Section 16
Instruments and Controls

BY

O. MULLER-GIRARD *Consulting Engineer, Rochester, NY.*
GREGORY V. MURPHY *Process Control Consultant, DuPont Co.*
W. DAVID TETER *Professor, Department of Civil Engineering, College of Engineering, University of Delaware.*

16.1 INSTRUMENTS

by Otto Muller-Girard

REFERENCES: ASME publications: ''Instruments and Apparatus Supplement to Performance Test Codes (PTC 19.1–19.20)''; ''Fluid Meters, pt. II, Application.'' ASTM, ''Manual on the Use of Thermocouples in Temperature Measurement,'' STP 470B. ISA publications: ''Standards and Recommended Practices for Instrumentation and Controls,'' 11 ed. Spitzer, ''Flow Measurement.'' Preston-Thomas, The International Temperature Scale of 1990 (ITS-90), *Metrologia,* **27,** 3–10 (1990), Springer-Verlag. NIST Monograph 175, ''Temperature-Electromotive Force Reference Functions and Tables for the Letter-Designated Thermocouple Types Based on the ITS-90,'' Government Printing Office, April 1993. Schooley, (ed.), ''Temperature, Its Measurement and Control in Science and Industry,'' Vol. 6, Pts. 1 and 2, American Institute of Physics. Time and frequency services offered by the National Institute of Standards and Technology (NIST). Lombardi and Beehler, NIST, paper 37-93. Beckwith, et al., ''Mechanical Measurements,'' Addison-Wesley. Considine, ''Encyclopedia of Instrumentation and Control,'' Krieger reprint. Considine, ''Handbook of Applied Instrumentation,'' McGraw-Hill, Krieger reprint. Dally, et al., ''Instrumentation for Engineering Measurements,'' Wiley. Doebelin, ''Measurement Systems, Application and Design,'' McGraw-Hill. Erikson and Graber, Harris et al., ''Shock and Vibration Control Handbook,'' McGraw-Hill. Holman, ''Experimental Methods for Engineers,'' McGraw-Hill. Jones (ed.), ''Instrument Science and Technology, Vol. 1, Measurement of Pressure, Level, Flow and Temperature,'' Heyden. Lion, ''Instrumentation in Scientific Research, Electrical Input Transducers,'' McGraw-Hill. Sheingold, (ed.), ''Transducer Interfacing Handbook,'' Analog Devices, Inc. Norwood, MA. Snell, ''Nuclear Instruments and Their Uses,'' Wiley. Spink, ''Principles and Practice of Flow Meter Engineering,'' Foxboro Co. Stout, ''Basic Electrical Measurements,'' Prentice-Hall. *Periodicals: Instruments & Control Systems,* monthly, Chilton Co. *InTech,* monthly, ISA. *Measurements & Control,* bimonthly, Measurements and Data Corp., Pittsburgh. *Sensors,* monthly, Helmers Publishing. *Test & Measurement World,* Cahners.

INTRODUCTION TO MEASUREMENT

An **instrument,** as referred to in the following discussion, is a device for determining the value or magnitude of a quantity or variable. The variables of interest are those which help describe or define an object, system, or process. Thus, in a manufacturing operation, product quality is related to measurements of its various dimensions and physical properties such as hardness and surface finish. In an industrial process, measurement and control of temperature, pressure, flow rates, etc., determine quality and efficiency of production.

Measurements may be direct, e.g., using a micrometer to measure a dimension, or indirect, e.g., determining moisture in steam by measuring the temperature in a throttling calorimeter.

Because of physical limitations of the measuring device and the system under study, practical measurements always have some error. The **accuracy** of an instrument is the closeness with which its reading approaches the true value of the variable being measured. Accuracy is commonly expressed as a percentage of measurement span, measurement value, or full-scale value. Span is the difference between the full-scale and the zero scale value. **Uncertainty,** the sum of the errors at work to make the measured value different from the true value, is the accuracy of measurement standards. Uncertainty is expressed in parts per million (ppm) of a measurement value. **Precision** refers to the reproducibility of the measurements, i.e., with a fixed value of the variable, how much successive readings differ from one another. **Sensitivity** is the ratio of output signal or response of the instrument to a change in input or measured variable. **Resolution** relates to the smallest change in measured value to which the instrument will respond.

Error may be classified as systematic or random. Systematic errors are those due to assignable causes. These may be static or dynamic. Static errors are caused by limitations of the measuring device or the physical laws governing its behavior. Dynamic errors are caused by the instrument not responding fast enough to follow the changes in measured variable. Random errors are those due to causes which cannot be directly established because of random variations in the system.

Standards for measurement are established by the National Institute of Standards and Technology. Secondary standards are prepared by very precise comparison with these primary standards and, in turn, form the basis for calibrating instruments in use. A well-known example is the use of precision gage blocks for the calibration of measuring instruments and machine tools.

There are three essential parts to an instrument: the **sensing element,** the **transmitting means,** and the **output** or **indicating element**. The sensing element responds directly to the measured quantity, producing a related motion, pressure, or electrical signal. This is transmitted by linkage, tubing, wiring, etc., to a device for display, recording, and/or control. Displays include motion of a pointer or pen on a calibrated scale, chart, oscilloscope screen, or direct numerical indication. Recording forms include writing on a chart and storage on magnetic tape or disk. The instrument may be actuated by mechanical, hydraulic, pneumatic, electrical, optical, or other energy medium. Often a combination of several energy modes is employed to obtain the accuracy, sensitivity, or form of output desired.

The transmission of measurements to distant indicators and controls is industrially accomplished by using the standardized electrical current signal of 4 to 20 mA; 4 mA represents the zero scale value and 20 mA the full-scale value of the measurement range. A pressure of 3 to 15 lb/in^2 is commonly used for pneumatic transmission of signals.

COUNTING EVENTS

Event counters are used to measure the number of items passing on a conveyor line, the number of operations of a machine, etc. Coupled with time measurements, they yield measures of average rate or frequency. They find important application, therefore, in inventory control, production analysis, and in the sequencing control of automatic machines.

Choice of the proper counting device depends on the kind of events being counted, the necessary counting speed, and the disposition of the measurement; i.e., whether it is to be indicated remotely, used to actuate a machine, etc. Errors in the total count may be introduced by events being too close together or by too much nonuniformity in the items being counted.

The **mechanical counter** is shown in Fig. 16.1.1. Motion of the event being counted deflects the arm, which through an appropriate linkage advances the count register one unit. Alternatively, motion of the actu-

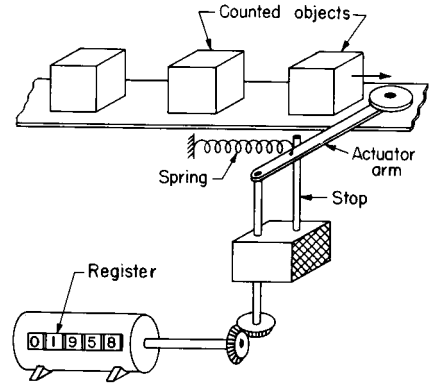


Fig. 16.1.1 Mechanical counter.

ating arm may close an electrical switch which energizes a relay coil to advance the count register one step.

Where there is a desire to avoid contact or close proximity with the object being counted, the photoelectric cell or diode, in conjunction with a lamp, a light-emitting diode (LED), or a laser light source, is employed in the transmitted or reflected light mode (Fig. 16.1.2). A signal to a counter is generated whenever the received light level is altered by the passing objects. Objects may be very small and very high counting speeds may be achieved with electronic counters.

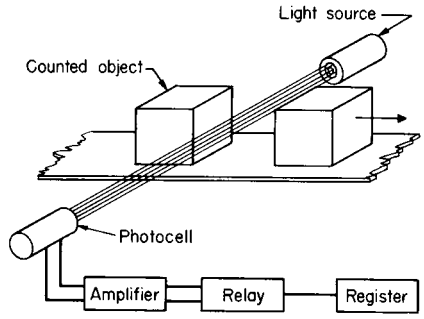


Fig. 16.1.2 Photoelectric counter.

Sensing methods based on electrical capacitance, magnetic, and eddy-current effects are extremely sensitive and fast acting, and are suitable for objects in close proximity to the sensor. The capacitive probe senses dielectrics other than air, such as glass and plastic parts. The magnetic pickup, by induction, responds to the motion of iron and nickel. The eddy-current sensor, by energy absorption, detects nonmagnetic conductors. All are suitable for counting machine operations.

The count is displayed by either a mechanical register as in Fig. 16.1.1, a dial-type register (as on the household watthour meter), or an electronic pulse counter with either number indicators or digital printing output. **Electronic counters** can operate accurately at rates exceeding 1 million counts per second.

TIME AND FREQUENCY MEASUREMENT

Measurement of time is basic to time and motion studies, time program controls, and the measurements of velocity, frequency, and flow rate. (See also Sec. 1.)

Mechanical clocks, chronometers, and stopwatches measure time in terms of the natural oscillation period of a system such as a pendulum, or hairspring balance-wheel combination. The minimum resolution is one-half period. Since this period is somewhat affected by temperature, precise timepieces employ a compensating element to maintain timing accuracies over long periods. Stopwatches may be obtained to read to better than 0.1 s. The major limitation, however, is in the response time of the user.

Electric timers are simple, inexpensive, and readily adaptable to remote-control operations. The majority of these are ac synchronous motors geared in the proper ratio to the indicator. These depend for their accuracy on the frequency of the line voltage. Consequently, care must be exercised in using such devices for precise short-time measurements.

Electronic timers are started and stopped by electrical pulses and hence are not limited by the observer's reaction time. They may be made extremely accurate and capable of measuring to less than 1 μs. These measure time by counting the number of cycles in a high-frequency signal generated internally by means of a quartz crystal. Stopwatch versions read at 0.01 s. Commercial instruments offer one or more functions: counting, measurement of frequency, period, and time intervals. Microprocessor-equipped versions increase versatility.

There are a variety of timing devices designed to indicate or control to a fixed time. These include timers based on the charging time of a condenser (e.g., type 555 integrated circuit), and the flow of oil or other fluid through a restriction.

Timing devices can be **calibrated** by comparison with a standard instrument or by reference to the National Institute of Standards and Technology timed radio signals, carrier frequencies and audio modulation of radio stations WWV and WWVB, Colorado, and WWVH, Hawaii. WWV and WWVH broadcast with carrier frequencies of 2.5, 5, 10, and 15 MHz. WWV also broadcasts on 20 MHz. Broadcasts provide second, minute, and hour marks with once-per-minute time announcements by voice and binary-coded decimal (BCD) signal on a 100-Hz subcarrier. Standard audio frequencies of 440, 500, and 600 Hz are provided. Station WWVB uses a 60-kHz carrier and provides second and minute marks and BCD time and date. Time services are also issued by NIST from geostationary satellites of the National Oceanographic and Atmospheric Administration (NOAA) on frequencies of 468.8375 MHz for the 75° west satellite and 468.825 MHz for the 105° west satellite. Automated Computer Time Service (ACTS) is available to 300- or 1200-baud modems via phone number 303-494-4774. (See also Sec. 1.2.)

Fast-moving, repetitive motions may be timed and studied by the use of stroboscopes which generate brilliant, very brief flashes of light at an adjustable rate.

The frequency of the observed motion is measured by adjusting the stroboscopic frequency until the system appears to stand still. The frequency of the motion is then equal to the stroboscope frequency or an integer multiple of it.

Many other means exist for **measuring vibrational or rotational frequencies.** These include timing a fixed number of rotations or oscillations of the moving member. Contact sensing can be done by an attached switch, or noncontact sensing can be done by magnetic or optical means. The pulses can be counted by an electronic counter or displayed on an oscilloscope or recorder and compared with a known frequency. Also used are reeds which vibrate when the measured oscillation excites their natural frequencies, flyball devices which respond directly to angular velocity, and generator-type tachometers which generate a voltage proportional to the speed.

MASS AND WEIGHT MEASUREMENT

Mass is the measure of the quantity of matter. The fundamental unit is the kilogram. The U.S. customary unit is the pound; 1 lb = 0.4536 kg (see Sec. 1.2, ''Measuring Units''). Weight is a measure of the force of gravity acting on a mass (see ''Units of Force and Mass'' in Sec. 4).

A general equation relating weight W and mass M is $W/g = M/g_c$, where g is the local acceleration of gravity, and $g_c = 32.174$ lbm · ft/(lbf) (s^2) [(1 kg · m/(N) (s^2)] is a property of the unit system. Then $W = Mg/g_c$. The specific weight w and the mass density p are related by $w = pg/g_c$. Masses are conveniently compared by comparing their weights, and masses are often loosely referred to as weights. Indeed, almost all practical measures of mass are based on weight.

Weighing devices fall into two major categories: balances and force-deflection systems. The device may be batch or continuous weighing, automatic or manual. Accuracies are expected to be of the order of 0.1 to better than 0.0001 percent, depending on the type and application of the scale. Calibration is normally performed by use of standard weights (masses) with calibrations traceable to the National Institute of Standards and Technology.

The **equal arm balance** compares the weight of an object with a set of standard weights. The laboratory balance shown in Fig. 16.1.3 is used for extreme precision and sensitivity. A chain poise provides fine adjustment of the final balance weight. The magnetic damper causes the balance to come to equilibrium quickly.

Large weighing scales operate on the same principle; however, the arms are unequal to allow multiplication between the tare and the measured weights. In this group are platform, track, hopper, and tank scales. Here balance is achieved by adjusting the position of one or more balance weights along a beam directly calibrated in weight units. In dial-

indicating-type scales, balance is achieved automatically through the deflection of calibrated pendulum weights from the vertical. The deflection is greatly magnified by the pointer-actuating mechanism, providing a direct-reading weight indication on the dial.

