
    resistive, 41-9
    voltage output, 41-8–41-9
    voltage sensitivity, 41-8
  strain-gage load cell
    beam cell, 41-6–41-7
    gage resistivity, 41-4–41-6
    ring cell, 41-7
  tactile sensor
    array sensor, 41-15–41-16
    requirements, 41-14–41-15
  torsional balances, 41-14
  types of, 41-4
Force sensing, 38-1
Force-sensing resistor (FSR), 41-13
Force sensors
  disadvantage, 12-1
  dynamic characteristics, 12-1–12-2
  impulse response
    Dirac delta function, 12-7
    experimental setup, 12-7–12-9
    FWHM, pulse duration, 12-10
    inertial force, 12-9
    instantaneous value, impact force, 12-9–12-10
    and momentum change, 12-10
    velocity and force measurement, 12-9
  LMM, 12-2
  oscillation response
    damping, 12-13
    experimental setup, 12-10–12-11
    $F_{inertial}$ and F_{sensor}, 12-12–12-13
    frequency, velocity, acceleration and force measurement, 12-11–12-12
    measurement, interferometer, 12-12–12-13
    validity and reproducibility, 12-14

S-shaped force sensor
corner cube (CC) prism, 12-3
displacement, 12-2
experimental setup, calibration, 12-2–12-3
force measurement, 12-4–12-6
linear bearing force, 12-4
rms value of error, 12-6
second-order derivative, 12-6–12-7
strain-gage sensor, 12-2–12-3
velocity, 12-4
Force transducer, 25-8–25-9
Forcing functions, 4-12–4-13
Fourier transform infrared analysis (FTIR), 79-11
Foursquare dynamometers, 44-14
Fraunhofer diffraction, 49-25
Free spectral range (FSR), 27-3
Frequency analyzers, 48-12
Frequency-based detection, 14-13
Frequency division multiplexing (FDM)
cross talk, 90-6
frequency and time graphs, 90-2–90-3
frequency bands, 90-5–90-6
measurement channel, 90-5
subcarrier modulation scheme
amplitude modulation, 90-6–90-7
CBW, 90-9–90-10
frequency modulation, 90-7–90-8
PBW, 90-8–90-9
Frequency-modulated continuous-wave (FMCW) radar, 18-11–18-12
Frequency-modulation method, 29-5
Frequency response, 4-2–4-3
Frequency-shift keying (FSK), 85-8
Frequency-to-voltage converter, 95-7
Fresnel diffraction, 49-24

G

Gage accuracy, 28-2
Gage factor, 39-5–39-7
Gamma-ray backscatter, 13-7–13-8
Gas dilution, 20-13
Gaseous detectors
charge transport, 81-2–81-3
electric field effects, 81-3
Geiger–Müller counters, 81-6–81-7
ionization chambers, 81-4
operating principles, 81-2
proportional counters, 81-5–81-6
scintillators (*see* Scintillators)
versatility, 81-2
W values, 81-2
Gas-filled detectors
advantage, 82-5
Geiger–Müller counter, 82-7, 82-10–82-11
ionization counter/chamber, 82-7
operating regions, 82-6
proportional counter
advantages, 82-7
applications, 82-11
design and operating principle, 82-7–82-8
drift chamber, 82-10
fill gases, 82-8–82-9
multiwire proportional chamber, 82-10
operation and readout, 82-9–82-10
signal amplitude, 82-7
time projection chamber, 82-10
Gas temperature sensor, 63-4–63-5
Gas thermometer
acoustics, 75-8, 75-10, 75-12
CVGT, 75-5, 75-7–75-8, 75-11–75-12
Gaussian distribution, 2-6
Geiger–Müller counters, 81-6–81-7, 82-7
Geodetic/differential leveling, 21-4–21-5
Geodetic/geophysical tiltmeters, 21-2
Geometrical product specification (GPS), 28-2
Giant magnetoresistance (GMR), 31-9
Glass capillary viscometer, 46-12
Global positioning systems (GPS), 14-7
Gradiometer sensor, 24-19
Graybody, 72-7
Gray code, 30-4–30-5, 31-10–31-11
Grid overlay, 19-6
Grindo-sonics, 49-22
Ground-based height estimation
dual-synchronized autonomous monitoring system, 15-8
inherent bias, 15-13
NAMS, 15-3–15-4
onboard-derived height estimation, 15-12
position fix, 15-14
quadrilateration, 15-8–15-10
single air traffic control radar, 15-5–15-6
SSR
omnidirectional radar, 15-6–15-7
pseudoranges, 15-10–15-11
RDHMU, 15-11–15-12
Gyroscopes
differenced/integrated accelerometer, 35-16
directional gyro analysis, 16-13–16-14
electrostatic gyro, 17-2–17-3
fiber optic/laser, 35-16
fluidic, 35-14
flywheel gyroscopes, 17-5–17-7
GyroChip™, 35-14–35-15
laser gyro, 17-2
mechanical gyroscopes, 17-4–17-5
MHD effect, 35-15–35-16
optical gyroscopes, 17-7–17-8
piezoelectric vibration, 35-14–35-15
spinning mass, 35-14
types, 16-1–16-2
vertical gyro analysis, 16-11–16-13

H

Hall effect, 35-12–35-13
Hall effect gage, 13-6
Hall effect sensor, 31-9–31-10
Heated-needle probe, 66-9
Heat flux
  calibration, 67-12–67-13
  calorimeter, 79-5
  heat transfer
    conduction, 67-2–67-3
    convection, 67-3
    definition, 67-2
    energy balance, 67-2
    first law of thermodynamics, 67-2
    radiation, 67-3
  measurement, 67-4–67-5
  surface heating, 67-12
  surface temperature history measurement
    calorimeter methods, 67-11
    semi-infinite surface temperature methods, 67-9–67-11
  temperature gradient
    insert gages, 67-8–67-9, 67-13
    one-dimensional planar sensors, 67-5–67-7
Heat transfer
  conduction, 67-2–67-3
  convection, 67-3
  definition, 67-2
  energy balance, 67-2
  first law of thermodynamics, 67-2
  radiation, 67-3
Helical turbine meters, 54-8–54-9
Helium–neon laser interferometer
  discharge tube, 27-2
  light emission, 27-2–27-3
  longitudinal Zeeman effect, 27-4–27-5
  resonant frequency, 27-3
Hidra nodes, 86-4–86-5
High-pressure rheometry, 46-26–46-27
High-purity germanium (HPGe) detectors
  coaxial detectors (*see* Coaxial detectors)
  cross section, 81-10–81-11
  depletion voltage, 81-12
  leakage current, 81-12–81-13
  process development, 81-17
  reliability, 81-17
  signal-to-noise ratio, 81-11
  x-ray detection, 81-16
High-temperature heat flux gage (HTHFS), 67-6–67-7
Hill-climbing technique
  gradient and random walk, 92-2
  local maxima and minima, 92-2–92-3
Hooke's law, 41-2–41-3
Horizon sensors, 16-4
Horizontal pendulum tiltmeters, 21-2
Hot cathode ionization gages
  Bayard–Alpert gage, 40-12–40-14
  triode, 40-11–40-12
Hot-disk instrument, 66-8–66-9
HPGe detectors, *see* High-purity germanium (HPGe) detectors
HUGIN 1000 autonomous underwater vehicle (AUV), 17-12
Hybrid elements, 23-3
Hybrid networks, 1-11
Hybrid stepping motors
  chopping current amplifiers, 95-2–95-3
  closed-loop control, 95-4
  full stepping mode, 95-2
  microstepping, 95-3
  open-loop operation, 95-2
  position measurement, 95-4–95-5
  torque-speed curve, 95-2
Hydraulic dynamometers, 44-12–44-13
Hydrometers, 45-4–45-5
Hydrostatic leveling, 21-5–21-6
Hydrostatic weighing densitometers, 45-5–45-6
Hymarc laser triangulation line scanner, 14-11–14-12
Hysteresis, 31-2–31-3

I

Immersion sensors, 68-7–68-8
Impedance plethysmography, 20-5–20-6, 20-16
Impediography, 49-8
Impeller flowmeters
  accuracy, 55-3
  applications, 55-5–55-6
  flow measurement capacity, 55-3
  flow ranges, 55-4
  in-line meters, 55-1–55-2
  insertion meters, 55-1–55-2
  installation
    accuracy, 55-5
    calibration, 55-4
    design features, 55-5
    operating pressure, 55-4
    pipe sizes, 55-4
    piping system restrictions, 55-4
  mechanical construction, 55-1
  principles, 55-3–55-4
  U.S. manufacturers, 55-6–55-7
Impulse response, 29-6
  Dirac delta function, 12-7
  experimental setup, 12-7–12-9
  FWHM, pulse duration, 12-10
  inertial force, 12-9
  instantaneous value, impact force, 12-9–12-10
  and momentum change, 12-10
  velocity and force measurement, 12-9

Indicator dilution methods
dynamic recording method, 20-13
first-pass method, 20-12–20-13
gas dilution, 20-13
principle, 20-11
radionuclide techniques, 20-12
thermodilution, 20-12
time–concentration curve, 20-11
Indirect reading gages
convection, 40-9–40-11
heat transfer regimes, 40-6–40-7
Pirani, 40-6–40-8
thermistor, 40-8
thermocouple, 40-8–40-9
Induction potentiometers, 32-1–32-2
Induction-type accelerometers, 36-11
Inductive displacement sensor
eddy current transducer, 24-17–24-19
electromagnetic interference shielding, 24-20
gradiometer, 24-19
linear-variable-reluctance transducer
single-coil, 24-3–24-5
tachogenerator, 24-6–24-7
variable-differential reluctance sensor, 24-5–24-6
LVDT
advantage of, 24-17
applications, 24-17
disadvantage, 24-17
ferromagnetic core/plunger, 24-12
frequency response, 24-16–24-17
phase-sensitive demodulator, 24-14–24-16
second-order system, 24-14
transfer function, 24-14
tubular ferromagnetic former, 24-12
voltages, 24-12–24-14
magnetic circuit, 24-1–24-3
microsyn, 24-7–24-8
moving-coil displacement
induction potentiometer, 24-11
linear-variable inductor, 24-9
variable-coupling transducer, 24-9–24-11
multiple-coil system, 24-19–24-20
rotary core differential transformer, 24-17–24-18
synchros, 24-8–24-9
types, 24-1–24-2
Inductive gaging, 13-5–13-6
Inductive method, 41-9
Inductive plethysmography, 20-6, 20-16
Inertia, 43-14
Inertial measurement units (IMU), 16-14
Inertial navigation
accelerometer, 17-8
in aircraft, 17-11–17-12
dead reckoning, 17-2
definition, 17-1
errors, 17-9
ESG, 17-2–17-3
flywheel gyroscopes, 17-5–17-7
inertial sensors, 17-2
laser gyro, 17-2
mechanical gyroscopes, 17-4–17-5
navigational platform, 17-2–17-3
optical gyroscopes, 17-7–17-8
in robotics, 17-13
stabilization, 17-9–17-10
strapdown system, 17-2–17-3
in underwater, 17-12
vehicular navigation, 17-10–17-11
Inertial sensors, 17-2
Influence parameters, 4-6
In-plane switching (IPS) LCD, 96-13
Input impedance, 3-6–3-7
Insert gages, 67-8–67-9, 67-13
Insertion axial turbine flowmeters, 54-10–54-11
Instrument-integrated closed-loop system, 95-7–95-8
Instruments
accuracy and errors, 1-4, 2-4–2-5
aerospace instrumentation, 1-13
analog, 1-6
calibration, 1-5–1-6, 2-3–2-4
construction structure, 1-2–1-3
definition, 1-2, 2-1
deflection method, 3-3–3-4
design, 1-2–1-3
digital, 1-7
drift, 4-6–4-7
environmental monitoring, 1-13
error reduction, 1-5, 2-7–2-8
estimation techniques, 2-8
industry and automobiles, 1-13
medical technology, 1-12–1-13
modifying and interfering inputs, 2-4–2-5
and networks, 1-10–1-11
passive and active sensors, 2-3
physical and signal variables, 2-1–2-2
power plants, 1-13
random-error sources (noise), 2-6–2-7
response and drift, 1-4
sensor fusion, 2-8
signal amplification, 2-2
simple model, 2-1–2-2
software support, 1-11–1-12
systematic error sources (bias)
compensation methods, 2-5
invasiveness, 2-6
miscalibration, 2-4
parallax error, 2-6
testing, 1-3, 2-8–2-9
use, 1-4
Integrated temperature sensor
analog-output sensor, 71-5–71-6
digital communication, 71-5

digital-output sensor, 71-5, 71-7
external circuitry, 71-7
Intelligent sensors
application, 9-5
communication, 9-4
definition, 9-1
IEEE 1451 standard
digital information, 9-3
NCAP, 9-4
object-oriented technology, 9-3
physical connection level, 9-2
seven sections, 9-2–9-3
simplified functionality, 9-2
TEDS, 9-3–9-4
intelligent instruments
architecture, 9-6
flowmeters, 9-7
functions, 9-6
microsurgical instruments, 9-7
operations, 9-6
pressure transmitters, 9-8
temperature transducer, 9-7
networking capability, 9-2
PnP capability, 9-4–9-5
Interfacial tension, 47-2
Interferometry, 13-7
International Standards Organization (ISO)
branches, 6-4
documents, 6-7
guides, 6-5
structure, 6-5–6-6
technical reports, 6-4–6-5
technical specifications and corrigendum, 6-5
ubiquitous, 6-7
worldwide proprietary standards, 6-4
International temperature scales, 65-9–65-11
Internet Calibration Services (iCals), 8-9
Interpolating constant-volume gas thermometer (ICVGT), 75-7–75-8, 75-11–75-12
Inter-Range Instrumentation Group (IRIG) standard, 90-3, 90-13–90-15
Intrinsic Fabry–Perot interferometric (IFPI) sensor, 33-4
Intrinsic safety (IS)
circuit, 94-6
field device, 94-6–94-7
ITC, 94-8
repeater barrier, 94-9
resistance temperature devices (RTD), 94-8
thermocouples, 94-8
Zener safety barrier, 94-7–94-8
Ionization chamber dosimeters, 84-7–84-8
Isolated temperature converters (ITC), 94-8
Isoperibol calorimeter, 79-5
Isothermal calorimeter, 79-4
iWave wireless bore gage, 28-7–28-8

K

Kerr effect, 77-9
Kinematic viscosity, 46-2
Kirchhoff's law, 72-4
Knife-edge photodetector, 34-2–34-4
KN-6050 underwater navigators, 17-12

L

LabVIEW programming, 10-6–10-7
Laplace pressure, 47-5
Laser-based active triangulation
light plane principle, 14-12
pinhole camera, 14-10–14-11
synchronized scanning principle, 14-11–14-12
Laser calipers, 13-5
Laser-Doppler anemometers, 59-1, 59-4, 59-12–59-13
Laser Doppler velocimetry (LDV)
advantages of, 64-2
data analysis, 64-12–64-13
flow velocity measurement, 64-2–64-3
laser beam, 64-2
operating principle
dual-beam system (*see* Dual-beam system)
frequency shifting, 64-6–64-7
multiple component velocity measurements, 64-2, 64-8–64-9
signal strength, 64-7–64-8
particle size, 64-13
phase Doppler system, 64-13–64-15
PIV, 64-1
scattering centers, 64-2
seeding, 64-11
signal processing, 64-9–64-11
Laser-flash instrument, 66-10–66-11
Laser heating, 84-5
Laser interferometers
applications, 27-13–27-15
error sources, 27-7
Fabry–Pérot, 27-12–27-13
helium–neon laser
discharge tube, 27-2
light emission, 27-2–27-3
longitudinal Zeeman effect, 27-4–27-5
resonant frequency, 27-3
Michelson interferometer, 27-7–27-9
refractive index, 27-5–27-7
signal processing, 27-12, 27-14
single-mode homodyne, 27-9–27-12
two-frequency heterodyne, 27-8, 27-10
Lateral force microscopy (LFM), 13-9
Lead magnesium niobate (PMN), 26-4
Lead zirconate titanate (PZT), 26-5–26-6
Level measurement
balance method, 18-5–18-6
capacitance-type level device, 18-13–18-14

conductive, 18-14
continuous level measurements, 18-1
displacers, 18-2–18-4
electrical properties, 18-13
float, 18-4–18-5
FMCW radar, 18-11–18-12
heat dissipation, 18-16
level switches, 18-1
microwave switch, 18-16
photoelectric barrier, 18-16
pressure gages, 18-4–18-6
pricing, 18-7
principal operational modes, 18-1–18-2
radiation attenuation, 18-14
radiometric level gaging, 18-15
techniques, 18-7
time-of-flight measurements
laser/light, 18-10–18-11
liquid, 18-12–18-13
microwaves, 18-8–18-10
operation principle, 18-7
ultrasonic waves, 18-8
ultrasonic switch, 18-15
viscosity, 18-16–18-7
Levitation mass method (LMM), 12-2
Lidar, 13-6, 14-7
AM LIDAR, 14-12–14-13
resonating LIDAR, 14-13
Light-emitting diode (LED) displays, 96-15
Linear array transducer, 49-6–49-7
Linear encoders, 30-18–30-19
Linear scale, 65-3
Linear-variable differential-transformer (LVDT), 36-20, 39-2–39-3
advantage of, 24-17
applications, 24-17
disadvantage, 24-17
ferromagnetic core/plunger, 24-12
frequency response, 24-16–24-17
phase-sensitive demodulator, 24-14–24-16
second-order system, 24-14
transfer function, 24-14
tubular ferromagnetic former, 24-12
voltages, 24-12–24-14
Linear-variable-reluctance transducer
single-coil, 24-3–24-5
tachogenerator, 24-6–24-7
variable-differential reluctance sensor, 24-5–24-6
Linear velocity
absolute angular velocity sensor, 35-14–35-16
angular velocity, 35-9–35-10
applications, 35-2
light interference, 35-5–35-7
measurement, 35-2–35-3
reference-based methods
average speed, 35-3–35-4
Doppler shift, 35-5
linear to rotational velocity, 35-5
parameters, 35-4
time interval, 35-3
velocity transducers (LVT), 35-4
relative angular sensor
counter types, 35-10–35-12
Hall effect, 35-12–35-13
optical methods, 35-11–35-12
tachometer generators, 35-10
Wiegand effect, 35-12–35-14
seismic devices, 35-8–35-9
Line-graduated instrument, 13-2
Liquid-crystal displays (LCD)
active-matrix LCD, 96-12
advantages, 96-12
IPS, 96-13
passive-matrix LCD, 96-12
performance, 96-12–96-13
twisted-nematic (TN) effect, 96-13
VA LCD, 96-13
Liquid crystal indicator, 76-14–76-16
Liquid-in-glass (LIG) thermometers
ASTM specifications, 74-3
components, 74-1–74-2
errors
hysteresis and stability, 74-9
immersion conditions, 74-5–74-7
parallax, 74-10
pressure, 74-10
scale markings, 74-4–74-5
separated columns, 74-11–74-12
stem correction, 74-7–74-9
stiction, 74-10
thermal capacity, 74-10–74-11
time-constant, 74-9–74-10
features, 74-2
operation of, 74-3–74-4
Liquid/pendulum tiltmeters, 21-2
Logarithmic natural scale, 65-2–65-3
Long-period grating sensor
fabrication procedure
fiber C, 33-12–33-13
highest-order cladding mode, 33-11
peak loss wavelength, 33-11–33-12
setup, 33-10
strain sensitivity, 33-11–33-12
transmission, 33-10–33-11
transmitted intensity, 33-12–33-13
types of, 33-11
operating principle, 33-9–33-10
temperature sensitivity, 33-13–33-14
Long-wave infrared (LWIR) detectors
atmospheric transmission, 78-5
electronics, 78-10
emission source, 78-4
photon detectors, 78-5
thermal imagers, 78-15

trade-offs, 78-17
visual optics, 78-10, 78-13
wave band, 78-2–78-3
LVDT, *see* Linear-variable differential-transformer (LVDT)
LWIR detectors, *see* Long-wave infrared (LWIR) detectors

M

Mach number, 62-8–62-10
Magnetic circuit, 24-1–24-3, 24-17–24-18
Magnetic displacement sensor
Hall effect, 31-9–31-10
magnetic encoder, 31-10–31-12
magnetic field, 31-2–31-4
magnetostrictive sensor (*see* Magnetostrictive sensor)
manufacturers, 31-12
position sensor, 31-2
Magnetic drives, 45-10
Magnetic encoders, 31-10–31-12
Magnetic field intensity, 31-2
Magnetic flux density, 31-2
Magnetic force microscopy (MFM), 13-9
Magnetic hysteresis, 31-3–31-4
Magnetic induction, 13-6
Magnetic permeability, 31-2
Magnetic position tracking, 14-15–14-16
Magnetic resonance imaging, 20-2
Magnetic saturation, 31-3
Magnetic sensors, 22-15
Magnetization curve, 31-3–31-4
Magnetoelastic force sensor, 41-14
Magnetohydrodynamic (MHD) effect, 35-15–35-16, 56-9
Magnetometers, 16-5
Magnetoresistive force sensor, 41-13
Magnetoresistive sensor, 31-8–31-9
Magnetostrictive sensor
dead zone, 31-6
distributed impedance sensor, 31-7–31-8
domains, 31-5
magnetoresistance, 31-8–31-9
operation of, 31-6
position magnet, 31-5
time measurement, 31-6–31-7
Villari effect, 31-6
Wiedemann effect, 31-6
Magnetostrictive transducer, 29-4–29-5
Manometric thermometers
gas thermometer
acoustics, 75-8, 75-10, 75-12
CVGT, 75-5, 75-7–75-8, 75-11–75-12
vapor pressure
design of, 75-4–75-7
equations, 75-2–75-4
temperature range, 75-2
Mass flow controller (MFC), 50-1–50-2; *see also* Capillary-type thermal mass flow meter
Mass traceability, 43-16
Mass, weights, and instrumentation
accuracy, 43-14–43-15
calibration standards
classification, 43-12
design, 43-13
handling, storage, packing, and shipping, 43-13
materials, 43-13
maximum permissible error (MPE), 43-13
definitions, 43-2–43-3
etymology, 43-2
inertia, 43-14
linearity, 43-15
mass traceability, 43-16
metric unit of, 43-17–43-18
range, 43-14
replaceability, 43-14
safety measures and precautions, 43-16
sensitivity, 43-15–43-16
stability, 43-14
test instrumentation
balances and mass comparators, 43-7
classification of, 43-5
equal arm balance, 43-6–43-7
mass measuring technique, 43-5–43-6
scales, 43-7
theory, 43-3–43-5
weighing method
equal arm balance, 43-8–43-10
single-pan optical-mechanical/electronic scales, 43-10
substitution method, 43-11
transposition, 43-12
Maximum bubble pressure method (MBPM), 47-8–47-9
Maximum permissible error (MPE), 43-13
Maximum workplace concentration (MWC), 27-5
Measurement accuracy
engineering classifications
random standard, 5-8
symmetrical systematic standard, 5-9
systematic standard, 5-8–5-9
error limits, 5-4
expanded uncertainty
degrees of freedom, 5-11
engineering system, 5-10
round-off errors, 5-12
temperature measurement, 5-11
ISO classifications
degrees of freedom, 5-5–5-6
dependent error sources, 5-7
equation, 5-8
independent sources of error, 5-6–5-7
sensitivity, 5-5–5-6