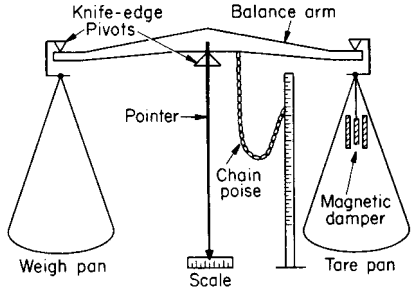


Fig. 16.1.3 Laboratory balance.

Since the deflection of a spring (within its design range) is directly proportional to the applied force, a calibrated spring serves as a simple and inexpensive weighing device. Applications include the **spring scale** and **torsion balance**. These are subject to hysteresis and temperature errors and are not used for precise work.

Other force-sensing elements are adaptable to weight measurement. Strain-gage load cells eliminate pivot maintenance and moving parts and provide an electrical output which can be used for direct recording and control purposes. Pneumatic pressure cells are also used with similar advantages.

In production processes, **continuous and automatic operating scales** are employed. In one type, the balancing weight is positioned by a reversible electric motor. Deflection of the beam makes an electrical contact which drives the motor in the proper direction to restore balance. The final balance position is translated by means of a potentiometer or digital encoding disk into a signal which is used for recording or control purposes.

The **batch-type scale** (Fig. 16.1.4) is adaptable to continuous flow streams of either liquids or solid particles. Material flows from the feed hopper through an adjustable gate into the scale hopper. When the weight in the scale hopper reaches that of the tare, the trip mechanism operates, closing the gate and opening the door. As soon as the scale hopper is empty, the weight of the tare forces the door closed again, resets the trip, and opens the gate to repeat the cycle. The agitator rotates

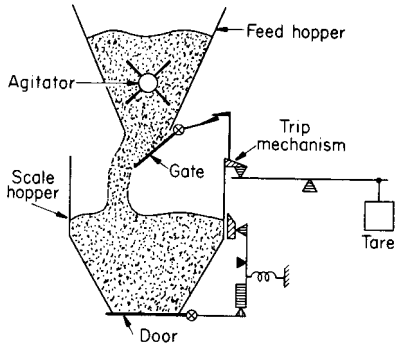


Fig. 16.1.4 Automatic batch-weighing scale.

while the gate is open, to prevent the solids from packing. Also, a ''dribble'' (partial closing of the gate just before the mechanism trips) is employed to minimize the error from the falling column of material at the instant balance is achieved. Since each dump of the scale represents a fixed weight, a counter yields the total weight of material passing through the scale.

In **continuous weighers,** a section of conveyor belt is balanced on a weigh beam (Fig. 16.1.5). The belt is driven at a constant speed; hence, if the total weight is held constant, the weight rate of material fed through the scale is fixed. Unbalance of the weight beam causes the rate of material flow onto the belt to be changed in the direction of restoring balance. This is accomplished by a mechanical adjustment of the feed gate or by varying the speed of a belt or screw feeder drive.

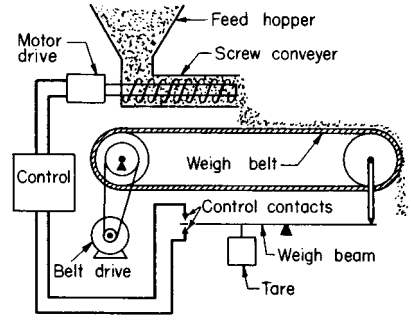


Fig. 16.1.5 Continuous-weighing scale.

If the density of the material is constant, **volume measurements** may be used to determine the mass. Thus, calibrated tanks are frequently used for liquids and vane and screw-type feeders for solids. Though often simpler to apply, these are not generally capable of as high accuracies as are common in weighing.

MEASUREMENT OF LINEAR AND ANGULAR DISPLACEMENT

Displacement-measuring devices are employed to measure dimension, distances between points, and some derived quantities such as velocity, area, etc. These devices fall into two major categories: those based on comparison with a known or reference length and those based on some fixed physical relationship.

The **measurement of angles** is closely related to displacement measurements, and indeed, one is often converted into the other in the process of measurement. The common unit is the degree, which represents $\frac{1}{360}$ of an entire rotation. The radian is used in mathematics and is related to the degree by π rad = 180°; 1 rad = 57.3°. The grad is an angle unit = $\frac{1}{400}$ rotation.

Figure 16.1.6 illustrates some methods of **rotary to linear conversion.** Figure 16.1.6*a* is a simple link and lever, Fig. 16.1.6*b* is a flexible link and sector, and Fig. 16.1.6*c* is a rack-and-pinion mechanism. These can be used to convert in either direction according to the relationship $D = RA/57.3$, where R = mean radius of the rotating element, in; D = displacement, in; and A = rotation, deg. (This equation holds for the link and lever of Fig. 16.1.6*a* only if the angle change from the perpendicular is small.)

Comparative devices are generally of the indicating type and include ruled or graduated devices such as the machinist's scale, folding rule, tape measure, digital caliper (Fig. 16.1.7), digital micrometer (Fig. 16.1.8), etc. These vary widely in their accuracy, resolution, and measuring span, according to their intended application. The manual readings depend for their accuracy on the skill and care of the operator.

The digital caliper and digital micrometer provide increased sensitivity and precision of reading. The stem of the digital caliper carries an embedded encoded distance scale. That scale is read by the slider. The distance so found shows on the digital display. The device is battery-operated and capable of displaying in inches or millimeters. Typical resolution is 0.0005 in or 0.01 mm.

The **digital micrometer,** employing rotation and translation to stretch the effective encoded scale length, provides resolution to 0.0001 in or 0.003 mm.

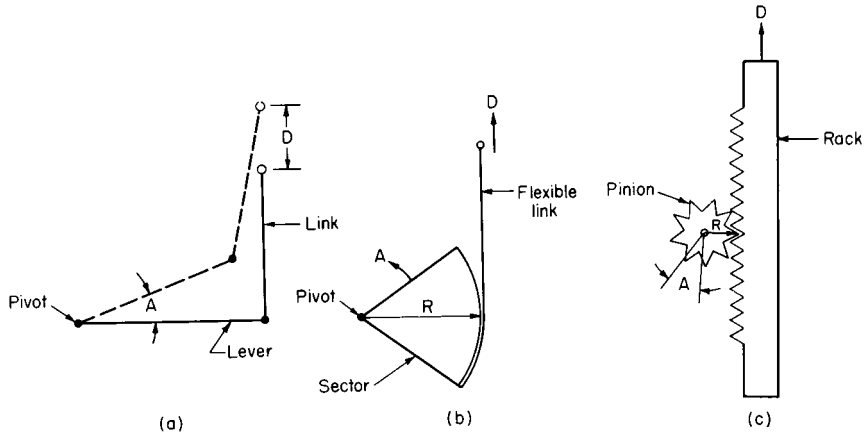


Fig. 16.1.6 Linear-rotary conversion mechanisms.

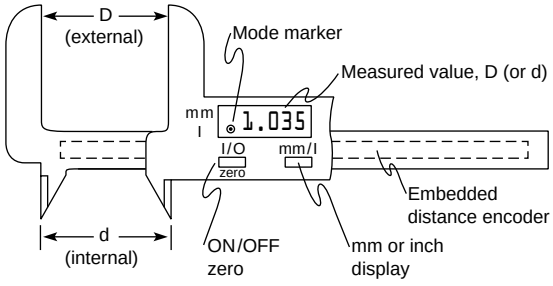


Fig. 16.1.7 Digital caliper.

Dial gages are also used to magnify motion. A rack and pinion (Fig. 16.1.6c) converts linear into rotary motion, and a pointer moves over a calibrated scale.

Various modifications of the above-mentioned devices are available for making special kinds of measurements; e.g., **depth gages** for measuring the depth of a hole or cavity, **inside and outside calipers** (Fig. 16.1.7) for measuring the internal and external dimensions respectively of an object, **protractors** for angular measurement, etc.

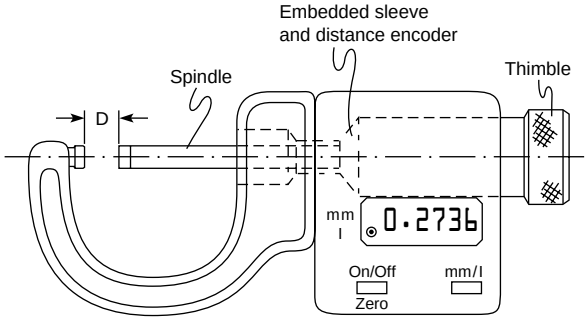


Fig. 16.1.8 Digital micrometer.

For line production and inspection work, **go no-go gages** provide a rapid and accurate means of dimension measurement and control. Since the measured values are fixed, the dependence on the operator's skill is considerably reduced. Such gages can be very complex in form to embrace a multidimensional object. They can also take the more general forms of the **feeler, wire, or thread-gage** sets. Of particular importance are **precision gage blocks,** which are used as standards for calibrating other measuring devices.

Displacement can be measured electrically through its effect on the resistance, inductance, reluctance, or capacitance of an appropriate sensing element.

The **potentiometer** is comparatively inexpensive, accurate, and flexible in application. It consists of a fixed linear resistance over which slides a rotating contact keyed to the input shaft (Fig. 16.1.9). The resistance or voltage (assuming constant voltage across terminals 1 and 3) measured across terminals 1 and 2 is directly proportional to the angle *A*. For straight-line motion, a mechanism of the type shown in Fig. 16.1.6 converts to rotary motion (or a rectilinear-type potentiometer can be used directly). (See also Sec. 15.) Versions with multiturns, straight-line motion, and special nonlinear resistance vs. motion are available.

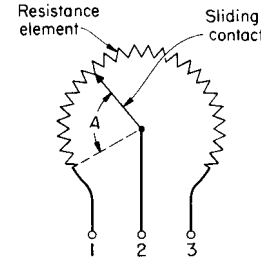


Fig. 16.1.9 Potentiometer.

The **synchro,** the **linear variable differential transformer (LVDT),** and the **E transformer** are devices in which the input motion changes the inductive coupling between primary and secondary coils. These avoid the limitations of wear, friction, and resolution of the potentiometer, but they require an ac supply and usually an electronic amplifier for the output. (See also Sec. 15.)

The **synchro** is a rotating device which is used to transmit rotary motions to a remote location for indication or control action. It is particularly useful where the rotation is continuous or covers a wide range. They are used in pairs, one transmitter and one receiver. For measurement of difference in angular position, the **control-transmitter** and **control-transformer** synchros generate an electrical error signal useful in control systems. A synchro differential added to the pair serves the same purpose as a gear differential.

The **linear variable differential transformer (LVDT)** consists of a primary and two secondary coils wound around a common core (Fig. 16.1.10). An armature (iron) is free to move vertically along the axis of the coils. An ac voltage is applied to the primary. A voltage is induced in each secondary coil proportional to the relative length of armature linking it with the primary. The secondaries are connected to oppose each other so that when the armature is centered, the output voltage is zero.

When the armature is displaced off center by an amount D, the output will be proportional to D (and phased to show whether D is above or below the center). These devices are very linear near the centered position, require negligible actuating force, and have spans ranging from 0.1 to several inches (0.25 cm to several centimeters).

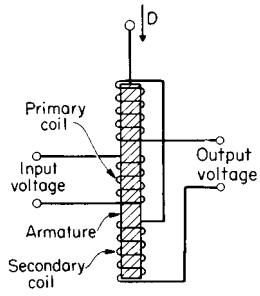


Fig. 16.1.10 Linear variable differential transformer (LVDT).

The **E transformer** is very similar to the above except that the coils are wound around a laminated iron core in the shape of an E (with the primary and secondaries occupying the center and outside legs respectively). The magnetic path is completed through an armature whose motion, either rotary or translational, varies the induced voltage in the secondaries, as in the device of Fig. 16.1.10. This, too, is sensitive to extremely small motions.

A method that is readily applied, if a strain-gage analyzer is handy, is to measure the deflection of a cantilever spring with strain gages bonded to its surface (see Strain Gages, Sec. 5).

The **change of capacitance** with the displacement of the capacitor plates is extremely sensitive and suitable to very small displacements or large rotation. Often, one plate is fixed within the instrument; the other is formed or rotated by the object being measured. The capacitance can be measured by an impedance bridge, by determining the resonant frequency of a tuned circuit or using a relaxation oscillator.

Many optical instruments are available for obtaining precise measurements. The **transit and level** are used in surveying for measuring angles and vertical distances (see Sec. 16.3). A telescope with fine cross hairs permits accurate sighting. The angle scales are generally equipped with verniers. The **measuring microscope** permits measurement of very small displacements and dimensions. The microscope table is equipped with micrometer screws for sensitive adjustment. In addition, templates of scales, angles, etc., are available to permit measurement by comparison. The **optical comparator** projects a magnified shadow image of an object on a screen where it can readily be compared with a reference template.

Light can be used as a standard for the measurement of distance, straightness, and related properties. The wavelength of light in a medium is the velocity of light in vacuum divided by the index of refraction n of the medium. For dry air $n - 1$ is closely proportional to air density and is about 0.000277 at 1 atm and 15°C for 550-nm green light. Since the wavelength changes about $+1$ ppm/°C, and about -0.36 ppm/mmHg, density gradients bend light slightly. A temperature gradient of 1°C/m (0.5°F/ft) will cause a deviation from a tangent line of about 0.05 mm (0.002 in) at 10 m (33 ft).

Optical equipment to establish and test alignment, plumb lines, squareness, and flatness includes **jig transits, alignment telescopes, collimators,** optical squares, mirrors, targets, and scales.

Interference principles can be used for distance measurements. An optical flat placed in close contact with a polished surface and illuminated perpendicular to the surface with a monochromatic light will show interference bands which are contours of constant separation distance between the surfaces. Adjacent bands correspond to separation differences of one-half wavelength. For 550-nm wavelength this is 275 nm (10.8 μin). This test is useful in examining surfaces for flatness and in length comparisons with gage blocks.

Laser beams can be used over great distances. Surveying instruments are available for measurements up to 40 mi (60 km). Accuracy is stated to be about 5 mm (0.02 ft) + 1 ppm. These instruments take several measurements which are processed automatically to display the distance directly. Momentary interruptions of the light beam can be tolerated.

A laser system for machine tools, measurement tables, and the like is available in modular form (Hewlett-Packard Co.). It can serve up to eight axes by using beam splitters with a combined range of 200 ft (60 m). Normal resolution of length is about one-fourth wavelength, with a digital display least count of 10 μin (0.1 μm). Angle-measurement display resolves 0.1 second of arc. Accuracy with proper environmental compensation is stated to be better than 1 ppm + 1 count in length measurement. Velocities up to 720 in/min (0.3 m/s) can be followed. Accessories are available for measuring straightness, parallelism, squareness and flatness, and for automatic temperature compensation. Various output options include displays and automatic computation and plots. The system can be used directly in measurement and control or to calibrate lead screws and other conventional measuring devices.

Pneumatic gaging finds an important place in line inspection and quality control. The device (Fig. 16.1.11) consists of a nozzle fixed in position relative to a stop or jig. Air at constant supply pressure passes through a restriction and discharges through the nozzle. The nozzle back pressure P depends on the gap G between the measured surface and the nozzle opening. If the measured dimension D increases, then G decreases, restricting the discharge of air, increasing P. Conversely, when D decreases, P decreases. Thus, the pressure gage indicates deviation of the dimension from some normal value. With proper design,

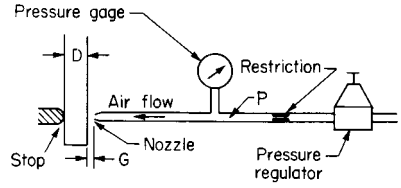


Fig. 16.1.11 Pneumatic gage.

this pressure is directly proportional to the deviation, limited, however, to a few thousandths of an inch span. The device is extremely sensitive [better than 0.0001 in (0.003 mm)], rugged, and, with periodic calibration against a standard, quite accurate. The gage is adaptable to automatic line operation where the pressure signal is recorded or used to actuate **"reject"** or **"accept"** controls. Further, any number of nozzles can be used in a jig to check a multiplicity of dimensions. In another form of this device, the flow of air is measured with a rotameter in place of the back pressure. The linear-variable differential transformer (LVDT) is also applicable.

The advent of automatically controlled machine tools has brought about the need for very accurate displacement measurement over a wide

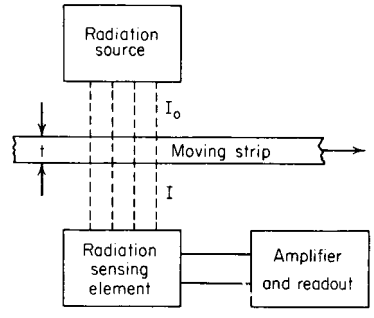


Fig. 16.1.12 Radiation-type thickness gage.

range. Most commonly applied for this purpose is the calibrated **lead screw** which measures linear displacement in terms of its angular rotation. **Digital** systems greatly extend the resolution and accuracy limitations of the lead screw. In these, a uniformly spaced optical or inductive grid is displaced relative to a sensing element. The number of grid lines counted is a direct measure of the displacement (see discussion of lasers above).

Measurement of strip thickness or coating thickness is achieved by **X-ray** or **beta-radiation-** type gages (Fig. 16.1.12). A constant radiation source (X-ray tube or radioisotope) provides an incident intensity I_0; the radiation intensity I after passing through the absorbing material is measured by an appropriate device (scintillation counter, Geiger-Müller tube, etc.). The thickness t is determined by the equation $I = I_0 e^{-kt}$, where k is a constant dependent on the material and the measuring device. The major advantage here is that measurements are continuous and nondestructive and require no contact. The method is extended to measure liquid level and density.

MEASUREMENT OF AREA

Area measurements are made for the purpose of determining surface area of an object or area inside a closed curve relating to some desired physical quantity. Dimensions are expressed as a length squared; e.g., in^2 or m^2. The areas of simple forms are readily obtained by formula. The area of a complex form can be determined by subdividing into simple forms of known area. In addition, various numerical methods are available (see **Simpson's rule,** Sec. 2) for estimating the area under irregular curves.

Area measuring devices include various mechanical, electrical, or electronic **flow integrators** (used with flowmeters) and the **polar planimeter**. The latter consists of two arms pivoted to each other. A tracer at the end of one arm is guided around the boundary curve of the area, causing rotation of a recorder wheel proportional to the area enclosed.

MEASUREMENT OF FLUID VOLUME

For a liquid of known density, volume is a quick and simple means of measuring the amount (or mass) of liquid present. Conversely, measuring the weight and volume of a given quantity of material permits calculation of its density. Volume has the dimensions of length cubed; e.g., cubic metres, cubic feet. The volume of simple forms can be obtained by formula.

A volumetric device is any container which has a known and fixed calibration of volume contained vs. the level of liquid. The device may be calibrated at only one point **(pipette, volumetric flask)** or may be graduated over its entire volume **(burette, graduated cylinder, volumetric tank)**. In the case of the tank, a sight glass may be calibrated directly in liquid volume.

Volumetric measure of continuous flow streams is obtained with the **displacement meter**. This is available in various forms: the nutating disk, reciprocating piston, rotating vane, etc. The **nutating-disk meter** (Fig. 16.1.13) is relatively inexpensive and hence is widely used (water meters, etc.). Liquid entering the meter causes the disk to nutate or ''roll'' as the liquid makes its way around the chamber to the outlet. A pin on the disk causes a counter to rotate, thereby counting the total number of rolls of the disk. Meter accuracy is limited by leakage past the disk and friction. The **piston meter** is like a piston pump operated backward. It is used for more precise measure (available to 0.1 percent accuracy).

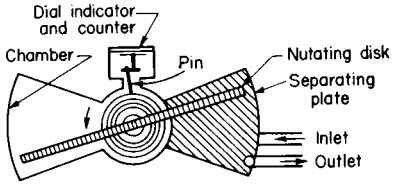


Fig. 16.1.13 Nutating-disk meter.

Volumetric gas measurement is commonly made with a **bellows meter**. Two bellows are alternately filled and exhausted with the gas. Motion of the bellows actuates a register to indicate the total flow. Various liquid-sealed displacement meters are also available for this purpose.

For precise volume measurements, corrections for temperature must be made (because of expansion of both the material being measured and the volumetric device). In the case of gases, the pressure also must be noted.

FORCE AND TORQUE MEASUREMENT

Force may be measured by the deflection of an elastic element, by balancing against a known force, by the acceleration produced in an object of known mass, or by its effects on the electrical or other properties of a stress-sensitive material. The common unit of force is the pound (newton). **Torque** is the product of a force and the perpendicular distance to the axis of rotation. Thus, torque tends to produce rotational motion and is expressed in units of pound feet (newton metres). Torque can be measured by the angular deflection of an elastic element or, where the moment arm is known, by any of the force measuring methods.

Since weight is the force of gravity acting on a mass, any of the weight-measuring devices already discussed can be used to measure force. Common methods employ the deflection of springs or cantilever beams.

The strain gage is an element whose electrical resistance changes with applied strain (see Sec. 5). Combined with an element of known force-strain, motion-strain, or other input-strain relationship it is a transducer for the corresponding input. The relation of gage-resistance change to input variable can be found by analysis and calibration. Measure of the resistance change can be translated into a measure of the force applied. The gage may be bonded or unbonded. In the bonded case, the gage is cemented to the surface of an elastic member and measures the strain of the member. Since the gage is very sensitive to temperature, the readings must be compensated. For this purpose, four gages are connected in a Wheatstone-bridge circuit such that the temperature effect cancels itself. A four-element unbonded gage is shown in Fig. 16.1.14. Note that as the applied force increases, the tension on two of the elements increases while that on the other two decreases. Gages subject to strain change of the same sign are put in opposite arms of the bridge. The zero adjustment permits balancing the bridge for zero output at any desired input. The e_1 and e_2 terminal pairs may be used interchangeably for the input excitation and the signal output.

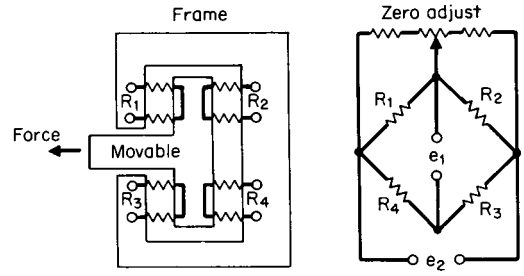


Fig. 16.1.14 Unbonded strain-gage board.

The **piezoelectric** effect is useful in measuring rapidly varying forces because of its high-frequency response and negligible displacement characteristics. Quartz rochelle salt, and barium titanate are common piezoelectric materials. They have the property of varying an output charge in direct proportion to the stress applied. This produces a voltage inversely proportional to the circuit capacitance. Charge leakage produces drifting at a rate depending on the circuit time constant. The

voltage must be measured with a device having a very high input resistance. Accuracy is limited because of temperature dependence and some hysteresis effect.

Forces may also be measured with any of the pressure devices described in the next section by balancing against a fluid pressure acting on a fixed area.

PRESSURE AND VACUUM MEASUREMENT

Pressure is defined as the force per unit area exerted by a fluid. Pressure devices normally measure with respect to atmospheric pressure (mean value = 14.7 lb/in^2), $p_a = p_g + 14.7$, where p_a = total or **absolute** pressure and p_g = **gage** pressure, both lb/in^2. Conventionally, gage pressure and vacuum refer to pressures above and below atmospheric, respectively. Common units are lb/in^2, in Hg, ftH_2O, kg/cm^2, bars, and mmHg. The mean SI atmosphere is 1.013 bar.

Pressure devices are based on (1) measure of an equivalent height of liquid column; (2) measure of the force exerted on a fixed area; (3) measure of some change in electrical or physical characteristics of the fluid.

The manometer measures pressure according to the relationship $p = wh = \rho gh/g_c$, where h = height of liquid of density ρ and specific weight w (assumed constants) supported by a pressure p. Thus, pressures are often expressed directly in terms of the equivalent height (head) of manometer liquid, e.g., inH_2O or inHg. Usual manometer fluids are water or mercury, although other fluids are available for special ranges.

The **U-tube manometer** (Fig. 16.1.15*a*) expresses the pressure difference $p_1 - p_2$ as the difference in levels h. If p_2 is exposed to the atmosphere, the manometer reads the gage pressure of p_1. If the p_2 tube is evacuated and sealed ($p_2 = 0$), the absolute value of p_1 is indicated. A common modification is the **well-type manometer** (Fig. 16.1.15*b*). The scale is specially calibrated to take into account changes of level inside the well so that only a single tube reading is required. In particular, Fig. 16.1.15*b* illustrates the form usually applied to measurement of atmospheric pressure **(mercury barometer)**.

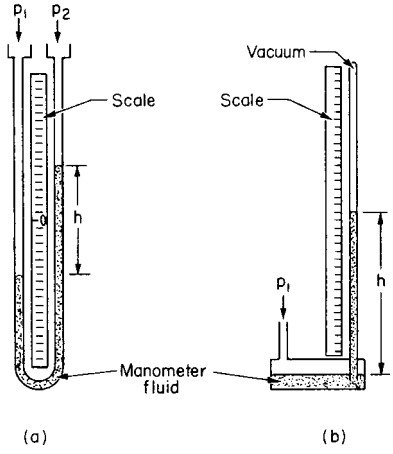


Fig. 16.1.15 Manometers. (*a*) U tube; (*b*) well type.

The sensitivity of readings can be increased by inclining the manometer tubes to the vertical **(inclined manometer),** by use of low-specific-gravity manometer fluids, or by application of optical-magnification or level-sensing devices. Accuracy is influenced by surface-tension effects (reading of the meniscus) and changes in fluid density (due to temperature changes and impurities).

By definition, pressure times the area acted upon equals the force exerted. The pressure may act on a diaphragm, bellows, or other element of fixed area. The force is then measured with any force-measuring device, e.g., spring deflection, strain gage, or weight balance. Very commonly, the unknown pressure is balanced against an air or hydraulic pressure, which in turn is measured with a gage. By use of unequal-area diaphragms, the pressure can thus be amplified or attenuated as required. Further, it permits isolating the process fluid which may be corrosive, viscous, etc.

The **Bourdon-tube gage** (Fig. 16.1.16) is the most commonly used pressure device. It consists of a flattened tube of spring bronze or steel bent into a circle. Pressure inside the tube tends to straighten it. Since one end of the tube is fixed to the pressure inlet, the other end moves proportionally to the pressure difference existing between the inside and outside of the tube. The motion rotates the pointer through a pinion-and-sector mechanism. For amplification of the motion, the tube may be bent through several turns to form spiral or helical elements as are used in pressure recorders.

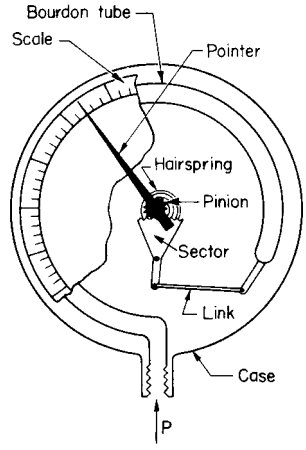


Fig. 16.1.16 Bourdon-tube gage.