standard deviation, 5-4–5-5
total uncertainty, 5-6
normal and uniform distribution
distribution curve, 5-2–5-3
error, definition, 5-1
limit estimation, 5-2
standard deviation, 5-2–5-4
Measurement standards
accuracy of measurement, 7-7
calibration, 7-7
calibration laboratory, 7-7
coherent unit, 7-8
conservation standard, 7-8
conversions of units, 7-7, App-4–App-8
CRM, 7-8
forensic metrology, 7-3
interlaboratory standard, 7-8
international standard, 7-8
international system of units, 7-8
legal metrology, 7-3
Lord Kelvin's oft-quoted statement, 7-4–7-5
measurand, 7-8
method of measurement, 7-9
metrology, 7-9
multiplication factors, 7-6, App-3
national reference standard, 7-9
national standard, 7-9
"natural" phenomena, 7-2
need for, 7-5
NIST, 7-9
numbers, dimensions, and units, 7-6
primary standard, 7-9
principle of measurement, 7-9
protocol standards, 7-3
reference standard and material, 7-9
repeatability, 7-9–7-10
reproducibility, 7-10
ruler, 7-1
secondary standard, 7-10
SRMs, 7-4
standardization organizations and standards structure, 7-2–7-3
standards laboratory, 7-10
tolerance, 7-10
traceability, 7-10–7-11
transfer standard, 7-11
traveling standard, 7-11
types
basic/fundamental standards, 7-5, App-2
derived standards, 7-5, App-3
measurement assurance system, 7-5–7-6
uncertainty of measurement, 7-11
value (of quantity), 7-11
weights and measures, 7-1
working standard, 7-11
Measurement system, 1-10
Measurement uncertainty, *see* Measurement accuracy
Mechanical gyroscopes, 17-4–17-5
Mechanical planimeters, 19-7
Medium-wave infrared (MWIR) detectors
atmospheric transmission, 78-5
electronics, 78-10
emission source, 78-4
photon detectors, 78-5–78-6
thermal imagers, 78-15
trade-offs, 78-17
visual optics, 78-10
wave band, 78-2–78-3
Mercury tiltmeter, 21-2
Mica mote, 86-4–86-5
Michelson interferometer, 27-7–27-9
Microaccelerometers, 36-24–36-26, 36-31–36-32
Micro-instrument, 1-8
Micrometer bore gages, 28-6–28-7
Microprocessor-based digital thermal anemometer, 59-8, 59-10
Microprocessor-controlled closed-loop system, 95-8–95-9
Microsyn, 24-7–24-8
Microsystems technology (MST), 1-8
Mineral-insulated, metal-sheathed (MIMS) thermocouple material, 70-24
Miniature-inductive displacement sensor, 24-19–24-20
Minimum resolvable temperature difference (MRTD), 78-14
M-mode display, 49-20
Mobile instrumentation
components, 87-2
EVAN, 87-8
advantages, 87-10
"always-on" mode, 87-8
design, 87-9–87-10
electric system behavior tracking, 87-8
microcontroller, 87-8
peripherals, 87-10
physical and electric footprint advantages, 87-11
power consumption, 87-8
power supply design considerations, 87-10–87-11
system-wide implementation, 87-3
PCVAN, 87-4–87-5
sneaker net, 87-2
telematics hub, 87-2–87-3
vehicle and mobile assets, 87-5–87-7
wireless wheel vibration and temperature sensor, 87-2–87-3
Modular azimuth position system (MAPS), 17-13
Modulation transfer function (MTF), 78-12–78-13
Modulus of rigidity, 44-3
Moiré fringe method, 30-14–30-16
Moiré interference, 37-16–37-17
Molecular drag gage (MDG), 40-16–40-18
Molecular viscosity, 46-2
Moment arm, 44-2

Momentum transport process, 46-1
Monochromatic-brightness radiation thermometer, *see* Disappearing filament pyrometer
Morphotropic phase boundary (MPB), 26-5–26-6
Movable vane meter, 52-2, 52-4–52-5
Moving-coil inductive displacement sensor
  induction potentiometer, 24-11
  linear-variable inductor, 24-9
  variable-coupling transducer, 24-9–24-11
Moving pointers, 96-10
Multijet turbine flowmeters, 54-12
Multipoint contact bore gages, 28-9
MWIR detectors, *see* Medium-wave infrared (MWIR) detectors

N

Nanoscale scanning probe microscopy, 13-8–13-9
Navigation accuracy measurement system (NAMS), 15-3–15-4
NBS, *see* Neutrally buoyant sensor (NBS)
Near-field acoustic holography (NAH)
  assumption, 48-16
  principle, 48-19–48-20
  spatial frequency, 48-16–48-18
Negative temperature coefficient (NTC) thermistors, 69-1, 69-11–69-13
Network Capable Application Processor (NCAP), 9-4
Network operating systems (NOSs), 1-11
Neutrally buoyant sensor (NBS)
  borehole-accessed caves, 89-8–89-9
  cave mapping
    components, 89-6
    3D conduit imagery, 89-7
    design trade-offs, 89-7–89-8
    electronics and housing, 89-7
    operations, 89-6–89-7
    SwRI®, 89-6
  culvert inspection, 89-9–89-10
Neutron measurement
  detectors
    capture reactions, 83-3
    characteristics, 83-10–83-11
    commercial suppliers, 83-11
    efficiency calculations and measurements, 83-7–83-10
    elastic scattering, 83-2–83-3
    fission process, 83-4
    inelastic scattering, 83-4
    (n,p)/(n,α) reactions, 83-4–83-6
    standards, 83-5
  nuclear reactions, 83-1–83-2
Newtonian fluid, 46-3
Niveltronic tiltmeter, 21-2
Noise-equivalent temperature difference (NETD), 78-14
Noncontact sensing, 38-1
Noncontact thermometers
  absolute zero, 72-2
  applications
    medical thermometers, 72-6, 72-18–72-20
    semiconductor material, measurement of, 72-18
  blackbody
    cavity effect, 72-7–72-8
    gold mirror, 72-9
    immersion, 72-9
    solid-state, 72-8–72-9
  disappearing filament pyrometer, 72-14–72-16
  electrodynamics law, 72-1–72-2
  emissivity, 72-4–72-6
  IR thermometry, 72-17
  Kirchhoff's law, 72-4
  Planck's law, 72-2–72-3
  quantum detectors, 72-9–72-10
  Stefan–Boltzmann law, 72-3–72-4
  thermal detectors
    bidirectional radiation flow, 72-11–72-12
    bolometer, 72-13
    operating ranges, 72-11
    output electric signal, 72-12
    radiative heat exchange, 72-11
    reference temperature sensor, 72-13
    temperature, 72-12
    thermal imaging camera, 72-14
    thermopile IR sensor, 72-13–72-14
  two-color pyrometer, 72-16–72-17
  Wien's law, 72-3
Nondestructive testing/evaluation, 49-2, 49-22
Non-Newtonian fluid, 46-3
Nozzle, 51-7, 51-11
Nutating-disk-type positive displacement meter, 53-5

O

OBD sensor, *see* Optical beam deflection (OBD) sensor
Occluder devices, 38-11–38-12
Omnidirectional radar, 15-6–15-7
One-dimensional planar sensors, 67-5–67-7
Open channel flow measurements
  applications, 57-9
  multipath, 57-13
  rating curve, 57-9
  sensor alignment, 57-5
  single path
    area–velocity method, 57-9
    corrective factor, 57-12
    flow rate, 57-9
    logarithmic k_h factor, 57-11
    power law, 57-11–57-12
    velocity profile, 57-10
    vertical velocity distribution, 57-9–57-10
Operational modes
  analog and digital readout, 3-5–3-6
  analog and digital sensors, 3-4–3-5

deflection instrument, 3-3–3-4
input impedance, 3-6–3-7
null instrument, 3-1–3-2
Optical beam deflection (OBD) sensor
bicell detector, 34-4–34-5
characteristics
operation, 34-6–34-7
sum voltage *vs*. displacement, 34-7–34-9
transimpedance amplifier, 34-6
UDT SPOT2D, 34-6–34-7
voltage difference, 34-6–34-7
continuous position sensor, 34-6
knife-edge photodetector, 34-2–34-4
Optical birefringence, 38-13
Optical comparator, 42-5
Optical encoders
absolute encoders
direct binary, 30-2–30-3
gray code, 30-4–30-5
optical resolving, 30-6
pseudorandom code, 30-4–30-7
Vee-Scan, 30-3–30-4
accuracy, 30-2
analog quadrature signal
resistor network, 30-10–30-12
sampling and numerical processing, 30-12–30-13
classification of, 30-1–30-2
components and technology, 30-19–30-21
digital quadrature signal
commercial chips and suppliers, 30-10, 30-21
count signal, 30-8–30-9
divide-by-four counter circuit, 30-10, 30-21
flip-flops, 30-8
rotary incremental encoder, 30-7–30-8
principles
canon laser rotary encoder, 30-17
diffraction-based encoders, 30-16–30-17
geometric masking, 30-13–30-14
linear, 30-18–30-19
Moiré fringe method, 30-14–30-16
rotary, 30-17
revolution per minute (RPM), 30-2
Optical fiber displacement sensor
EFPI, 33-2–33-3
FBG sensor
fabrication techniques, 33-6–33-8
limitation of, 33-8–33-9
operating principle, 33-5–33-6
wavelength, 33-8
IFPI, 33-4
interferometric *vs*. grating-based sensors, 33-14
long-period gratings
fabrication procedure, 33-10–33-13
operating principle, 33-9–33-10
temperature sensitivity, 33-13–33-14
Optical fiber strain gage
intensity variation, 61-7
optical root strain magnitude, 61-7
optical strain direction, 61-8
principle, 61-7
with rubber beam, 61-8
wind velocity, 61-7
without rubber beam, 61-8–61-9
Optical gyroscopes, 17-7–17-8
Optical pyrometer, *see* Disappearing filament pyrometer
Optical responsivity (K), 34-7
Optical transducer, 38-10–38-12
Optimal control
calculus variations, 92-6–92-7
cost functions/performance index
control vector, 92-3
definition, 92-3
integral absolute error (IAE), 92-3–92-4
integral squared error (ISE), 92-4
integral time absolute error (ITAE), 92-4
integral time squared error (ITSE), 92-4
mean square error (MSE), 92-4
minimum energy consumption, 92-5
minimum fuel consumption, 92-5
minimum time control, 92-5
hill-climbing technique
gradient and random walk, 92-2
local maxima and minima, 92-2–92-3
optimization, 92-1
Riccati equation, 92-8–92-11
software support and applications, 92-14–92-15
state-feedback matrix, 92-11–92-14
OPTOTRAK position-tracking system, 14-14–14-15
Orifice/cup, short capillary viscometer, 46-12
Orifice plate
advantages, 51-1, 51-6
correction factors, 51-5
disadvantages, 51-6
expansibility factor, 51-5
schematic diagram, 51-4–51-5
standards, 51-5–51-6
vena contracta, 51-5
volumetric flow rate, 51-5
Oscillating cup viscometer, 46-17–46-20
liquid metals, 46-20
logarithmic decrement of oscillations (δ), 46-18–46-19
working principle, 46-17–46-18
working Roscoe's formula, 46-19
Oscillation response
damping, 12-13
experimental setup, 12-10–12-11
$F_{inertial}$ and F_{sensor}, 12-12–12-13
frequency, velocity, acceleration and force measurement, 12-11–12-12

measurement, interferometer, 12-12–12-13
validity and reproducibility, 12-14
Ostwald viscometer, 46-12
Oval-gear positive displacement meter, 53-4–53-5