In the **diaphragm gage,** the pressure acts on a diaphragm in opposition to a spring or other elastic member. The deflection of the diaphragm is therefore proportional to the pressure. Since the force increases with the area of the diaphragm, very small pressures can be measured by the use of large diaphragms. The diaphragm may be metallic (brass, stainless steel) for strength and corrosion resistance, or nonmetallic (leather, neoprene, silicon, rubber) for high sensitivity and large deflection. With a stiff diaphragm, the total motion must be very small to maintain linearity.

The **bellows gage** (Fig. 16.1.17) is somewhat similar to the diaphragm gage, with the advantage, however, of providing a much wider range of motion. The force acting on the bottom of the bellows is balanced by the deflection of the spring. This motion is transmitted to the output arm, which then actuates a pointer or recorder pen.

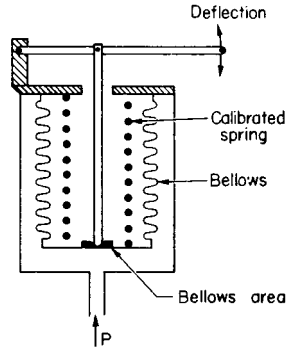


Fig. 16.1.17 Bellows gage.

The motion (or force) of the pressure element can be converted into an electrical signal by use of a differential transformer or strain-gage element or into an air-pressure signal through the action of a nozzle and pilot. The signal is then used for transmission, recording, or control.

The **dead-weight tester** is used as a standard for calibrating gages. Known hydraulic or gas pressures are generated by means of weights loaded on a calibrated piston. The useful range is from 5 to 5,000 lb/in^2 (0.3 to 350 bar). For low pressures, the water or mercury manometer serves as a reference.

For many applications (fluid flow, liquid level), it is important to measure the **difference between two pressures**. This can be done directly with the manometer. Other pressure devices are available as differential devices where (1) the case is made pressure-tight so that the second pressure can be applied external to the pressure element; (2) two identical pressure elements are mounted so that their outputs oppose each other.

Similar devices to those discussed are used to measure **vacuum,** the only difference being a shift in range or at most a relocation of the zeroing spring. When the vacuum is high (absolute pressure near zero) variations in atmospheric pressure become an important source of error. It is here that absolute-pressure devices are employed.

Any of the differential-pressure elements can be converted to an **absolute-pressure device** by sealing one pressure side to a perfect vacuum. A common instrument for the range 0 to 30 inHg employs two bellows of equal area set back to back. One bellows is completely evacuated and sealed; the other is connected to the measured pressure. The output is a bellows displacement, as in Fig. 16.1.17.

There are many instruments for high-vacuum work (0.001 to 10,000 μm range). These kinds of devices are based on the characteristic properties of gases at low pressures. The **McLeod gage** amplifies the pressure to be measured by compressing the gas a known amount and then measuring its pressure with a mercury manometer. The ratio of initial to final pressure is equal to the ratio of final to initial volume (for common gases). This gage serves as a standard for low pressures.

The **Pirani gage** (Fig. 16.1.18) is based on the change of heat conductivity of a gas with pressure and the change of electrical resistance of a wire with temperature. The wire is electrically heated with a constant current. Its temperature changes with pressure, producing a voltage across the bridge network. The compensating cell corrects for room-temperature changes.

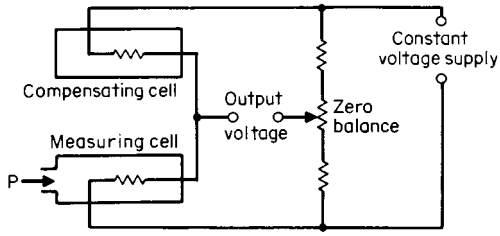


Fig. 16.1.18 Pirani gage.

The **thermocouple gage** is similar to the Pirani gage, except that a thermocouple is used to measure the temperature difference between the resistance elements in the measuring and compensating cells, respectively.

The **ionization gage** measures the ion current generated by bombardment of the molecules of the gas by the electron stream in a triode-type tube. This gage is limited to pressures below 1 μm. It is, however, extremely sensitive.

LIQUID-LEVEL MEASUREMENT

Level instruments are used for determining (or controlling) the height of liquid in a vessel or the location of the interface between two liquids of different specific gravity. In large storage tanks the level is indicated by a **calibrated tape or chain** which is attached to a float riding the liquid surface or by converting the signal reflection time of a radar or ultrasonic beam radiated onto the surface of the liquid into a level indication. For measuring small changes in level, the **fixed displacer** is common (Fig. 16.1.19). The buoyant force is proportional to the volume of displacer submerged and hence changes directly with the level. The force is balanced by the air pressure acting in the bellows, which in turn is generated by the flapper and nozzle. A pressure gage (or recorder) indicates the level.

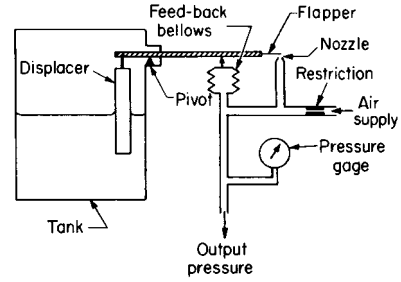


Fig. 16.1.19 Displacer-type level meter.

The level is often measured by means of a **differential-pressure meter** connected to taps in the top and bottom of the tank. As indicated in the discussion on manometers, the pressure difference is the height times the specific weight of the liquid. Where the liquid is corrosive or contains solids, then liquid seals, water purge, or air purge may be used to isolate the meter from the process.

For special applications, the dielectric, conducting, or absorption properties of the liquid can be used. Thus, in one model the liquid rises between two plates of a condenser, producing a **capacitance change** proportional to the change in level, and in another the **radiation** from a small radioactive source is measured. Since the liquid has a high absorption for the rays (compared with the vapor space), the intensity of the measured radiation decreases with the increase in level. An important advantage of this type is that it requires no external connections to the process.

TEMPERATURE MEASUREMENT

The common temperature scales (Fahrenheit and Celsius) are based on the freezing and boiling points of water (see Sec. 4 for discussion of temperature standards, units, and conversion equations).

Temperature is measured in a number of different ways. Some of the more useful are as follows.

1. *Thermal expansion of a gas* **(gas thermometer).** At constant volume, the pressure p of an (ideal) gas is directly proportional to its absolute temperature T. Thus, $p = (p_0/T_0)T$, where p_0 is the pressure at some known temperature T_0.

2. *Thermal expansion of a liquid or solid* **(mercury thermometer, bimetallic element).** Substances tend to expand with temperature. Thus, a change in temperature $t_2 - t_1$ causes a change in length $l_2 - l_1$ or a change in volume $V_2 - V_1$, according to the expressions.

$$l_2 - l_1 = a'(t_2 - t_1)l_1 \quad \text{or} \quad V_2 - V_1 = a'''(t_2 - t_1)V_1$$

where a' and $a''' =$ linear and volumetric coefficients of thermal expansion, respectively (see Sec. 4). For many substances, a' and a''' are reasonably constant over a limited temperature range. For solids, $a''' = 3a'$. For mercury at room temperature, a''' is approximately $0.00018°C^{-1}$ ($0.00010°F^{-1}$).

3. *Vapor pressure of a liquid* **(vapor-bulb thermometer).** The vapor pressure of all liquids increases with temperature. The Clapeyron equation permits calculation of the rate of change of vapor pressure with temperature.

4. *Thermoelectric potential* **(thermocouple).** When two dissimilar metals are brought into intimate contact, a voltage is developed which depends on the temperature of the junction and the particular metals

used. If two such junctions are connected in series with a voltage-measuring device, the measured voltage will be very nearly proportional to the temperature difference of the two junctions.

5. *Variation of electrical resistance* **(resistance thermometer, thermistor).** Electrical conductors experience a change in resistance with temperature which can be measured with a Wheatstone- or Mueller-bridge circuit, or a digital ohmmeter. The platinum resistance thermometer (PRT) can be very stable and is used as the temperature scale interpolation standard from − 160 to 660°C. Commercial resistance temperature detectors (RTD) using copper, nickel, and platinum conductors are in use and are characterized by a polynomial resistance-temperature relationship, such as

$$t = A + B \times R_t + C \times R_t^2 + D \times R_t^3 + E \times R_t^4$$

where R_t = resistance at prevailing temperature t in °C. A, B, C, D, and E are range- and material-dependent coefficients listed in Table 16.1.1. R_0, also shown in the table, is the base resistance at 0°C used in the identification of the sensor.

The thermistor has a large, negative temperature coefficient of resistance, typically − 3 to − 6 percent/°C, decreasing as temperature increases. The temperature-resistance relation is approximated (to perhaps 0.01° in range 0 to 100°C) by:

$$R_t = \exp\left(A_0 + A_1/t + \frac{A_2}{t^2} + \frac{A_3}{t^3}\right)$$

and

$$\frac{1}{t} = a_0 + a_1 \ln R_t + a_2(\ln R_t)^2 + a_3(\ln R_t)^3$$

with the constants chosen to fit four calibration points. Often a simpler form is given:

$$R = R_0 \exp\left\{\beta\left[\left(\frac{1}{t}\right) - \left(\frac{1}{t_0}\right)\right]\right\}$$

Typically β varies in the range of 3,000 to 5,000 K. The reference temperature t_0 is usually 298 K(= 25°C, 77°F), and R_0 is the resistance at that temperature. The error may be as small as 0.3°C in the range of 0° to 50°C. Thermistors are available in many forms and sizes for use from − 196 to + 450°C with various tolerances on interchangeability and matching. (See ''Catalog of Thermistors,'' Thermometrics, Inc.) The AD590 and AD592 integrated circuit (Analog Devices, Inc.) passes a current of 1 μA/°K very nearly proportional to absolute temperature. All these sensors are subject to self-heating error.

6. *Change in radiation* **(radiation and optical pyrometers).** A body radiates energy proportional to the fourth power of its absolute temperature. This principle is particularly adaptable to the measurement of very high temperatures where either the total quantity of radiation or its intensity within a narrow wavelength band may be measured. In the former type (radiation pyrometer), the radiation is focused on a heat-sensitive element, e.g., a thermocouple, and its rise in temperature is measured. In the latter type (optical pyrometer) the intensity of the radiation is compared optically with a heated filament. Either the filament brightness is varied by a control calibrated in temperature, or a fixed brightness filament is compared with the source viewed through a calibrated optical wedge.

The infrared thermometer accepts radiation from an object seen in a definite field of view, filters it to select a portion of the infrared spectrum, and focuses it on a sensor such as a blackened thermistor flake, which warms and changes resistance. Electronic amplification and signal processing produce a digital display of temperature. Correct calibration requires consideration of source emissivity, reflection, and transmission from other radiation sources, atmospheric absorption between the source object and the sensor, and compensation for temperature variation at the sensor's immediate surroundings.

Electrical nonconductors generally have fairly high (about 0.95) emissivities, while good conductors (especially smooth, reflective metal surfaces), do not; special calibration or surface conditioning is then needed. Very wide band (0.7 to 20 μm) instruments gather relatively large amounts of energy but include atmospheric absorption bands which reduce the energy received from a distance. The band 8 to 14 μm is substantially free from atmospheric absorption and is popular for general use with source objects in the range 32 to 1,000°F (0 to 540°C). Other bands and two-color instruments are used in some cases. See Bonkowski, Infrared Thermometry, *Measurements and Control*, Feb. 1984, pp. 152–162.

Fiber-optics probes extend the use of radiation methods to hard-to-reach places.

Important relationships used in the design of these instruments are the Wien and Stefan-Boltzmann laws (in modified form):

$$\lambda_m = k_1/T \qquad q = k_2 \varepsilon A(T_2^4 - T_1^4)$$

where λ_m = wavelength of maximum intensity, μm (nm); q = radiant energy flux, Btu/h (W); A = radiation surface, ft^2 (m^2); ε = mean emissivity of the surfaces; T_2, T_1 = absolute temperatures of radiating and receiving surfaces, respectively, °R (K); k_1 = 5215 μm. °R (2898 μm·K); k_2 = 0.173 × 10^{-8} Btu/(h·ft^2·°R^4) [5.73 × 10^{-8} W/(m^2·K^4)]. The emissivity depends on the material and form of the surfaces involved (see Sec. 4). Radiation sensors with scanning capability can produce maps, photographs, and television displays showing temperature-distribution patterns. They can operate with resolutions to under 1°C and at temperatures below room temperature.

7. *Change in physical or chemical state* **(Seger cones, Tempilsticks).** The temperatures at which substances melt or initiate chemical reaction are often known and reproducible characteristics. Commercial products are available which cover the temperature range from about 120 to 3600°F (50 to 2000°C) in intervals ranging from 3 to 70°F (2 to 40°C). The temperature-sensing element may be used as a solid which softens and changes shape at the critical temperature, or it may be applied as a paint, crayon, or stick-on label which changes color or surface appearance. For most the change is permanent; for some it is reversible. Liquid crystals are available in sheet and liquid form: these change reversibly through a range of colors over a relatively narrow temperature range. They are suitable for showing surface-temperature patterns in the range 20 to 50°C (68 to 122°F).

An often used temperature device is the **mercury-in-glass thermometer**. As the temperature increases, the mercury in the bulb expands and rises through a fine capillary in the graduated thermometer stem. Useful range extends from − 30 to 900°F (− 35 to 500°C). In many applications of the mercury thermometer, the stem is not exposed to the mea-

Table 16.1.1 Polynomial Coefficients for Resistance Temperature Detectors

Material of conductor	ID R_0, Ω	Useful range, °C	Polynomial coefficients					Typical accuracy,* °C
			A, °C	B, °C/Ω	C, °C/Ω^2	D, °C/Ω^3	E, °C/Ω^4	
Copper	9.042	− 70 to 0	− 225.64	23.30735	+ 0.246864	− 0.00715		1.5
10Ω @25°C		0 to 150	− 234.69	25.95508				1.5
Nickel	120	− 80 to 320	− 199.47	1.955336	− 0.00266	1.88E − 6		1
Platinum	100	− 200 to 0	− 241.86	2.213927	0.002867	− 9.8E − 6	1.64E − 8	1
DIN/IEC		0 to 850	− 236.06	2.215142	0.001455			0.5
α = 0.00385/°C								

* For higher accuracy consult the table or equation furnished by the manufacturer of the specific RTD being used. Temperatures per ITS-90, resistances per SI-90.

sured temperature; hence a correction is required (except where the thermometer has been calibrated for **partial immersion**). Recommended formula for the correction K to be added to the thermometer reading is $K = 0.00009\ D(t_1 - t_2)$, where D = number of degrees of exposed mercury filament, °F; t_1 = thermometer reading, °F; t_2 = the temperature at about middle of the exposed portion of stem, °F. For Celsius thermometers the constant 0.00009 becomes 0.00016.

For **industrial applications** the thermometer or other sensor is encased in a metal or ceramic protective well and case (Fig. 16.1.20). A threaded union fitting is provided so that the thermometer can be installed in a line or vessel under pressure. Ideally the sensor should have the same temperature as the fluid into which the well is inserted. However, heat

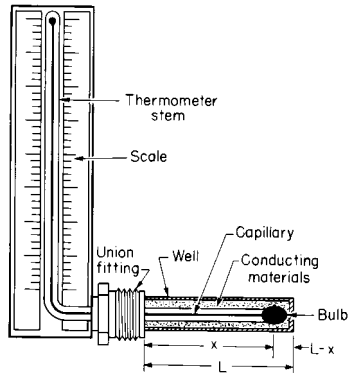


Fig. 16.1.20 Industrial thermometer.

conduction to or from the pipe or vessel wall and radiation heat transfer may also influence the sensor temperature (see ASME PTC 19.3-1974 Temperature Measurement, on well design). An approximation of the conduction error effect is

$$T_{sensor} - T_{fluid} = (T_{wall} - T_{fluid})E$$

For a sensor inserted to a distance $L - x$ from the tip of a well of insertion length L, $E = \cosh[m(L - x)]/\cosh mL$, where $m = (h/kt)^{0.5}$; x and L are in ft (m); h = fluid-to-well conductance, Btu/(h) (ft²)(°F) [J/(h) (m²)(°C)]; k = thermal conductivity of the well-wall material. Btu/(h)(ft)(°F) [J/(h)(m)(°C)]; and t = well-wall thickness, ft (m). Good thermal contact between the sensor and the well wall is assumed. For $(L - x)/L = 0.25$:

mL	1	2	3	4	5	6	7
E	0.67	0.30	0.13	0.057	0.025	0.012	0.005

Radiation effects can be reduced by a polished, low-emissivity surface on the well and by radiation shields around the well. Concern with mercury contamination has made the bimetal thermometer the most commonly used expansion-based temperature measuring device. Differential thermal expansion of a solid is employed in the simple **bimetal** (used in thermostats) and the bimetallic helix (Fig. 16.1.21). The bimetallic element is made by welding together two strips of metal having different coefficients of expansion. A change in temperature then causes the element to bend or twist an amount proportional to the temperature.

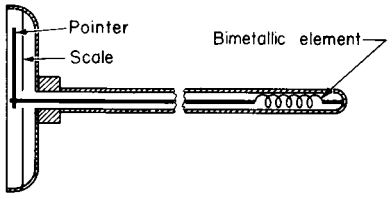


Fig. 16.1.21 Bimetallic temperature gage.

A common bimetallic pair consists of invar (iron-nickel alloy) and brass.

For control or alarm indications at fixed temperatures, thermometers may be equipped with electrical contacts such that when the temperature matches the contact point, an external relay circuit is energized.

A popular industrial-type instrument employs the deflection of a **pressure-spring** to indicate (or record) the temperature (Fig. 16.1.22). The sensing element is a metal bulb containing some specific gas or liquid. The bulb connects with the pressure spring (in the form of a spiral or helix) through a capillary tube which is usually enclosed in a

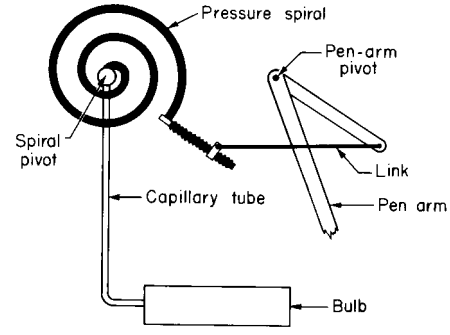


Fig. 16.1.22 Pressure-spring element.

protective sheath or armor. Increasing temperature causes the fluid in the bulb to expand in volume or increase in pressure. This forces the pressure spring to unwind and move the pen or pointer an appropriate distance upscale.

The **bulb fluid** may be mercury (mercury system), nitrogen under pressure (gas system), or a volatile liquid (vapor-pressure system). Mercury and gas systems have linear scales; however, they must be compensated to avoid ambient temperature errors. The capillary may range up to 200 ft in length with, however, considerable reduction in speed of response.

For transmitting temperature readings over any distance (up to 1,000 ft), the **pneumatic transmitter** (Fig. 16.1.23) is better suited than the methods outlined thus far. This instrument has the additional advantages of greater compactness, higher response speeds, and generally better accuracy. The bulb is filled with gas under pressure which acts on the diaphragm. An increase in bulb temperature increases the upward force acting on the main beam, tending to rotate it clockwise. This causes the baffle or flapper to move closer to the nozzle, increasing the nozzle back pressure. This acts on the pilot, producing an increase in output pressure, which increases the force exerted by the feedback bellows. The system returns to equilibrium when the increase in bellows pressure exactly balances the effect of the increased diaphragm pressure. Since the lever ratios are fixed, this results in a direct proportionality between bulb temperature and output air pressure. For precision,

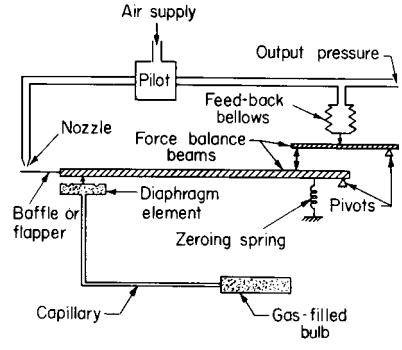


Fig. 16.1.23 Pneumatic temperature transmitter.

compensating elements are built into the instrument to correct for the effects of changes in barometric pressure and ambient temperature.

Electrical systems based on the thermocouple or resistance thermometer are particularly applicable where many different temperatures are to be measured, where transmission distances are large, or where high sensitivity and rapid response are required. The thermocouple is used with high temperatures; the resistance thermometer for low temperatures and high accuracy requirements.

The **choice of thermocouple** depends on the temperature range, desired accuracy, and the nature of the atmosphere to which it is to be exposed. The temperature-voltage relationships for the more common of these are given by the curves of Fig. 16.1.24. Table 16.1.2 gives the recommended temperature limits, for each kind of couple. Table 16.1.3 gives polynomials for converting thermocouple millivolts to temperature. The thermocouple voltage is measured by a digital or deflection millivoltmeter or null-balance type of potentiometer. Completion of the thermocouple circuit through the instrument immediately introduces one or more additional junctions. Common practice is to connect the thermocouple (hot junction) to the instrument with special lead wire (which may be of the same materials as the thermocouple itself). This assures that the **cold junction** will be inside the instrument case, where compensation can be effectively applied. Cold junction compensation is typically achieved by measuring the temperature of the thermocouple wire to copper wire junctions or terminals with a resistive or semiconductor thermometer and correcting the measured terminal voltage by a derived equivalent millivolt cold junction value. Figure 16.1.25 shows a digital temperature indicator with correction for different ANSI types of thermocouple voltage to temperature nonlinearities being stored in and applied to the analog-to-digital converter (A/D) by a read-only memory (ROM) chip.

The **resistance thermometer** employs the same circuitry as described above, with the resistance element (RTD) being placed external to the instrument and the cold junction being omitted (Fig. 16.1.26). Three types of RTD connections are in use: two wire, three wire, and four wire. The two-wire connection makes the measurement sensitive to lead

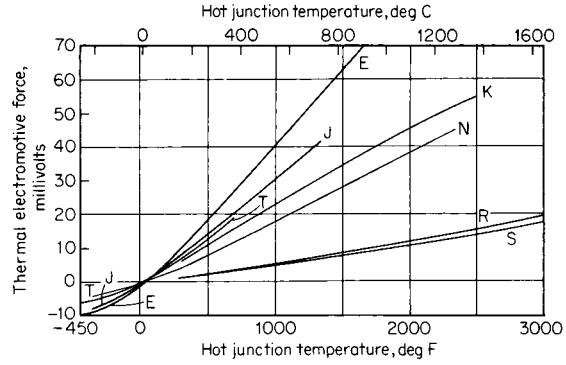


Fig. 16.1.24 Thermocouple voltage-temperature characteristics [reference junction at 32°F (0°C)].

Table 16.1.2 Limits of Error on Standard Wires without Selection*†

ANSI	Materials and polarities		°F: −150	−75	32	200	530	600	700	1,000	1,400	2,300	2,700
symbol‡	Positive	Negative	°C: −101	−59	0	93	277	316	371	538	760	1,260	1,482
T	Cu	Constantan§		2%	1.5°F (0.8°C)			¾%					
E	Ni-Cr	Constantan				3°F (1.7°C)				½%			
J	Fe	Constantan				4°F (2.2°C)			¾%				
K	Ni-Cr	Ni-Al				4°F (2.2°C)			¾%				
N	Ni-Cr-Si	Ni-SiMn				4°F (2.2°C)			¾%				
R	Pt-13% Rh	Pt				3°F (1.5°C)					¼%		
S	Pt-10% Rh	Pt				3°F (1.5°C)					¼%		

* Protect copper from oxidation above 600°F; iron above 900°F. Protect Ni-Al from reducing atmospheres. Protect platinum from nonreducing atmospheres. Type B (Pt-30% Rh versus Pt-6%) is used up to 3,200°F(1,700°C). Its standard error is ½ percent above 1,470°F (800°C).

† Closer tolerances are obtainable by selection and calibration. Consult makers' catalogs. Tungsten-rhenium alloys are in use up to 5,000°F (2,760°C). For cryogenic thermocouples see Sparks et al., Reference Tables for Low-Temperature Thermocouples. *Natl. Bur. Stand. Monogr.* 124.

‡ Individual wires are designated by the ANSI symbol followed by P or N; thus iron is JP.

§ Constantan is 55% Cu, 45% Ni. The nickel-chromium and nickel-aluminum alloys are available as Chromel and Alumel, trademarks of Hoskins Mfg. Co.

Table 16.1.3 Polynomial Coefficients for Converting Thermocouple emf to Temperature*

Range	Type E	Type J	Type K	Type N	Type S	Type T
mV	0 to 76.373	0 to 42.919	0 to 20.644	0 to 47.513	1.874 to 11.95	0 to 20.872
°C	0 to 1000°	0 to 760°	0 to 500°	0 to 1300°	250 to 1200°	0 to 400°
°F	32 to 1832°	32 to 1400°	32 to 932°	32 to 2372°	482 to 2192°	32 to 752°
α_0	0	0	0	0	1.291507177E + 01	0
α_1	1.7057035E + 01	1.978425E + 01	2.508355E + 01	3.8783277E + 01	1.466298863E + 02	2.592800E + 01
α_2	−2.3301759E − 01	−2.001204E − 01	7.860106E − 02	−1.1612344E + 00	−1.534713402E + 01	−7.602961E − 01
α_3	6.5435585E − 03	1.036969E − 02	−2.503131E − 01	6.9525655E − 02	3.145945973E + 00	4.637791E − 02
α_4	−7.3562749E − 05	−2.549687E − 04	8.315270E − 02	−3.0090077E − 03	−4.163257839E − 01	−2.165394E − 03
α_5	−1.7896001E − 06	3.585153E − 06	−1.228034E − 02	8.8311584E − 05	3.187963771E − 02	6.048144E − 05
α_6	8.4036165E − 08	−5.344285E − 08	9.804036E − 04	−1.6213839E − 06	−1.291637500E − 03	−7.293422E − 07
α_7	−1.3735879E − 09	5.099890E − 10	−4.413030E − 05	1.6693362E − 08	2.183475087E − 05	
α_8	1.0629823E − 11		1.057734E − 06	−7.3117540E − 11	−1.447379511E − 07	
α_9	−3.2447087E − 14		−1.052755E − 08		8.211272125E − 09	
Maximum deviation	± 0.02°C	± 0.04°C	− 0.05 to + 0.04°C	± 0.06°C	± 0.01°C	± 0.03°C

* $T(°C) = \sum_{i=0}^{n} a_i \times (mV)^i$

All temperatures are ITS-1990 and all voltages are SI-1990 values. Maximum deviation is that from the ITS-1990 tables; thermocouple wire error is additional. Computed temperature deviates greatly outside of given ranges. Consult source for thermocouple types B and R and for other millivolt ranges.