P

Paddle wheel flowmeters, *see* Impeller flowmeters
Panoramic annular lens (PAL), 16-17–16-18
Paralleled thermoelements, 70-16–70-17
Parallel-plate instrument
fluid measurement, 66-4–66-5
schematic diagram, 66-3
solid measurement, 66-5–66-6
Partial immersion thermometer, 74-2, 74-6–74-7
Particle emission radioactivity
alpha particles, 80-3–80-4
beta particles, 80-4–80-6
fission fragments, 80-7
neutrons, 80-5–80-6
protons, 80-7
Particle image velocimetry (PIV), 64-1
Particle-tracking microrheology, 46-23–46-24
Passive sensors, 2-3
PC-based digital thermal anemometer system, 59-8–59-9
PC-based hub design approach (PCVAN)
development tools, 87-4
form factor, 87-4
I/O considerations, 87-5
memory, 87-5
microprocessors, 87-5
wireless network support, 87-4–87-5
PCM, *see* Pulse code modulation (PCM)
Pellets, 76-6, 76-8–76-9
Pendant drop method, 47-9–47-10
Penstock leak detection, 57-13–57-15
Permanent magnet accelerometer, 36-8–36-9
Perovskite, 26-5–26-6
Phase-angle method, 29-5
Phase angle voltmeter (PAV), 32-5–32-6
Phase-based detection, 14-12–14-13
Phased array transducer, 49-6
Phase Doppler system, 64-13–64-15
Phase-modulated (PM) proximity sensor
amplitude, 22-5
configuration, 22-3
LED driving signals, 22-4–22-5
measurement parameters, 22-4–22-5
phase shift, 22-5
sensor head, 22-4
signal attenuation, 22-4
Phase-shift keying (PSK), 85-8
Photoelastic transducer, 38-13–38-14
Photoelectric barrier, 18-16
Photon detector, 78-5–78-6
Photothermal optical beam deflection (PTOBD) spectroscopy, 34-1
Phototransferred thermoluminescence (PTTL), 84-6–84-7
Piezoelectric accelerometer, 36-11–36-13, 36-29–36-31
Piezoelectric drives, 45-10
Piezoelectricity
applications
active modes, 26-10
displacement sensor, 26-10–26-11
passive modes, 26-10
polymer, 26-11
suppliers, 26-12
Young's moduli, 26-10
definition of, 26-2
direct and converse effect, 26-2
elastic and electric stresses, 26-1–26-2
electromechanical coupling coefficient, 26-3
materials
ceramics, 26-4–26-7
perovskite, 26-5–26-6
polymers, 26-6–26-7
single-crystal, 26-3–26-4
measurements
equivalent circuit, 26-7–26-9
FET-input operational amplifier, 26-9–26-10
frequency-independent response, 26-10
resonance technique, 26-7
subresonance technique, 26-8
vibration modes, 26-8
third-rank tensor, 26-2
Piezoelectric method
crystals, 41-9
electrorestrictive elements, 41-7
inductive, 41-9
magnitude and polarity, 41-7–41-8
materials, 41-8
multicomponent dynamometers, quartz crystals, 41-10–41-12
operating modes, 41-8
piezotransistor, 41-10
resistive, 41-9
voltage output, 41-8–41-9
voltage sensitivity, 41-8
Piezoelectric strain gages
advantage of, 37-14
electrometer and charge amplifier, 37-13–37-14
FET input, 37-13
lead zirconate titanate (PZT), 37-13
polarization, 37-12
PVDF film, 37-11–37-12
Taylor's series, 37-12
Piezoelectric transducer, 38-8–38-10
Piezoresistive accelerometer, 36-14, 36-31
Piezoresistive effect, 39-5
Piezoresistive foil/wire strain gages

application, 37-8
area strain, 37-7–37-8
construction of, 37-6
gage factor, 37-6
heat sinking ability, 37-10
Mohr's circle, 37-10–37-11
Poisson's ratio, 37-7–37-8
relative resistivity, 37-8
resistance change, 37-6–37-7
self-temperature compensation (STC), 37-7
sensitivity, 37-10
thermal drift, 37-9
Wheatstone bridge, 37-9
Piezoresistive pressure sensor, 39-5–39-6
Piezoresistive semiconductor strain gages, 37-11
Piezotransistor method, 41-10
Pinhole camera, 14-10–14-11
Pirani gage, 40-6–40-8
Pistol-grip bore gages, 28-8–28-9
Piston and flight control surface
aircraft short-period approximation, 93-9
flexible hinge joint model, 93-6–93-7
hinge moments, 93-8
load components, 93-6
stiff hinge joint model, 93-7–93-8
Piston-type positive displacement meter, 53-3–53-4
Pitot probe anemometer
hybrid system, 62-12–62-14
manufacturers/vendors, 62-14–62-15
theory
angle of attack, 62-6–62-7
Bernoulli's equation, 62-1–62-2
design parameters, 62-7
fluid velocity, 62-4
nonrotating solid body, 62-3
pressure difference, 62-3
pressure tap flow field, 62-6
stagnation point, 62-3
static holes, 62-5
tube arrangement, 62-4
tube method, 62-5
variable density
binomial theorem, 62-8
compressibility factor, 62-9
high-velocity gas stream, 62-7–62-8
Mach number, 62-8–62-10
Rayleigh's formula, 62-10
shock wave, 62-10
subsonic flow, 62-8
supersonic flow, 62-9
volumetric flow measurements, 62-10–62-11
Pitot tubes, 59-1, 59-12–59-13
Planck's law, 72-2–72-3
Planimeter
basic wheel and bar mechanism, 19-2
battery-operated planimeters, 19-8
electronic planimeters, 19-7–19-8
evaluation, 19-10–19-11
manufacturers, 19-10
measuring wheel, 19-1–19-2
mechanical planimeters, 19-7
operation, 19-3
polar planimeter, 19-1–19-2
Plasma displays, 96-13–96-14
Platinum resistance thermometer (PRT), 68-3–68-4
Plethysmography
acoustic plethysmography, 20-16
air/water plethysmography, 20-3–20-4, 20-15
impedance plethysmography, 20-5–20-6, 20-16
inductive plethysmography, 20-6, 20-16
strain-gage plethysmography, 20-4
Plug-and-play (PnP) capability, 9-4–9-5
Pneumatic controllers, 91-6
Pneumatic gages, 13-5
Point spread function (PSF), 78-12–78-13
Polariscope, 38-14
Polar planimeter, 19-1–19-2
Policy development committees, 6-5
Polymer capacitive humidity sensor, 25-12–25-13
Popular pad digitizer, 19-9
Position-sensing detector (PSD), *see* Optical beam deflection (OBD) sensor
Position tracking
active target approaches, 14-14
devices, 14-7, 14-15
magnetic position tracking, 14-15–14-16
OPTOTRAK position-tracking system, 14-14–14-15
Positive displacement (PD) flowmeters
accessories, 53-8
advantages/disadvantages, 53-7
birotor type, 53-3–53-4
CVM, 53-6
design and construction, 53-1–53-2
diaphragm, 53-6–53-7
differential pressure, 53-8
dimensional control, 53-7
fluid viscosity, 53-7
leakage flow, 53-7
nutating-disk-type, 53-5
oval-gear, 53-4–53-5
piston-type, 53-3–53-4
price and performance, 53-8
primary losses, 53-7
roots, 53-5–53-6
sliding-vane type, 53-2–53-3
trirotor type, 53-3
Positive temperature coefficient (PTC) thermistors, 69-13–69-14
Potentiometer accelerometer, 36-20
Power measurements
absorption dynamometers
drag torque, 44-11–44-12
eddy current dynamometer, 44-13
friction brake, 44-11

Froude brake, 44-12
hydraulic dynamometers, 44-12–44-13
prony brake, 44-11–44-12
water brake, 44-12
costs, 44-15
driving and universal dynamometers, 44-14
foursquare dynamometers, 44-14
measurement accuracy, 44-15
Precision potentiometers
electrical characteristics
AC excitation, 23-6
electrical travel, 23-2, 23-4
linearity, 23-4–23-5
loading, 23-2, 23-4–23-5
power rating, 23-6
resistance, 23-6
resolution, 23-5–23-6
taper, 23-2, 23-4
temperature coefficient, 23-2, 23-6
terminals and taps, 23-3–23-4
mechanical characteristics
contamination and seals, 23-8
lifetime, 23-8
mechanical loading, 23-6–23-7
mechanical travel, 23-2, 23-7
operating temperature, 23-7
shaft misalignment, 23-8
speed, 23-7
vibration, shock, and acceleration, 23-7
mounting method, 23-8–23-9
resistive element, 23-2–23-3
types of, 23-1–23-2
Pressure gages, 18-4–18-6
Pressure measurement
definitions, 39-1–39-2
detection methods
capacitive, 39-3–39-5
LVDT, 39-2–39-3
piezoresistive, 39-5–39-6
elements, 39-2–39-3
silicon micromachined sensor
construction of, 39-6
limitations, 39-7–39-9
wafer-to-wafer bonding, 39-7
Primary batteries, 88-2
Probability density function, 28-13–28-15
Programmable logic controller (PLC), 32-9
Propeller meters, 54-8
Proportional-bandwidth modulation (PBW), 90-8–90-9
Proportional-integral-derivative (PID) controller
block diagram, 91-5–91-6
closed-loop responses, 91-3
derivative time, 91-5
digital controllers, 91-7
electronic controllers, 91-6–91-7
expression, 91-5
functional blocks, 91-1–91-2
integral controller, 91-4
integral time, 91-5
open-loop responses, 91-2
PD controller, 91-5
PI controller, 91-4–91-5
pneumatic controllers, 91-6
proportional controller, 91-4
tuning controllers, 91-7–91-8
Protractor, 42-5–42-6
Proximity sensors, robotics
amplitude modulation
calibration-per-surface strategy, 22-9
configuration, 22-6, 22-9
distance measurement, 22-8
light intensities, 22-7–22-8
symmetrical three-fiber constellation, 22-7
Y-guide transducer, 22-6–22-7
capacitive sensors, 22-13–22-14
characteristics, 22-2
electro-optical sensors, 22-3
geometrical techniques, 22-10–22-13
magnetic sensors, 22-15
phase modulation
amplitude, 22-5
configuration, 22-3
LED driving signals, 22-4–22-5
measurement parameters, 22-4–22-5
phase shift, 22-5
sensor head, 22-4
signal attenuation, 22-4
photothermal effect, 22-13
proximity parameters, 22-1–22-2
time-of-flight-measuring electro-optical sensors, 22-13
ultrasonic sensors, 22-14–22-15
Pseudoplasticity, 46-4
Pseudorandom binary sequence (PRBS), 30-4–30-6
Pseudorandom code, 30-4–30-7
Published standards
definition, 6-1
design and operations, 6-8
instrumentation, measurement, and related fields, 6-8–6-9
ISO
branches, 6-4
documents, 6-7
guides, 6-5
structure, 6-5–6-6
technical reports, 6-4–6-5
technical specifications and corrigendum, 6-5
ubiquitous, 6-7
worldwide proprietary standards, 6-4
online purchase, 6-7
software standards
instrumentation and measurements, 6-10–6-11
software, definition, 6-8