SOURCE: NIST Monograph 175, April 1993.

wire temperature changes. The three-wire connection, preferred in industrial applications, eliminates the lead wire effect provided the leads are of the same gage and length, and subject to the same environment. The four-wire arrangement makes no demands on the lead wires and is preferred for scientific measurements.

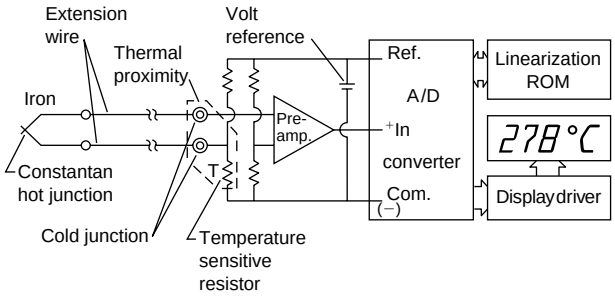


Fig. 16.1.25 Temperature measurement with thermocouple and digital millivoltmeter.

The resistance bulb consists of a copper or platinum wire coil sealed in a protective metal tube. The **thermistor** has a very large temperature coefficient of resistance and may be substituted in low-accuracy, low-cost applications.

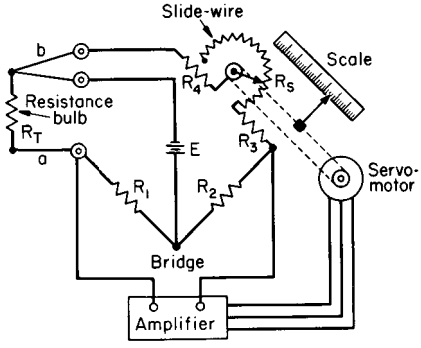


Fig. 16.1.26 Three-wire resistance thermometer with self-balancing potentiometer recorder.

By use of a **selector switch,** any number of temperatures may be measured with the same instrument. The switch connects in order each thermocouple (or resistance bulb) to the potentiometer (or bridge circuit) or digital voltmeter. When balance is achieved, the recorder prints the temperature value, then the switch advances on to the next position.

Optical pyrometers are applied to high-temperature measurement in the range 1000 to 5000°F (540 to 2760°C). One type is shown in Fig. 16.1.27. The surface whose temperature is to be measured (target) is focused by the lens onto the filament of a calibrated tungsten lamp. The light intensity of the filament is kept constant by maintaining a constant

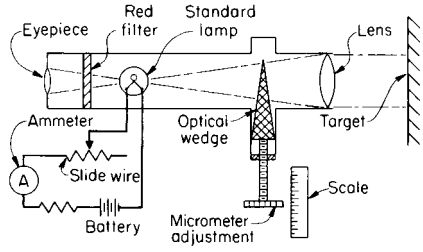


Fig. 16.1.27 Optical pyrometer.

current flow. The intensity of the target image is adjusted by positioning the optical wedge until the image intensity appears exactly equal to that of the filament. A scale attached to the wedge is calibrated directly in temperature. The red filter is employed so that the comparison is made at a specific wavelength (color) of light to make the calibration more reproducible. In another type of optical pyrometer, comparison is made by adjusting the current through the filament of the standard lamp. Here, an ammeter in series is calibrated to read temperature directly. **Automatic operation** may be had by comparing filament with image intensities with a pair of photoelectric cells arranged in a bridge network. A difference in intensity produces a voltage, which is amplified to drive the slide wire or optical wedge in the direction to restore zero difference.

The **radiation pyrometer** is normally applied to temperature measurements above 1000°F. Basically, there is no upper limit; however, the lower limit is determined by the sensitivity and cold-junction compensation of the instrument. It has been used down to almost room temperature. A common type of radiation receiver is shown in Fig. 16.1.28. A lens focuses the radiation onto a thermal sensing element. The temperature rise of this element depends on the total radiation received and the conduction of heat away from the element. The radiation relates to the temperature of the target; the conduction depends on the temperature of the pyrometer housing. In normal applications the latter factor is not very great; however, for improved accuracy a compensating coil is added to the circuit. The sensing element may be a thermopile, vacuum

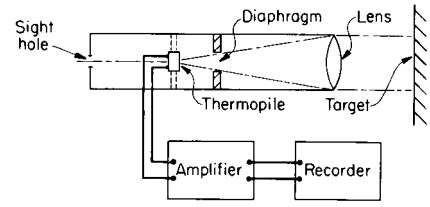


Fig. 16.1.28 Radiation pyrometer.

thermocouple, or bolometer. The **thermopile** consists of a number of thermocouples connected in series, arranged so that all the hot junctions lie in the field of the incoming radiation; all of the cold junctions are in thermal contact with the pyrometer housing so that they remain at ambient temperature. The **vacuum thermocouple** is a single thermocouple whose hot junction is enclosed in an evacuated glass envelope. The **bolometer** consists of a very thin strip of blackened nickel or platinum foil which responds to temperature in the same manner as the resistance thermometer. The **thermal sensing element** is connected to a potentiometer or bridge network of the same type as described for the self-balance thermocouple and resistance-thermometer instruments. Because of the nature of the radiation law, the scale is nonlinear.

Accuracy of the optical- and radiation-type pyrometers depends on:

1. *Emissivity of the surface being sighted on.* For closed furnace applications, blackbody conditions can be assumed (emissivity = 1). For other applications corrections for the actual emissivity of the surface must be made (correction tables are available for each pyrometer model). Multiple color or wavelength sensing is used to reduce sensitivity to hot object emissivity. For measuring hot fluids, a target tube immersed in the fluid provides a target of known emissivity.
2. *Radiation absorption between target and instrument.* Smoke, gases, and glass lenses absorb some of the radiation and reduce the incoming signal. Use of an enclosed (or purged) target tube or direct calibration will correct this.
3. *Focusing of the target on the sensing element.*

MEASUREMENT OF FLUID FLOW RATE

(See also Secs. 3 and 4.)

Flow is expressed in volumetric or mass units per unit time. Thus gases are generally measured in ft^3/min (m^3/min) or ft^3/h (m^3/h), steam in lb/h

(kg/h), and liquid in gal/min (L/min) or gal/h (L/h). Conversion between volumetric flow Q and mass flow m is given by $m = K\rho Q$, where ρ = density of the fluid and K is a constant depending on the units of m, Q, and ρ. Flow rate can be measured directly by attaching a rate device to a volumetric meter of the types previously described, e.g., a tachometer connected to the rotating shaft of the nutating-disk meter (Fig. 16.1.13).

Flow is most frequently measured by application of the principle of conservation of mechanical energy through conversion of fluid velocity to pressure (head). Thus, if the fluid is forced to change its velocity from V_1 to V_2, its pressure will change from p_1 to p_2 according to the equation (neglecting friction, expansion, and turbulence effects):

$$V_2^2 - V_1^2 = \frac{2g_c}{\rho}(p_1 - p_2) \qquad (16.1.1)$$

where g = acceleration due to gravity, ρ = fluid density, and g_c = 32.184 lbm · ft/(lbf)(s²) [1.0 kg · m/(N)(s²)]. **Caution:** If the flow pulsates, the average value of $p_1 - p_2$ will be greater than that for steady flow of the same average flow.

See ''ASME Pipeline Flowmeters,'' and ''Pitot Tubes'' in Sec. 3 for coverage of venturi tubes, flow nozzles, compressible flow, orifice meters, ASME orifices, and Pitot tubes.

The tabulation orifice coefficients apply only for **straight pipe** upstream and downstream from the orifice. In most cases, satisfactory results are obtained if there are no fittings closer than 25 pipe diameters upstream and 5 diameters downstream from the orifice. The upstream limitation can be reduced a bit by employing **straightening vanes**. Reciprocating pumps in the line may introduce serious errors and require special efforts for their correction.

A wide variety of differential pressure meters is available for measuring the orifice (or other primary element) pressure drop.

Figure 16.1.29 shows the **diaphragm, or ''dry,'' meter**. The orifice differential acts across a metal or rubber diaphragm, generating a force which tends to rotate the lever clockwise, moving the baffle toward the nozzle. This increases the nozzle back pressure, which acts on the pilot diaphragm to open the air supply port and increase the output pressure.

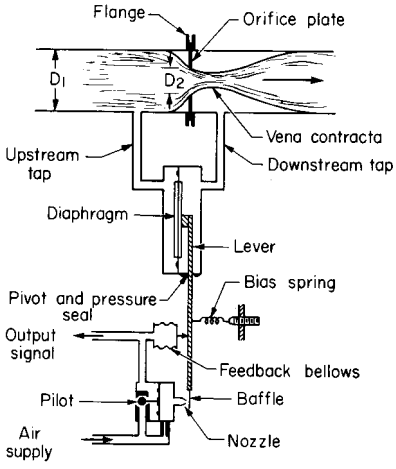


Fig. 16.1.29 Orifice plate and diaphragm-type meter.

This increases the force exerted by the feedback bellows, which generates a force opposing the motion of the main diaphragm. Equilibrium is reached when a change in orifice differential is exactly balanced by a proportionate change in output pressure. Often a damping device in the form of a simple oil dashpot is attached to the lever to reduce output fluctuations.

The flowmeter normally exhibits a square-root flow calibration. Some meters are designed to take out the square root by use of **cams, characterized floats or displacers (Ledoux bell)** or devices which describe a square-root behavior. These methods do not improve accuracy or performance but merely provide the convenience of a linear scale.

The meters described thus far are termed **variable-head** because the pressure drop varies with the flow, orifice ratio being fixed. In contrast, the **variable-area** meter maintains a constant pressure differential but varies the orifice area with flow.

The **rotameter** (Fig. 16.1.30) consists of a float positioned inside a tapered tube by action of the fluid flowing up through the tube. The flow restriction is now the annular area between the float and the tube (area increases as the float rises). The pressure differential is fixed, determined by the weight of the float and the buoyant forces. To satisfy the

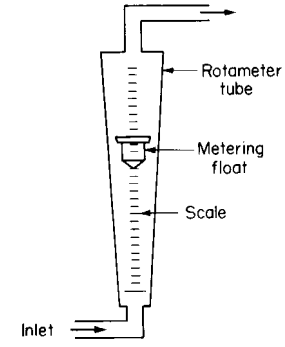


Fig. 16.1.30 Rotameter.

volumetric flow equation then, the annular area (hence the float level) must increase with flow rate. Thus the rotameter may be calibrated for direct flow reading by etching an appropriate scale on the surface of the glass tube. The calibration depends on the float dimensions, tube taper, and fluid properties. The equation for volumetric flow is

$$Q = C_R(A_T - A_F)\left[\frac{2gV_F}{\rho A_F}(\rho_F - \rho)\right]^{1/2}$$

where A_T = cross-sectional area of tube (at float position), A_F = effective float area, V_F = float volume, ρ_F = float density, ρ = fluid density, and C_R = rotameter coefficient (usually between 0.6 and 0.8). The coefficient varies with the fluid viscosity; however, special float designs are available which are relatively insensitive to viscosity effects. Also, fluid density compensation can be obtained.

The **rotameter reading may be transmitted** for recording and control purposes by affixing to the float a stem which connects to an armature or permanent magnet. The armature forms part of an inductance bridge whose signal is amplified electronically to drive a pen-positioning motor. For pneumatic transmission, the magnet provides magnet coupling to a pneumatic motion transmitter external to the rotameter tube. This generates an air pressure proportional to the height of the float.

The **area meter** is similar to the rotameter in operation. Flow area is varied by motion of a piston in a straight cylinder with openings cut into the wall. The piston position is transmitted as above by an armature and inductance bridge circuit.

Primary elements for **flow in open channels** usually employ **weirs** or **open nozzles** to restrict the flow. Weir designs include the rectangular slot; the V notch; and for a linear-flow characteristic, the parabolically shaped weir (Sutro weir). The flow rate is determined from the height of the liquid surface relative to the base of the weir. This height is measured by a liquid-level device, usually float-actuated. A still well (float chamber or open standpipe) connected to the bottom of the weir or the nozzle tap is used to avoid errors in float displacement due to the motion of the flowing fluid or to the buildup of solids. (See also Sec. 3.)

There are many other kinds of flow instruments which serve special purposes of accuracy, response, or application. The **propeller type** (Fig. 16.1.31) responds linearly to the average velocity in the path of the propeller, assuming negligible friction. The propeller may be mechanically geared to a tachometer to indicate flow rate and to a counter to

show total quantity flow. The magnetic pickup (Fig. 16.1.31) generates a pulse each time a propeller tip passes. The frequency of pulses (measured by means of appropriate electronic circuitry) is then proportional to the local stream velocity. If the propeller occupies only part of the flow stream, an individual calibration is necessary and the velocity distribution must remain constant. The **turbine** type is similar, but is fabricated as a unit in a short length of pipe with vanes to guide the flow approaching the rotor. Its magnetic pickup permits hermetic sealing. A minimum flow is needed to overcome magnetic cogging and start the rotor turning.

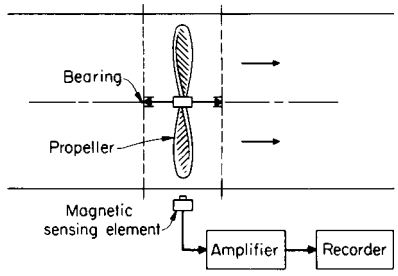


Fig. 16.1.31 Propeller-type flowmeter.

The **metering pump** is an accurately calibrated positive-displacement pump which provides both measurement and control of fluid-flow rate. The pump may be either fixed volumetric displacement-variable speed or constant speed-variable displacement.

For air flow, a **vane-type meter (anemometer)** is often used. A mechanical counter counts the number of revolutions of the vane shaft over a timed interval. Instantaneous airflow readings are more readily obtained with the **hot-wire anemometer**. Here, a resistance wire heated by an electric current is placed in the flow stream. The temperature of the wire depends on the current and the rate at which heat is conducted away from it. This latter factor is related to the thermal properties of the air and its velocity past the wire. Airflow can be measured in terms of (1) the current through the wire to maintain a fixed temperature, (2) temperature of the wire for a fixed current, or (3) temperature rise of the air passing the wire for fixed current. The wire temperature is readily measured in terms of its resistance. The anemometer must be specially calibrated for the application. Lasers have also been applied to anemometer use.

The **electromagnetic flowmeter** has no moving parts and does not require any insertions in the flow stream. It is based on the voltage induced by the flow of charged particles of the fluid past a strong magnetic field. It is suitable for liquids having resistivities of 50 $k\Omega \cdot cm$ or less. The **vortex-shedding meter** has a flow obstruction in the pipe; vortices form behind it at a rate nearly proportional to the volume flow rate. Vortex-formation-rate data give flow rate; a counter gives the integrated flow.

Doppler-effect flowmeters depend on reflection from particles moving with the fluid being metered; the shift in frequency of the reflected wave is proportional to velocity. Two transducers are used side by side, directed so that there is a large component of flow velocity along the sound path. One transmits and one receives.

Transit-time **ultrasonic flowmeters** use one or more pairs of transducers on opposite sides of the pipe, displaced along the length of the pipe. The apparent velocity of sound is $c \pm v$, where c is the speed of sound with no flow, and v is the component of flow velocity in the direction of the sound propagation path. The difference in sound velocity in the two directions is proportional to the flow's velocity component along the sound propagation path. The transit time difference is $2vl/(c^2 + v^2)$, where l is the path length. For $v \ll c$, the factor $(c^2 + v^2)$ is nearly constant. These meters cause no pressure drop and can be applied to pipes up to very large diameters. Multipath meters improve accuracy.

Mass flowmeters measure changes in momentum related to the mass flow rate.

Flowmeters measure rate of flow. To measure the total quantity of fluid flowing during a specified interval of time, the flow rate must be integrated over that interval. The integration may be done manually by estimating from the chart record the hourly flow averages or by measuring the area under the flow curve with a special square-root planimeter. **Mechanical integrators** use a constant-speed motor to rotate a counter. A cam converts the square-root meter reading into a linear displacement such that the fraction of time that the motor is engaged to the counter is proportional to the flow rate, resulting in a counter reading proportional to the integrated flow. **Electrical integrators** are similar in principle to the watthour meter in that the speed of the integrating motor is made proportional to the magnitude of the flow signal (see Sec. 15).

POWER MEASUREMENT

Power is defined as the rate of doing work. Common units are the horsepower and the kilowatt: 1 hp = 33,000 $ft \cdot lb/min$ = 0.746 kW. The power input to a rotating machine in hp (W) = $2\pi nT/k$, where n = r/min of the shaft where the torque T is measured in $lbf \cdot ft$ ($N \cdot m$), and k = 33,000 $ft \cdot lbf/hp \cdot min$ [60 $N \cdot m/(W \cdot min)$]. The same equation applies to the power output of an engine or motor, where n and T refer to the output shaft. Mechanical power-measuring devices **(dynamometers)** are of two types: (1) those absorbing the power and dissipating it as heat and (2) those transmitting the measured power. As indicated by the above equation, two measurements are involved: shaft speed and torque. The speed is measured directly by means of a tachometer. Torque is usually measured by balancing against weights applied to a fixed lever arm; however, other force measuring methods are also used. In the **transmission dynamometer,** the torque is measured by means of strain-gage elements bonded to the transmission shaft.

There are several kinds of **absorption dynamometers**. The **Prony brake** applies a friction load to the output shaft by means of wood blocks, flexible band, or other friction surface. The **fan brake** absorbs power by ''fan'' action of rotating plates on surrounding air. The **water brake** acts as an inefficient centrifugal pump to convert mechanical energy into heat. The pump casing is mounted on antifriction bearings so that the developed turning moment can be measured. In the **magnetic-drag** or **eddy-current brake,** rotation of a metal disk in a magnetic field induces eddy currents in the disk which dissipate as heat. The field assembly is mounted in bearings in order to measure the torque.

One type of **Prony brake** is illustrated in Fig. 16.1.32. The torque developed is given by $L(W - W_0)$, where L is the length of the brake arm, ft; W and W_0 are the scale loads with the brake operating and with the brake free, respectively. The brake horsepower then equals $2\pi nL(W - W_0)/33{,}000$, where n is shaft speed, r/min.

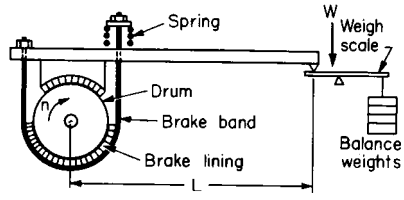


Fig. 16.1.32 Prony brake.

In addition to eddy-current brakes, **electric dynamometers** include calibrated generators and motors and cradle-mounted generators and motors. In calibrated machines, the efficiency is determined over a range of operating conditions and plotted. Mechanical power measurement can then be made by measuring the electrical power input (or output) to the machine. In the electric-cradle dynamometer, the motor or generator stator is mounted in trunnion bearing so that the torque can be measured by suitable scales.

The **engine indicator** is a device for plotting cylinder pressure as a function of piston (or volume) displacement. The resulting p-v diagram (Fig. 16.1.33) provides both a measure of the work done in a reciprocat-

ing engine, pump, or compressor and a means for analyzing its performance (see Secs. 4, 9, and 14). If A_d is the area inside the closed curve drawn by the indicator, then the indicated horsepower for the cylinder under test $= KnA_pA_d$ where K is a proportionality factor determined by the scale factors of the indicator diagram, n = engine speed, r/min, A_p = piston area.

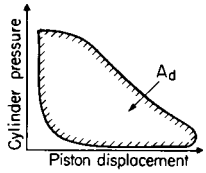


Fig. 16.1.33 Indicator diagram.

Completely mechanical indicators can be used only for low-speed machines. They have largely been superseded by electrical transducers using strain gages, variable capacitance, piezoresistive, and piezoelectric principles which are suitable for **high-speed** as well as low-speed pressure changes (the piezoelectric principle has low-speed limitations). The usual diagram is produced on an oscilloscope display as pressure vs. time, with a marker to indicate some reference event such as spark timing or top dead center. Special transducers can be coupled to a crank or cam shaft to give an electrical signal representing piston motion so that a *p-v* diagram can be shown on an oscilloscope.

ELECTRICAL MEASUREMENTS

(See also Sec. 15.)

Electrical measurements serve two purposes: (1) to measure the electrical quantities themselves, e.g., line voltage, power consumption, and (2) to measure other physical quantities which have been converted into electrical variables, e.g., temperature measurement in terms of thermocouple voltage.

In general, there is a sharp distinction between ac and dc devices used in measurements. Consequently, it is often desirable to transform an ac signal to an equivalent dc value, and vice versa. An ac signal is converted to dc (rectified) by use of **selenium rectifiers, silicon or germanium diodes, or electron-tube diodes**. Full-wave rectification is accomplished by the **diode bridge,** shown in Fig. 16.1.34. The rectified signal may be passed through one or more low-pass filter stages to smooth the waveform to its average value. Similarly, there are many ways of modulating a dc signal (converting it to alternating current). The most common method used in instrument applications is a solid-state oscillator.

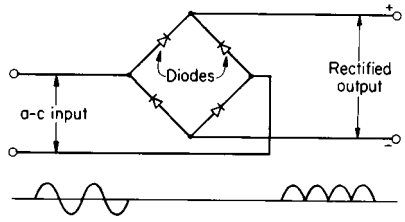


Fig. 16.1.34 Full-wave rectifier.

The **galvanometer** (Fig. 16.1.35), recently supplanted by the direct-reading **digital voltmeter** (DVM), is basic to dc measurement. The input signal is applied across a coil mounted in jeweled bearings or on a taut-band suspension so that it is free to rotate between the poles of a permanent magnet. Current in the coil produces a magnetic moment which tends to rotate the coil. The rotation is limited, however, by the restraining torque of the hairsprings. The resulting deflection of the coil θ is proportional to the current I:

$$\theta = \frac{NBWL}{K} I$$

where N = number of turns in coil; W, L = coil width and length, respectively; B = magnetic field intensity; K = spring constant of the hairsprings. Galvanometer deflection is indicated by a balanced pointer attached to the coil. In very sensitive elements, the pointer is replaced by a mirror reflecting a spot of light onto a ground-glass scale; the bearings and hairspring are replaced by a torsion-wire suspension.

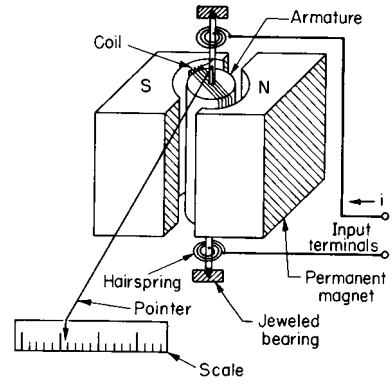


Fig. 16.1.35 D'Arsonval galvanometer.

The galvanometer can be converted into a **dc voltmeter, ammeter, or ohmmeter** by application of Ohm's law, $IR = E$, where I = current, A; E = electrical potential, V; and R = resistance, Ω.

For a **voltmeter,** a fixed resistance R is placed in series with the galvanometer (Fig. 16.1.36*a*). The current i through the galvanometer is proportional to the applied voltage E: $i = E/(r + R)$, where r = coil resistance. Different voltage ranges are obtained by changing the series resistance.

An **ammeter** is produced by placing the resistance in parallel with the galvanometer or DVM (Fig. 16.1.36*b*). The current then divides between the galvanometer coil or DVM and the resistor in inverse ratio to their resistance values (r and R, respectively); thus, $i = IR/(r + R)$, where i = current through coil and I = total current to be measured. Different current ranges are obtained by using different shunt resistances.

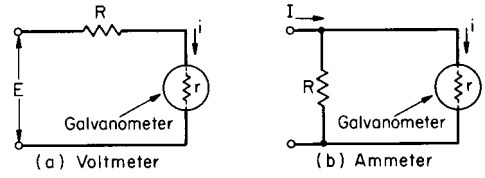


Fig. 16.1.36 (*a*) Voltmeter; (*b*) ammeter.

The common ohmmeter consists of a battery, a galvanometer with a shunt rheostat, and resistance in series to total Ri Ω. The shunt is adjusted to give a full-scale (0 Ω) reading with the test terminals shorted. When an unknown resistance R is connected, the deflection is $Ri/(Ri + R)$ fraction of full scale. The scale is calibrated to read R directly. A half-scale deflection indicates $R = Ri$. Alternatively, the galvanometer is connected to read voltage drop across the unknown R while a known current flows through it. This principle is used for low-value resistances and in digital ohmmeters.

Digital instruments are available for all these applications and often offer higher resolution and accuracy with less circuit loading. Fluctuating readings are difficult to follow, however.

Alternating current and voltage must be measured by special means. A dc instrument with a rectifier input is commonly used in applications requiring high input impedance and wide frequency range. For precise measurement at power-line frequencies, the electrodynamic instrument is used. This is similar to the galvanometer except that the permanent magnet is replaced by an electromagnet. The movable coil and field

coils are connected in series; hence they respond simultaneously to the same current and voltage alternations. The pointer deflection is proportional to the square of the input signal. The moving-iron-type instruments consist of a soft-iron vane or armature which moves in response to current flowing through a stationary coil. The pointer is attached to the iron to indicate the deflection on a calibrated nonlinear scale. For measuring at very high frequencies, the **thermocouple voltmeter or ammeter** is used. This is based on the heating effect of the current passing through a fixed resistance R. Heat is liberated at the rate of E^2/R or $I^2 R W$.

DC **electrical power** is the product of the current through the load and the voltage across the load. Thus it can be simply measured using a voltmeter and ammeter. AC power is directly indicated by the wattmeter, which is similar to the electrodynamic instrument described above. Here the field coils are connected in series with the load, and the movable coil is connected across the load (to measure its voltage). The deflection of the movable coil is then proportional to the effective load power.

Precise voltage measurement (direct current) can be made by balancing the unknown voltage against a measured fraction of a known reference voltage with a **potentiometer** (Fig. 16.1.9). Balance is indicated by means of a sensitive current detector placed in series with the unknown voltage. The potentiometer is calibrated for angular position vs. fractional voltage output. Accuracies to 0.05 percent are attainable, dependent on the linearity of the potentiometer and the accuracy of the reference source. The reference standard may be a Weston standard cell or a regulated voltage supply (based on diode characteristics). The balance detector may be a galvanometer or electronic amplifier.

Precision resistance and general impedance measurements are made with bridge circuits (Fig. 16.1.37) which are adjusted until no signal is detected by the null detector (bridge is balanced). Then $Z_1Z_3 = Z_2Z_4$. The basic Wheatstone bridge is used for resistance measurement where all the impedances (Z's) are resistances (R's). If R_1 is to be measured, $R_1 = R_2R_4/R_3$ when balanced. A sensitive galvanometer for the null detector and dc voltage excitation is usual. All R_2, R_3, R_4 must be calibrated, and some adjustable. For general impedance measurement, ac voltage excitation of suitable frequency is used. The null detector may be a sensitive ac meter, oscilloscope, or, for audio frequencies, simple earphones. The basic balance equation is still valid, but it now

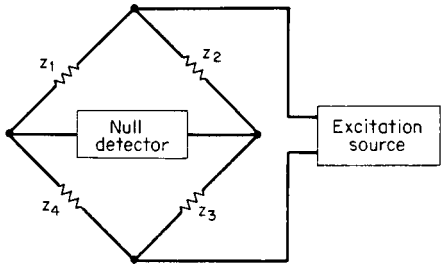


Fig. 16.1.37 Impedance bridge.

requires also that the sum of the phase angles of Z_1 and Z_3 equal the sum of the phase angles of Z_2 and Z_4. As an example, if Z_1 is a capacitor, the bridge can be balanced if Z_2 is a known capacitor while Z_3 and Z_4 are resistances. The phase-angle condition is met, and $Z_1Z_3 = Z_2Z_4$ becomes $(\frac{1}{2}\pi fC_1)R_3 = (\frac{1}{2}\pi fC_2)R_4$ and $C_1 = C_2R_3/R_4$. Variations on the basic principle include the Kelvin bridge for measurement of low resistance, and the Mueller bridge for platinum resistance thermometers.