standard organizations, 6-8
types, 6-10
standards developing organizations (SDOs)
development, 6-4
partial list, national/international organization, 6-2
preparation process, 6-3
relations, 6-3–6-4
World standards organizations, 6-2
Pulse amplitude modulation (PAM), 90-12
Pulse code modulation (PCM)
basic elements, 90-13
bit error rate *vs*. SNR, 90-15–90-16
delta modulation, 90-13
differential pulse code modulation, 90-13
frame structure, 90-13–90-15
IRIG standard, 90-13–90-15
specifications, 90-13–90-14
Pulse duration modulation (PDM), 90-12
Pulse-echo method, 29-5
Pulse position modulation (PPM), 90-12
Pulse-width-modulated (PWM) power amplifier, 95-5–95-6
Pycnometric densitometers, 45-3–45-4
Pyroelectric detectors
applications, 73-8–73-9
capacitance, 73-6
coefficients, 73-6
effects, 73-1–73-2
equivalent circuit, 73-5
heat function, 73-7–73-8
manufacturing of, 73-4–73-5
material, 73-2–73-3
polarization, 73-6
thermal radiation, 73-6–73-8
time constant, 73-7
Pyroelectric effect, 73-1–73-2

Q

Quadrilateration, 15-8–15-10
Quantitative and status displays, 96-10
Quantitative feedback theory (QFT) compensator, 93-9–93-11
Quantum detectors, 72-9–72-10
Quarter wavelength impedance matching, 49-14

R

Rack-and-pinion mechanisms, 23-10
Radar, 13-6, 14-6
Radar altimeter, 15-2
Radiation dosimetry
absorbed dose, 84-2
bubble dosimetry, 84-11
dose equivalent, 84-2
electronic personal dosimeter, 84-12
film dosimetry, 84-8–84-9
goal, 84-1
ionization chamber dosimeters, 84-7–84-8
quality factor, 84-2–84-3
radiation weighting factors, 84-2
TL dosimetry
activators, 84-3
instrumentation, 84-5
laser heating, 84-5–84-6
phosphors, 84-3–84-5
PTTL, 84-6–84-7
RPL, 84-7
signal-to-noise ratio, 84-6
traps and luminescence centers, 84-3
vendors, 84-6–84-7
track-etch dosimetry, 84-10–84-11
units and standards, 84-1
Radioactive densitometers, 45-13–45-14
Radioactivity
annihilation radiation, 80-9
bremsstrahlung
beam transmission equation, 80-10
mean range, 80-11–80-13
narrow-beam attenuation coefficient, 80-9–80-10
production, 80-9
stopping power, 80-11–80-12
cosmogenic and naturally radionuclides, 80-1–80-2
definition, 80-1
electromagnetic emission radioactivity
conversion electrons, 80-8
gamma rays, 80-7–80-8
x-rays, 80-7
gaseous detectors
charge transport, 81-2–81-3
electric field effects, 81-3
Geiger–Müller counters, 81-6–81-7
ionization chambers, 81-4
operating principles, 81-2
proportional counters, 81-5–81-6
scintillators (*see* Scintillators)
versatility, 81-2
W values, 81-2
HPGe detectors
coaxial detectors (*see* Coaxial detectors)
cross section, 81-10–81-11
depletion voltage, 81-12
leakage current, 81-12–81-13
process development, 81-17
reliability, 81-17
signal-to-noise ratio, 81-11
x-ray detection, 81-16
measurements, 80-14–80-17
measurement unit, 81-2
neutron interactions, 80-13–80-14
nuclear radiations types, 80-1–80-2

particle emission radioactivity
alpha particles, 80-3–80-4
beta particles, 80-4–80-6
fission fragments, 80-7
neutrons, 80-5–80-6
protons, 80-7
prices and availability, 81-25
room-temperature semiconductors
energy gap, 81-22
operating principles, 81-23–81-24
performance, 81-24
thickness and surface area limitations, 81-25
silicon detectors
CCD detectors, 81-21–81-22
charged-particle spectroscopy, 81-22
diffused junction detector, 81-18
drift detectors, 81-21
energy resolution, 81-19–81-20
impurity requirements, 81-17
ion-implanted detector, 81-18–81-19
radiation damage, 81-22
spatial resolution, 81-20
strip detectors, 81-20
surface barrier detector, 81-18
temperature, 81-17
specific activity, 80-3
stable nuclei, 80-3
Radiometric level gaging, 18-15
Radiophotoluminescence (RPL), 84-7
Random error, 2-7
Range differences height monitoring unit (RDHMU), 15-11–15-12
Range imaging devices, 14-7–14-8, 14-15
Ratio thermometer, *see* Two-color pyrometer
Rayleigh's formula, 62-10
Rectangular distribution, 5-3–5-4
Rectified signal, 4-8–4-9
Refraction densitometers, 45-14–45-15
Refractive index, 27-5–27-7
Relative angular sensor
counter types, 35-10–35-12
Hall effect, 35-12–35-13
optical methods, 35-11–35-12
tachometer generators, 35-10
Wiegand effect, 35-12–35-14
Relaxation absorption, 49-14
Rene mote, 86-4–86-5
Repeatability, 4-3
Repeatability and reproducibility (R&R) methodology
application, 28-15–28-16
probability density functions, 28-13–28-15
Replaceability, 43-14
Reproducibility, 4-4
Research thermal anemometer, 59-3–59-5
Residual gas analyzer (RGA), 40-18
Residual nitrogen analyzer (RNA), 40-18
Resistance temperature detector (RTD), 63-3–63-4
calibration, 68-17, 68-19
construction, 68-4–68-5
errors sources, 68-10–68-11
immersion sensors, 68-7–68-8
insulation resistance, 68-12
interchangeability, 68-11–68-12
interpolation, 68-17
lead wires, 68-6, 68-17–68-18
maintenance, 68-19
performance, 68-10
placement, 68-9
price, 68-10
protection, 68-9–68-10
protective housing, 68-6
repeatability, 68-14–68-15
resistive material, 68-1
R vs.T relationship
alpha values, 68-2–68-3
copper, 68-2
definition, 68-1–68-2
fixed-point calibration, 68-3
nickel, 68-2
platinum, 68-2–68-3
PRT, 68-3–68-4
self-heating, 68-18–68-19
sensing elements, 68-5–68-6
service life, 68-10
stability, 68-13–68-14
stem conduction, 68-16–68-17
surface-mounted sensors, 68-8–68-9
thermal EMF, 68-19
thermal hysteresis, 68-15–68-16
time response, 68-19
Resistive displacement sensor
characteristics, 23-1–23-2
costs and sources, 23-12–23-13
external shaft, 23-1–23-2
implementation, 23-9, 23-11–23-12
measurement techniques, 23-9–23-10
precision potentiometers (*see* Precision potentiometers)
voltage divider, 23-1–23-2
Resistive method, 41-9
Resistive strain gages, 61-5–61-7
Resistive transducer, 38-5–38-6
Resolver
amplitude modulation, 32-2–32-3
applications
bandwidth, 32-11
encoder emulation, 32-11
industrial application, 32-10
large step response, settling time for, 32-12–32-13
position lag error, acceleration, 32-12
resolver-to-digital converter, 32-10–32-11

benefits, 32-6–32-7, 32-9
control transmitter, 32-2–32-3
errors, 32-4
functional parameters, 32-5
induction potentiometers, 32-1–32-2
modular approach, 32-7
multislot lamination, 32-2
phase analog, 32-4
rotor winding, 32-2–32-3
Scott "T" device, 32-7
shaft angle encoding, 32-8–32-9
single-speed stator output, 32-2–32-3
testing, 32-5–32-6
type II servo loop, 32-9
Resonance-type vacuum gage, 40-15
Rheology, 46-3
Ring-type load cells, 41-7
Rising bubble/droplet viscometer, 46-16–46-17
Robotic tactile sensing
classification, 38-1
mechanical effects, 38-2
requirements, 38-4–38-5
technologies
capacitive, 38-6–38-8
optical, 38-10–38-12
photoelastic, 38-13–38-14
piezoelectric, 38-8–38-10
resistive, 38-5–38-6
theory for, 38-2–38-4
Rockwell WINS node, 86-4–86-5
Room-temperature semiconductors
energy gap, 81-22
operating principles, 81-23–81-24
performance, 81-24
thickness and surface area limitations, 81-25
Roots positive displacement meter, 53-5–53-6
Rotameter
advantages, 52-4
applications, 52-3
bob geometries, 52-3
characteristics, 52-4
computation of internal flow, 52-14–52-15
computed flow field, 52-15–52-18
cross section, 52-1–52-2
flow rate analysis, 52-8–52-9
prices, 52-2
Reynolds number, 52-10
Ruppel number, 52-10–52-11
Rotary core differential transformer, 24-17–24-18
Rotary encoders, 30-17
Rotating magnet sensor, 35-11
Rotational motion, 44-1–44-2
Rotational-type servo accelerometer, 36-9–36-10
Rotational viscometers
advantage, 46-5
apparent fluid viscosity, 46-9
Couette-type viscometer, 46-7
fluid shear stress, 46-7
representative radius, 46-8
Reynolds number, 46-9
Searle-type viscometer, 46-7
shear rate, 46-7
RTD, *see* Resistance temperature detector (RTD)

S

Sagnac interferometer, 17-7–17-8
Saturation/clipping signal, 4-8–4-9
Saybolt viscometer, 46-12
SCADA, *see* Supervisory control and data acquisition (SCADA)
Scale markings, 96-10
Scanning calorimeter
DSC, 79-7–79-9
heat-conducting disk, 79-6–79-7
measurement curve, 79-9
Scanning probe microscopes (SPMs), 13-8
Scanning tunneling microscope (STM), 13-8–13-9
Scintillators
efficiency, 81-8
inorganic group, 81-8–81-10
organic group, 81-10
property, 81-7–81-8
pulse decay time, 81-8
Searle-type viscometer, 46-7
Secondary batteries, 88-7
Secondary surveillance radar (SSR)
omnidirectional radar, 15-6–15-7
pseudoranges, 15-10–15-11
RDHMU, 15-11–15-12
Seebeck thermoelectric effect, 70-2–70-5
Seger cones, 76-2–76-4
Seismic accelerometers
displacement, 36-19
LVDT, 36-20
Newton's second law of motion, 36-18
potentiometer accelerometer, 36-20
schematic diagram, 36-17
strain-gage seismic instrument, 36-20–36-21
Semiconductor junction thermometers
applications for, 71-7–71-10
base–emitter voltage *vs.* ambient temperature, 71-2
base–emitter voltage *vs.* collector–emitter current, 71-3
bipolar transistor, 71-1–71-2
electronic components
Arrhenius equation, 71-12
junction temperature measurement, 71-12–71-13
mechanical stress, 71-11
integrated temperature sensors
analog-output sensor, 71-5–71-6
digital communication, 71-5

digital-output sensor, 71-5, 71-7
external circuitry, 71-7
linearity error *vs*. temperature, 71-2–71-3
sensor packages, 71-13
smart power IC, fault conditions, 71-10–71-11
thermal properties, 71-3–71-4
Semi-infinite surface temperature methods, 67-9–67-11
Sensor fusion, 2-8
Sensor nodes, 89-2
components
communication subsystem, 86-7–86-9
computing subsystem, 86-5–86-7
power subsystem, 86-9
sensing subsystem, 86-9–86-10
Dot mote, 86-4–86-5
Epic mote, 86-11
hardware, 86-4
list, 86-2–86-3
Mica mote, 86-4–86-5
Rene mote, 86-4–86-5
Rockwell WINS and Hidra nodes, 86-4–86-5
Shimmer2R mote, 86-10
smart dust sensor node, 86-4
SoC wireless sensor nodes, 86-11
Sun SPOT mote, 86-11
WeC mote, 86-4–86-5
Servoaccelerometer-type inclinometers, 21-4
Servo mount, 23-8–23-9
Servo-type accelerometer, 16-2–16-3
Servo-type inclinometers, 16-2–16-3
Sessile drop method, 47-9–47-10
Shaft runout, 23-8
Shear
dynamic viscosity, 46-2
strain, 44-3, 46-2
stress, 44-3, 46-2
thickening, 46-4
thinning, 46-4
Shimmer2R mote, 86-10
Ship's inertial navigational system (SINS), 17-12
Signal conditioning, 1-6
Signal generator
configuration circuit, 11-6–11-7
FILT, 11-5–11-6
safety characteristics, 11-5
Signal processing
feedback, 25-17
laser Doppler velocimetry (LDV), 64-9–64-11
laser interferometers, 27-12, 27-14
operational amplifiers, 25-14–25-15
oscillator circuits, 25-17
pulse width modulation, 25-15–25-16
square-wave linearization, 25-15–25-16
Signal-to-noise ratio (SNR), 2-7
digital signal processing, 64-11
Doppler shift, 64-4
imager passband, 78-14
nature of, 64-10
photodetector signal, 64-7
quality, 64-11
serial scanning, 78-5
trade-offs, 78-16–78-17
Silicon capacitive humidity sensor, 25-12
Silicon detectors
CCD detectors, 81-21–81-22
charged-particle spectroscopy, 81-22
diffused junction detector, 81-18
drift detectors, 81-21
energy resolution, 81-19–81-20
impurity requirements, 81-17
ion-implanted detector, 81-18–81-19
radiation damage, 81-22
spatial resolution, 81-20
strip detectors, 81-20
surface barrier detector, 81-18
temperature, 81-17
Silicon micromachined pressure sensor
construction of, 39-6
limitations, 39-7–39-9
wafer-to-wafer bonding, 39-7
Silicon thermistors, 69-14
Sine bar, 42-5–42-6
Sine plate, 42-6–42-7
Single air traffic control radar, 15-5–15-6
Single-crystal materials, 26-3–26-4
Single-mode homodyne
interferometer, 27-9–27-12
Single-reference junction circuit, 70-7–70-8,
70-15–70-16
Sleeve couplings, 23-9, 23-11
Sliding-vane type positive displacement meter,
53-2–53-3
Smart dust sensor node, 86-4
Smart sensors, 1-9
Smarttorque®, 44-7
Smart transducer interface module (STIM), 9-4
Sneaker net, 87-2
SoC wireless sensor nodes, 86-11
Software utility, 10-3
Sol–gel spin-casting deposition method, 73-2–73-3
Solid-state detectors
advantages, 82-5
categories, 82-16–82-17
front-end electronics, 82-20
manufacturers and prices, 82-17, 82-20
modern silicon processing technologies, 82-20
multichannel analyzer (MCA), 82-17–82-18
preamplifiers, 82-19
pulse height deficit, 82-19–82-20
signal generation, 82-15–82-16
silicon strip detector, 82-18–82-19
thin transmission-type (ΔE) detector, 82-18
Sonic digitizer, 19-5
Sound intensity measurement, 48-14–48-16

Sound level meter
definition, 48-9
function, 48-10–48-12
parts, 48-10–48-11
requirements, 48-10
types, 48-9
Southwest Research Institute® (SwRI®), 89-6
Spinning rotor gage, *see* Molecular drag gage (MDG)
Spirit thermometers, 74-4, 74-11
Spreadsheets, 8-5–8-6
Square-edged orifice plate, *see* Orifice plate
S-shaped force sensor
corner cube (CC) prism, 12-3
displacement, 12-2
experimental setup, calibration, 12-2–12-3
force measurement, 12-4–12-6
linear bearing force, 12-4
rms value of error, 12-6
second-order derivative, 12-6–12-7
strain-gage sensor, 12-2–12-3
velocity, 12-4
Stability, 43-14
Standard dispersion-shifted fiber (DSF), 33-11
Standard reference materials (SRMs), 7-4
Star sensors, 16-4–16-5
Static characteristics
bias, 4-9–4-10
drift, 4-6–4-7
error of nonlinearity, 4-10–4-11
hysteresis and backlash, 4-7–4-8
output/input relationship
amplification factor/gain, 4-5
dynamic range, 4-6
humidity sensor, 4-4
mathematical model, 4-5
static sensitivity, 4-5
transfer characteristic, 4-5–4-6
transfer function, 4-5
saturation, 4-7–4-9
Stefan–Boltzmann law, 72-3–72-4
Stem conduction, 63-4
Step-section planimetry, 20-2
Strain-gage accelerometers, 36-17
Strain-gage load cell
beam cell, 41-6–41-7
gage resistivity, 41-4–41-6
ring cell, 41-7
Strain-gage plethysmography, 20-4
Strain-gage seismic instrument, 36-20–36-21
Strain-gage sensor, *see* Piezoresistive pressure sensor
Strain measurement
definitions
characteristics of, 37-3–37-5
elastic material properties, 37-1–37-2
forces, 37-1–37-2
homogeneous, 37-1
Hooke's law, 37-3
isotropic, 37-1
manufacturers, 37-5–37-6
sources and prices, 37-5
temperature change, 37-2
tensile load, 37-1–37-2
types of, 37-3
uniaxial tension, 37-3
operating principle
birefringent film, 37-14–37-15
fiber-optic gages, 37-14
Moiré interference, 37-16–37-17
piezoelectric gages, 37-11–37-14
piezoresistive foil gages (*see* Piezoresistive foil/wire strain gages)
piezoresistive semiconductor gages, 37-11
Subcarrier modulation scheme
amplitude modulation, 90-6–90-7
CBW, 90-9–90-10
frequency modulation, 90-7–90-8
PBW, 90-8–90-9
Subsonic flow, 62-8
Substitution method, 43-11
Sun sensors, 16-4
Sun SPOT mote, 86-11
Supercapacitors, 88-8–88-9
Supersonic flow, 62-9
Supervisory control and data acquisition (SCADA)
components, 10-9–10-10
monitoring and controlling, 10-8
software, 10-10–10-12
typical network, 10-8, 10-10
Surface active agents, *see* Surfactants
Surface-mounted sensors, 68-8–68-9
Surface tension measurement
capillary rise method, 47-6–47-7
commercially available instruments, 47-6
contact angles and wetting, 47-2
definition, 47-12
description, 47-1
drop weight/volume method, 47-10–47-11
du Noüy ring method, 47-8
dynamic surface tension, 47-11
at extreme temperatures and pressures, 47-12
interfacial tension, 47-12
manufacturers and suppliers of instruments, 47-6
maximum bubble pressure method, 47-8–47-9
mechanical model, 47-1–47-2
mechanics of fluid
axisymmetric surfaces, 47-4–47-5
curved surfaces, 47-3–47-4
forces acting on contact line, 47-2–47-3
pendant drop and sessile drop method, 47-9–47-10
spinning drop method, 47-11
surface energy, 47-2
surface viscoelasticity, 47-11–47-12
surfactants, 47-2
Wilhelmy plate method, 47-7–47-8

Surfactants, 47-2
Swirl-type axial turbine meters, 54-9
Synchros, 24-8–24-9; *see also* Resolver

T

Tablet digitizer, 19-5, 19-9
Tactile sensor
  array sensor, 41-15–41-16
  requirements, 41-14–41-15
Tangential force, 44-2
Tantalum capacitive humidity sensor, 25-12
Taper, 42-7
Target mirror, 27-7
TDM, *see* Time division multiplexing (TDM)
Technical advisory groups (TAGs), 6-5
Telematics hub, 87-2
Telemetry
  aeronautical telemetry, 90-3–90-4
  baseband telemetry systems, 90-2
  defense and aerospace applications, 90-1
  definition, 90-1
  electric telemetry methods, 90-1
  frequency allocations, 90-3–90-4
  IRIG standard, 90-3
  medical applications, 90-3
  multiple-channel telemetry
    FDM (*see* Frequency division multiplexing)
    TDM (*see* Time division multiplexing)
  space applications, 90-2
  transmission channel, 90-4
  wireless telemetry, 90-1–90-3
  wire telemetry, 90-1–90-2
Temperature indicator
  color-change indicator
    characteristics, 76-12
    crayons, 76-13
    irreversible paints, 76-12–76-13
    liquid crystals, 76-14–76-16
    operating principle, 76-11
    reversible paints, 76-14
  shape/size deformation
    Bullers ring, 76-5–76-8
    crayons, 76-9–76-10
    liquids, 76-9–76-10
    pellets, 76-6, 76-8–76-9
    Seger cones, 76-2–76-4
    temperature monitors, 76-10–76-11
    thermoscope bars, 76-4–76-5
Temperature-modulated differential scanning calorimeter (TM-DSC), 79-7–79-8
Temperature scale
  Carnot cycle, 65-1–65-2
  empirical temperature scale
    calibration, 65-6
    fixed point, 65-8
    interpolating instrument, 65-8
    nonuniqueness, 65-8
    semi-empirical scales, 65-6–65-7
    smoothness, 65-9
  international temperature scales, 65-9–65-11
  linear scale, 65-3
  logarithmic natural scale, 65-2–65-3
  thermodynamic temperature scale, 65-4–65-7
  thermometers, 65-14–65-16
  T-T_{90} functions, 65-10, 65-12–65-14
  unit intervals, 65-3
Temperature sensor
  application, 77-10
  distributed and quasi-distributed optical fiber
    Bragg gratings, 77-9–77-10
    Brillouin scatter, 77-8–77-9
    core/cladding indices, 77-8
    heat and cold alarm system, 77-8
    Kerr effect, 77-9
    mixed pump and probe signal, 77-9
    principle, 77-7
    Raman distributed temperature probe, 77-7
    simple step index fiber, 77-8
    spatial resolutions, 77-7
  dyes, 77-4–77-5
  Fabry–Perot probe, 77-5
  fluorescence decay time measurements, 77-4
  microresonator, 77-6
  nanoprobes, 77-5–77-6
  passive interference immune measurement, 77-6
  probe principles, 77-4
  refractometry and interferometry, 77-5
  techniques for, 77-3
Thermal anemometer
  advantages and disadvantages, 59-1
  categories, 59-2
  commercial anemometers, 59-2–59-4
  first law of thermodynamics
    calibration constants, 59-7
    constant-temperature anemometer, 59-8–59-10
    film temperature, 59-7
    generalized expression, 59-11
    heat transfer, 59-6
    microprocessor-based digital thermal anemometer, 59-8, 59-10
    PC-based digital thermal anemometer system, 59-8–59-9
    point velocity determination, 59-5
    surface temperature, 59-6
  flow-calibration, 59-11–59-13
  instrument description, 59-16
  manufacturers, 59-16–59-17
  operation principle, 59-1–59-2
  point velocity, 59-13–59-14
  research anemometers, 59-3–59-5
  temperature compensation, 59-11
  turbulence measurements, 59-14–59-16

Thermal conductivity
fundamental equations, 66-2–66-3
heated-needle probe, 66-9
hot-disk instrument, 66-8–66-9
laser-flash instrument, 66-10–66-11
parallel-plate instrument
fluid measurement, 66-4–66-5
schematic diagram, 66-3
solid measurement, 66-5–66-6
steady-state methods, 66-3–66-4
suppliers, 66-11
transient hot bridge, 66-9
transient hot-wire technique, 66-6–66-8, 66-11
transient methods, 66-5
Thermal detectors, 78-7
bidirectional radiation flow, 72-11–72-12
bolometer, 72-13
operating ranges, 72-11
output electric signal, 72-12
radiative heat exchange, 72-11
reference temperature sensor, 72-13
temperature, 72-12
thermal imaging camera, 72-14
thermopile IR sensor, 72-13–72-14
Thermal dispersion mass flow meters
description
flow sensor, 63-2–63-4
gas temperature sensor, 63-4–63-5
in-line and insertion, 63-2
insertion meters, 63-3
RTD, 63-3–63-4
skin resistance, 63-3–63-4
stem conduction, 63-4
transmitter, 63-2, 63-5
transport property, 63-5
velocity sensor, 63-3–63-4
dry-sensor microprocessor-based system, 63-3, 63-11–63-12
flow meter selection, installation, and flow calibration, 63-12–63-14
operating principle
convection coefficient, 63-8
correlation, 63-9
dry velocity sensor, 63-3, 63-5–63-6
first law of thermodynamics, 63-7
gas law equation, 63-9–63-10
gas temperature, 63-6–63-7
mass velocity, 63-10
microprocessor-based system, 63-5, 63-7
output variables, 63-10–63-11
pipe/flow conduit, 63-10
potted velocity, 63-3, 63-8
quadratic formula, 63-6
Reynolds number, 63-9
skin resistance, 63-8
Thermal hysteresis, 68-15–68-16
Thermal imaging
atmospheric transmission, 78-3, 78-5
components, 78-1–78-2
detectors
cooling methods, 78-9
performance measures, 78-7–78-8
photon, 78-5–78-6
thermal, 78-7
electronics, 78-9–78-10
emission source, 78-4
fourier transform spectroscopy, 78-18
frame-sequential filtering, 78-18
MRTD, 78-14
optics and scanning
design, 78-11–78-12
field of view (FOV), 78-10–78-11
linear array, 78-11–78-12
numerical aperture (NA), 78-11
optical parameters, 78-11
PSF/MTF, 78-12–78-13
staring arrays, 78-10–78-11
SNR/NETD, 78-14
temperature, 78-13–78-14
thermal imagers, 78-15
trade-offs, 78-16–78-17
wave band, 78-2–78-3
Thermal radiation, 73-6–73-8
Thermistor gage, 40-8
Thermistor thermometers
β-values, 69-3
calibration equations, 69-4–69-5
epoxy-encapsulated thermistors, 69-10
errors
insulation resistance, 69-10
lead resistances, 69-9–69-10
offset voltages, 69-10
resistor instability, 69-9
self-heating error, 69-7–69-8
settling time, 69-8
stray thermal effect, 69-8
thermistor instability, 69-9
manufacturing process, 69-10
mobile charge carriers, 69-2
narrow-range thermistor thermometer, 69-5
nonlinearity, 69-6
NTC thermistors, 69-1, 69-11–69-13
PTC thermistors, 69-13–69-14
resistance–temperature characteristics, 69-2
sensitivities, 69-3
silicon thermistors, 69-14
temperature range, 69-10
three-element linear thermistor network, 69-6–69-7
zero output voltage condition, 69-5

Thermocouple gage, 40-8–40-9
Thermocouple thermometry
arbitrary thermoelement materials, 70-23
color codes, 70-17, 70-19
corresponding-type extensions, 70-7
definition, 70-2
differential thermocouple thermometer, 70-8–70-9
dual-reference junction circuit, 70-5–70-6, 70-13–70-15
electromagnetic noise, 70-39
electrostatic noise, 70-39
elementary thermocouple thermometry, 70-2
failure modes, 70-41
filtering, 70-40–70-41
grounding and shielding, 70-39–70-40
hardware
bar stock, 70-23
cable, 70-24
connectors, 70-25
contact junctions, 70-25–70-26
enclosed junctions, 70-26–70-27
foil and film, 70-23
fused junctions, 70-26
insulation, 70-24–70-25
intrinsic thermocouples, 70-26
junction assemblies, 70-27–70-28
material sources, 70-23
MIMS material, 70-24
quality assurance, 70-23
twisted junctions, 70-26
wire, 70-23–70-24
information sources, 70-41–70-43
instrumentation systems
computer-based systems, 70-36
CoMux, 70-36–70-37
data acquisition systems, 70-36
indicators and monitors, 70-35
modular components, 70-36
recorders, 70-35
signal transmission, 70-34–70-35
junction, 70-2
law of Seebeck emf, 70-10–70-11
letter-designated types, 70-22
mixed-type extensions, 70-7
paralleled thermoelements, 70-16–70-17
passive-type extensions, 70-6–70-7
quality assurance
active tests, 70-29–70-31
calibration, 70-31–70-34
magnetic identification, 70-28
resistive identification, 70-28
thermal identification, 70-29
visual identification, 70-28
reference temperature compensation, 70-6
Seebeck cell, 70-9–70-10
Seebeck's circuit, 70-5
Seebeck thermoelectric effect, 70-3–70-5
selection, 70-20–70-21
sensitivity, 70-2
single-reference junction circuit, 70-7–70-8, 70-15–70-16
standardized types, 70-17–70-18
temperature ranges, 70-37–70-38
thermal coupling, 70-38
thermocouple states, 70-2
thermoelectric schematics, 70-5
tolerances, 70-22
transient thermometry, 70-38–70-39
T/X sketch, 70-10–70-13
Thermodilution, 20-12
Thermodynamic temperature scale, 65-4–65-7
Thermogravimetry (TG), 79-11
Thermoluminescence (TL) dosimetry
activators, 84-3
instrumentation, 84-5
laser heating, 84-5–84-6
phosphors, 84-3–84-5
PTTL, 84-6–84-7
RPL, 84-7
signal-to-noise ratio, 84-6
traps and luminescence centers, 84-3
vendors, 84-6–84-7
Thermopile heat flux sensor, 67-5–67-6
Thermopower, 70-5
Thermoscope bars, 76-4–76-5
Thickness measurement
beta-backscatter, 13-7–13-8
caliper precision, 13-3
capacitive gaging, 13-5
comparators, 13-5
concerning capacitive sensors, 13-7
destructive techniques, 13-8
dial indicators, 13-3, 13-5
electronic gages, 13-5
ellipsometry, 13-7
gamma-ray backscatter, 13-7–13-8
graduated spacing, 13-2
hall effect gage, 13-6
inductive gaging (eddy-current sensing), 13-5–13-6
infrared absorption, 13-7–13-8
instruments, 13-2, 13-4
interferometry, 13-7
laser calipers, 13-5
lidar, 13-6
line-graduated instrument, 13-2
magnetic induction, 13-6
measuring techniques, 13-1–13-2
microscopes, 13-5
nanoscale scanning probe microscopy, 13-8–13-9
pneumatic gages, 13-5
radar, 13-6
THz technology, 13-8
ultrasound, 13-6–13-7

vendor addresses, 13-2–13-3
weighing, 13-5
x-ray fluorescence, 13-7–13-8
Thixotropic fluids, 46-4
Tilt measurement
geodetic/differential leveling, 21-4–21-5
hydrostatic leveling, 21-5–21-6
suspended and inverted plumb lines, 21-7–21-9
tiltmeters/inclinometers
angular resolution, 21-2
borehole inclinometers, 21-4
geodetic/geophysical tiltmeters, 21-2
ground subsidence, 21-3
horizontal pendulum tiltmeters, 21-2
liquid-bubble type, 21-2
liquid/pendulum tiltmeters, 21-2
mercury tiltmeter, 21-2
Niveltronic tiltmeter, 21-2
operation principle, 21-1
servoaccelerometer-type inclinometers, 21-4
temperature changes, 21-2–21-3
Torpedo-shaped biaxial inclinometer, 21-4
Tiltmeters/inclinometers
angular resolution, 21-2
borehole inclinometers, 21-4
geodetic/geophysical tiltmeters, 21-2
ground subsidence, 21-3
horizontal pendulum tiltmeters, 21-2
liquid-bubble type, 21-2
liquid/pendulum tiltmeters, 21-2
mercury tiltmeter, 21-2
Niveltronic tiltmeter, 21-2
operation principle, 21-1
servoaccelerometer-type inclinometers, 21-4
temperature changes, 21-2–21-3
Torpedo-shaped biaxial inclinometer, 21-4
Time division multiplexing (TDM)
advantages, 90-10
block diagram, 90-10–90-11
cross talk, 90-10
frequency and time graphs, 90-2–90-3
PCM
basic elements, 90-13
bit error rate *vs.* SNR, 90-15–90-16
delta modulation, 90-13
differential pulse code modulation, 90-13
frame structure, 90-13–90-15
IRIG standard, 90-13–90-15
specifications, 90-13–90-14
subcarrier modulation schemes
block diagram, 90-10–90-11
characteristics, 90-11
PAM, 90-12
PDM, 90-12
PPM, 90-12
Time-gain control (TGC), 49-20
Time-of-flight-measuring electro-optical sensors, 22-13
Time-of-flight (TOF) technique, 14-3, 14-5
advantages and drawbacks, 29-6
impulse response, 29-6
industrial acoustic noise, 29-6
methods, 29-5–29-6
suppliers and vendors, 29-6–29-7
Torductor®-S, 44-7
Torpedo-shaped biaxial inclinometer, 21-4
TorqSense, 44-7
Torque measurement
angular acceleration, 44-2
angular displacement, 44-1
angular velocity, 44-2
equilibrium, 44-2
kinetic energy, 44-3
measurement apparatus, 44-4–44-5
modulus of rigidity, 44-3
moment arm, 44-2
nonzero resultant torque, 44-2
power, 44-3
reaction torque, 44-2
resultant effect, 44-2
shear strain, 44-3, 44-5–44-6
shear stress, 44-3, 44-7
tangential force, 44-2
transducer
costs and options, 44-10
electric considerations, 44-9–44-10
features, 44-7–44-8
mechanical considerations, 44-9
smarttorque®, 44-7
Torductor®-S, 44-7
TorqSense, 44-7
torque-flange, 44-7–44-8
torquetronic, 44-7
twist angle, 44-5–44-6
work, 44-3
Touchdown temperature, 76-3–76-4
Track-etch dosimetry, 84-10–84-11
Transducer Electronic Data Sheet (TEDS), 9-3–9-4
Transduction process, 4-1
Transfer function, 1-2
Transient hot bridge, 66-9
Transient hot-wire technique, 66-6–66-8, 66-11
Transient thermometry, 70-38–70-39
Transit time flowmeter
axial sensor type, 57-4
clamp-on systems, 57-5
closed conduits flow measurements (*see* Closed conduits flow measurements)
field-mountable transducers, 57-4
manufacturers/distributors, 57-14–57-15
open channel flow measurements (*see* Open channel flow measurements)

operation principles, 57-2–57-3
penstock leak detection, 57-13–57-15
radial sensor type, 57-4–57-5
ultrasonic transducer, 57-3–57-4
Transposition method, 43-12
Triangulation techniques, 14-4–14-5
Triglycine sulfate (TGS), 73-1–73-3
Triode gage, 40-11–40-12
Trirotor type positive displacement meter, 53-3
Tuning controllers
autotuning, 91-8
manual tuning, 91-7–91-8
self-tuning, 91-8
Tuning fork densitometers, 45-12–45-13
Turbine and vane flowmeters
angular momentum, 54-11
axial
aerodynamic lift, 54-3
applications, 54-1
calibration, 54-6
design and construction, 54-7–54-8
dual-rotor axial, 54-9–54-10
flat-bladed rotor, 54-3–54-4
fluid driving torque, 54-3–54-5
helical turbine meters, 54-8–54-9
insertion flowmeter, 54-10–54-11
installation, 54-7
maintenance, 54-7
momentum exchange approach, 54-3
performance characteristics, 54-2–54-3
propeller, 54-8
schematic longitudinal section, 54-1–54-2
swirl-type axial, 54-9
cylindrical rotor, 54-12
manufacturers, 54-12–54-15
multijet, 54-12
Twisted-nematic (TN) effect LCD, 96-13
Two-color pyrometer, 72-16–72-17
Two-frequency heterodyne interferometer, 27-8, 27-10

U

Ultrasonic flowmeters
cross correlation, 57-1
Doppler, 57-1
drift, 57-1
phase shift, 57-1
transit time flowmeter (*see* Transit time flowmeter)
Ultrasonic pulse detection system, 62-14
Ultrasonic ranging techniques, 14-6
Ultrasonic sensors, 22-14–22-15
Ultrasonic switch, 18-15
Ultrasonic transducer, 57-3–57-4
Ultrasound, 13-6–13-7
characteristics of, 29-2–29-4
imaging, 20-2
measurement (*see* Ultrasound measurement)
pulse-echo system, 29-1–29-2
time-of-flight technique
advantages and drawbacks, 29-6
impulse response, 29-6
industrial acoustic noise, 29-6
methods, 29-5–29-6
suppliers and vendors, 29-6–29-7
transducer, 29-4–29-5
Ultrasound measurement
acoustic diffraction
Fraunhofer diffraction, 49-25
Fresnel diffraction, 49-24
overview, 49-24
pressure function, 49-25–49-26
acoustic intensity and sound levels, 49-11–49-12
acoustic microscopy, 49-23
application, 49-1–49-2
array transducers, 49-6–49-7
attenuation, 49-8
description, 49-14
pulse-echo measurement, 49-15–49-16
transmission measurement, 49-14–49-15
B-mode imaging, 49-4
CW fields, 49-16–49-17
Doppler flow measurement, 49-21–49-22
flaws, voids, and debonds detection, 49-22
frequency, selection criteria for, 49-7
medical ultrasound, 49-23
porosity and grain size estimation, 49-22–49-23
pulse-echo measurement
block diagram, 49-4–49-5
display formats for, 49-20
range measurement, 49-21
reflection coefficient, 49-8
reflection/transmission detection, 49-7
signal processing parameters, 49-8
single-element transducers, 49-6
solid elastic property, 49-22
sound velocity, 49-8–49-9
testing and process control, 49-22
transducer
magnetostrictive materials, 49-19
piezoelectric materials, 49-18–49-19
specification, 49-19
transmission measurement, 49-5–49-6
travel time, 49-7
ultrasound parameters, 49-2–49-4
ultrasound product manufacturer, 49-23
volume flow measurement, 49-22
wave propagation
homogeneous media, 49-9–49-11
layers and boundaries, 49-12–49-14
User calibration, 8-2
User machine interface, 10-11

V

Vacuum measurement
  cathode ionization (*see* Cathode ionization gages)
  classification of, 40-1–40-2
  direct reading gages
    liquid wall, 40-2–40-6
    solid wall, 40-3–40-6
  indirect reading gages
    convection, 40-9–40-11
    heat transfer regimes, 40-6–40-7
    Pirani, 40-6–40-8
    thermistor, 40-8
    thermocouple, 40-8–40-9
  MDG, 40-16–40-18
  operating range, 40-1, 40-3
  partial pressure measurements and mass spectrometers, 40-18–40-19
  resonance, 40-15
Vapor pressure thermometer
  design of, 75-4–75-7
  equations, 75-2–75-4
  temperature range, 75-2
Variable area flowmeters
  defintion, 52-1
  flumes
    channel width, 52-7
    characteristics, 52-8
    flow rate analysis, 52-14
    maximum discharge, 52-7–52-8
  movable vane meter, 52-2, 52-4–52-5
  rotameter (*see* Rotameter)
  weir (*see* Weir)
Variance, 4-3
Vector network analyzers (VNAs), 8-9
Vee-Scan method, 30-3–30-4
Vehicular navigation, 17-10–17-11
Velocity interference system for any reflector (VISAR) system, 35-6–35-7
Velocity sensors, 36-2
Venturi tube, 51-6–51-7, 51-11
Vertical alignment (VA) LCD, 96-13
Vertical gyro (VG) analysis, 16-11–16-13
Vibrating cylinder densitometers, 45-11–45-12
Vibrating element densitometers, 45-9–45-10
Vibrating-object viscometer, 46-20
Vibrating string accelerometer, 36-21–36-22
Vibrating tube densitometers, 45-10–45-11
Vibrating-wire pressure sensor, 39-3
Vibrating-wire viscometer, 46-20
Villari effect, 31-6, 41-9
Virtual instrument (VI)
  architecture, 10-1–10-2
  SCADA
    components, 10-9–10-10
    monitoring and controlling, 10-8
    software, 10-10–10-12
    typical network, 10-8, 10-10
  software
    application driver (*see* Application software)
    DAQ, 10-2–10-3
    device driver (*see* Device driver software)
Viscoelasticity, 46-4
Viscosity, 46-1; *see also* Fluid viscosity measurement
Visible-light photon counter (VLPC), 82-14
Visual optics/scanning
  design, 78-11–78-12
  field of view (FOV), 78-10–78-11
  linear array, 78-11–78-12
  numerical aperture (NA), 78-11
  optical parameters, 78-11
  PSF/MTF, 78-12–78-13
  staring arrays, 78-10–78-11
Voltage-controlled oscillator (VCO), 32-9, 32-11
Voltage-to-frequency converter, 95-7
Volume measurement
  indicator dilution methods
    dynamic recording method, 20-13
    first-pass method, 20-12–20-13
    gas dilution, 20-13
    principle, 20-11
    radionuclide techniques, 20-12
    thermodilution, 20-12
    time–concentration curve, 20-11
  plethysmography
    acoustic plethysmography, 20-16
    air/water plethysmography, 20-3–20-4, 20-15
    impedance plethysmography, 20-5–20-6, 20-16
    inductive plethysmography, 20-6, 20-16
    strain-gage plethysmography, 20-4
  of prostate, 20-7–20-10
  techniques, applications, and equipment, 20-15
  water displacement volumetry, 20-13–20-15
Volumetric flow measurements, 62-10–62-11
Vortex-shedding flowmeter
  configuration, 58-11
  flow ranges, 58-1
  flow tube
    flowmeter body, 58-4–58-5
    sensors, 58-5–58-6
    shedder, 58-5
  installation, 58-10–58-11
  mass flow determination, 58-4
  multisensor, 58-12
  operation principles, 58-2–58-3
  selection, 58-9
  single sensor, 58-12
  standard volume determination, 58-4
  transmitter
    analog transmitters, 58-7–58-8
    digital signal, 58-7

intelligent transmitters, 58-8–58-9
4–20 mA signal, 58-6–58-7
pulse signal, 58-7

W

Wafer-level packaging (WLP) techniques, 85-9, 85-14–85-15
Water displacement volumetry, 20-2, 20-13–20-15
Weber's law, 96-2
WeC mote, 86-4–86-5
Weir
broad-crested, 52-5, 52-7
characteristics, 52-8
flow rate analysis, 52-11–52-14
full-width thin-plate, 52-5–52-6
thin-plate rectangular notch, 52-5–52-6
thin-plate V-notch, 52-5, 52-7
Welch–Satterthwaite approximation, 5-9
Wiedemann effect, 31-6, 41-14
Wiegand effect, 35-12–35-14
Wien's law, 72-3
Winston cone, 82-12–82-13
Wired networks, 1-10
Wireless electroencephalogram (EEG) system, 85-19–85-20
Wireless instrumentation
ASK, 85-8
biomedical applications
cardiovascular stents, 85-21–85-22
glucose measurement, 85-21
sensium TZ1030 sensor interface, 85-20–85-21
wireless EEG system, 85-19–85-20
body movements monitoring, 85-11
communication protocols, 85-5–85-7
design, 85-1–85-2
free-space model, 85-9
frequency bands, 85-9–85-10
FSK, 85-8
generic microsystem architecture, 85-10
implantable microsystems, 85-11, 85-13
measurement systems, 85-2–85-3
multiplexing operation, 85-3–85-5
operation frequency selection, 85-8
power consumption, 85-13–85-15
PSK, 85-8
sensor RF interface, 85-11–85-12
wireless sensors network
Bluetooth protocols, 85-17
IEEE 802.11 protocol, 85-17
IEEE 802.15.4 protocol, 85-17–85-18
quality of service (QoS), 85-17
schematic diagram, 85-16
ZigBee protocol, 85-18–85-19
WLP techniques, 85-9
Wireless networks, 1-10–1-11
Wireless sensor network (WSN)
Bluetooth protocols, 85-17
IEEE 802.11 protocol, 85-17
IEEE 802.15.4 protocol, 85-17–85-18
localization for landslide measurements, 89-3–89-5
NBS variant
borehole-accessed caves, 89-8–89-9
cave mapping, 89-5–89-8
culvert inspection, 89-9–89-10
Perch and Move program, 89-3–89-4
processors, 89-1
quality of service (QoS), 85-17
reliability, 89-1
schematic diagram, 85-16
sensor nodes (*see* Sensor nodes)
trade-offs and design considerations, 89-10–89-11
ZigBee protocol, 85-18–85-19
Wireless sensors, 1-9–1-10
Wirewound elements, 23-3

X

X-ray fluorescence, 13-7–13-8

Z

Zeeman effect, 27-4–27-5, 27-10