Voltage measurement requires a meter of substantially higher impedance than the impedance of the source being measured. The vacuum-tube cathode follower and the field-effect transistor are suitable for high-impedance inputs. The following circuitry may be a simple amplifier to drive a pointer-type meter, or may use a digital technique to produce a digital output and display. Digital counting circuits are capable of great precision and are widely adapted to measurements of time, frequency, voltage, and resistance. Transducers are available to convert temperature, pressure, flow, length and other variables into signals suitable for these instruments.

The charge amplifier is an example of an operational amplifier application (Fig. 16.1.38). It is used for outputs of piezoelectric transducers in which the output is a charge proportional to input force or other input converted to a force. Several capacitors switchable across the feedback path provide a range of full-scale values. The output is a voltage.

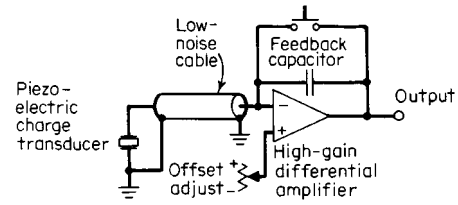


Fig. 16.1.38 Charge amplifier application.

The **cathode-ray oscilloscope** (Fig. 16.1.39) is an extremely useful and versatile device characterized by high input impedance and wide frequency range. An electron beam is focused on the phosphor-coated face of the cathode-ray tube, producing a visible spot of light at the point of impingement. The beam is deflected by applying voltages to vertical and horizontal deflector plates. Thus, the relationship between two varying voltages can be observed by applying them to the vertical and horizontal plates. The horizontal axis is commonly used for a linear time base generated by an internal sawtooth-wave generator. Virtually any desired sweep speed is obtainable as a calibrated sweep. Sweeps which change value part way across the screen are available to provide localized time magnification. As an alternative to the time base, any arbitrary voltage can be applied to drive the horizontal axis. The vertical axis is usually used to display a dependent variable voltage. **Dual-beam** and

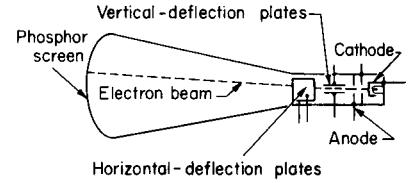


Fig. 16.1.39 Cathode-ray tube.

dual-trace instruments show two waveforms simultaneously. Special long-persistence and storage screens can hold transient waveforms for from seconds to hours. Greater versatility and unique capabilities are afforded by use of **digital-storage** oscilloscope. Each input signal is sampled, digitized, and stored in a first-in–first-out memory. Since a record of the recent signal is in memory when a trigger pulse is received, the timing of the end of storing new data into memory determines how much of the stored signal was before, and how much after, the trigger. Unlike storage screens, the stored signal can be amplified and shifted on the screen for detailed analysis, accompanied by numerical display of voltage and time for any point. Care must be taken that enough samples are taken in any waveform; otherwise aliasing results in a false view of the waveform.

The stored data can be processed mathematically in the oscilloscope or transferred to a computer for further study. Accessories for microcomputers allow them to function as digital oscilloscopes and other specialized tasks.

VELOCITY AND ACCELERATION MEASUREMENT

Velocity or speed is the time rate of change of displacement. Consequently, if the displacement measuring device provides an output signal which is a continuous (and smooth) function of time, the velocity can be measured by **differentiating** this signal either graphically or by use of a differentiating circuit. The accuracy may be very limited by noise

(high-frequency fluctuations), however. More commonly, the output of an accelerometer is integrated to yield the velocity of the moving member. **Average speed** over a time interval can be determined by measuring the time required for the moving body to pass two fixed points a known distance apart. Here photoelectric or other rapid sensing devices may be used to trigger the start and stop of the timer. **Rotational speed** may be similarly measured by counting the number of rotations in a fixed time interval.

The **tachometer** provides a direct measure of angular velocity. One form is essentially a small permanent-magnet-type generator coupled to the rotating element; the voltage induced in the armature coil is directly proportional to the speed. The principle is also extended to rectilinear motions (restricted to small displacements) by using a straight coil moving in a fixed magnetic field.

Angular velocity can also be measured by magnetic drag-cup and **centrifugal-force** devices (flyball governor). The force may be balanced against a spring with the resulting deflection calibrated in terms of the shaft speed. Alternatively, the force may be balanced against the air pressure generated by a pneumatic nozzle-baffle assembly (similar to Fig. 16.1.23).

Vibration velocity pickups may use a coil which moves relative to a magnet. The voltage generated in the coil has the same frequency as the vibration and, for sine motion, a magnitude proportional to the product of vibration frequency and amplitude. Vibration **acceleration pickups** commonly use strain-gage, piezoresistive, or piezoelectric elements to sense a force $F = Ma/g_c$. The maximum usable frequency of an accelerometer is about one-fifth of the pickup's natural frequency (see Sec. 3.4). The minimum usable frequency depends on the type of pickup and the associated circuitry. The output of an accelerometer can be integrated to obtain a velocity signal; a velocity signal can be integrated to obtain a displacement signal. The **operational amplifier** is a versatile element which can be connected as an integrator for this use.

Holography is being applied to the study of surface vibration patterns.

MEASUREMENT OF PHYSICAL AND CHEMICAL PROPERTIES

Physical and chemical measurements are important in the control of product quality and composition. In the case of manufactured items, such properties as color, hardness, surface, roughness, etc., are of interest. Color is measured by means of a **colorimeter,** which provides comparison with color standards, or by means of a **spectrophotometer,** which analyzes the color spectrum. The **Brinell and Rockwell testers** measure surface hardness in terms of the depth of penetration of a hardened steel ball or special stylus. Testing machines with **strain-gage** elements provide measurement of the strength and elastic properties of materials. **Profilometers** are used to measure surface characteristics. In one type, the surface contour is magnified optically and the image projected onto a screen or viewer; in another, a stylus is employed to translate the surface irregularities into an electrical signal which may be recorded in the form of a highly magnified profile of the surface or presented as an averaged roughness-factor reading.

For liquids, attributes such as density, viscosity, melting point, boiling point, transparency, etc., are important. Density measurements have already been discussed. **Viscosity** is measured with a **viscosimeter,** of which there are three main types: flow through an orifice or capillary (Saybolt), viscous drag on a cylinder rotating in the fluid (MacMichael), damping of a vibrating reed (Ultrasonic) (see Secs. 3 and 4). **Plasticity** and consistency are related properties which are determined with special apparatus for heating or cooling the material and observing the temperature-time curve. The **photometer, reflectometer, and turbidimeter** are devices for measuring transparency or turbidity of nonopaque liquids and solids.

A variety of properties can be measured for determining chemical composition. **Electrical properties** include pH, conductivity, dielectric constant, oxidation potential, etc. **Physical properties** include density, refractive index, thermal conductivity, vapor pressure, melting and boiling points, etc. Of increasing industrial application are **spectroscopic measurements: infrared absorption spectra, ultraviolet and visible emission spectra, mass spectrometry,** and **gas chromatography**. These are specific to particular types of compounds and molecular configurations and hence are very powerful in the analysis of complex mixtures. As examples, infrared analyzers are in use to measure low-concentration contaminants in engine oils resulting from wear and in hydraulic oils to detect deterioration. **X-ray diffraction** has many applications in the analysis of crystalline solids, metals, and solid solutions.

Of special importance in the realm of composition measurements is the determination of **moisture content**. A common laboratory procedure measures the loss of weight of the oven-dried sample. More rapid methods employ electrical conductance or capacitance measurements, based on the relatively high conductivity and dielectric constant values for ordinary water.

Water vapor in air (humidity) is measured in terms of its physical properties or effects on materials (see also Secs. 4 and 12). (1) The **psychrometer** is based on the cooling effect of water evaporating into the airstream. It consists of two thermal elements exposed to a steady airflow; one is dry, the other is kept moist. See Sec. 4 for psychrometric charts. (2) The **dew-point recorder** measures the temperature at which water just starts to condense out of the air. (3) The **hygrometer** measures the change in length of such humidity-sensitive elements as hair and wood. (4) **Electric sensing elements** employ a wire-wound coil impregnated with a hygroscopic salt (one that maintains an equilibrium between its moisture content and the air humidity) such that the resistance of the coil is related to the humidity.

The **throttling calorimeter** (Fig. 16.1.40) is most commonly used for determining the moisture in steam. A sampling nozzle is located preferably in a vertical section of steampipe far removed from any fittings. Steam enters the calorimeter through a throttling orifice and into a well-insulated expansion chamber. The steam quality x (fraction dry steam) is determined from the equation $x = (h_c - h_f) / h_{fg}$, where h_c is the enthalpy of superheated steam at the temperature and pressure measured

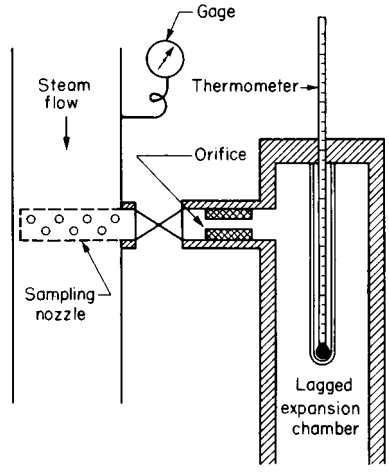


Fig. 16.1.40 Throttling calorimeter.

in the calorimeter; h_f and h_{fg} are, respectively, the liquid enthalpy and the heat of vaporization corresponding to line pressure. The chamber is conveniently exhausted to atmospheric pressure; then only line pressure and temperature of the throttled steam need be measured. The range of the throttling calorimeter is limited to small percentages of moisture; a **separating calorimeter** may be employed for larger moisture contents.

The **Orsat** apparatus is generally used for chemical analysis of flue gases. It consists of a graduated tube or burette designed to receive and measure volumes of gas (at constant temperature). The gas is analyzed for CO_2, O_2, CO, and N_2 by bubbling through appropriate absorbing reagents and measuring the resulting change in volume. The reagents normally employed are KOH solution for CO_2, pyrogallic acid and

KOH mixture for O_2, and cuprous chloride (Cu_2Cl_2) for CO. The final remaining unabsorbed gas is assumed to be N_2. The most common errors in the Orsat analysis are due to **leakage** and poor **sampling**. The former can be checked by simple test; the latter factor can only be minimized by careful sampling procedure. Recommended procedure is the taking of several simultaneous samples from different points in the cross-sectional area of the flue-gas stream, analyzing these separately, and averaging the results.

There are many instruments for **measuring CO_2 (and other gases) automatically**. In one type, the CO_2 is absorbed in KOH, and the change in volume determined automatically. The more common type, however, is based on the difference in **thermal conductivity** of CO_2 compared with air. Two thermal conductivity cells are set into opposing arms of a Wheatstone-bridge circuit. Air is sealed into one cell (reference), and the CO_2-containing gas is passed through the other. The cell contains an electrically heated resistance element; the temperature of the element (and therefore its resistance) depends on the thermal conductivity of the gaseous atmosphere. As a result, the unbalance of the bridge provides a measure of the CO_2 content of the gas sample.

The same principle can be employed for analyzing other constituents of gas mixtures where there is a significant thermal-conductivity difference. A modification of this principle is also used for determining CO or other combustible gases by mixing the gas sample with air or oxygen. The combustible gas then burns on the heated wire of the test cell, producing a temperature rise which is measured as above.

Many other physical properties are employed in the determination of specific components of gaseous mixtures. An interesting example is the **oxygen analyzer,** based on the unique paramagnetic properties of oxygen.

NUCLEAR RADIATION INSTRUMENTS

(See also Sec. 9.)

Nuclear radiation instrumentation is increasing in importance with two main areas of application: (1) measurement and control of radiation variables in nuclear reactor-based processes, such as nuclear power plants and (2) measurement of other physical variables based on radioactive excitation and tracer techniques. The instruments respond in general to electromagnetic radiation in the gamma and perhaps X-ray regions and to beta particles (electrons), neutrons, and alpha particles (helium nuclei).

Gas Ionization Tubes The **ion chamber, proportional counter,** and **Geiger counter** are common instruments for radiation detection and measurement. These are different applications of the gas-ionization tube distinguished primarily by the amount of applied voltage.

A simple and very common form of the instrument consists of a gas-filled cylinder with a fine wire along the axis forming the anode and the cylinder wall itself (at ground potential) forming the cathode, as shown in Fig. 16.1.41. When a radiation particle enters the tube, its collision with gas molecules causes an ionization consisting of electrons (negatively charged) and positive ions. The electrons move very rapidly toward the positively charged wire; the heavier positive ions move relatively slowly toward the cathode. The above activity is detected by the resulting current flow in the external circuitry.

When the voltage applied across the tube is relatively low, the number of electrons collected at the anode is essentially equal to that produced by the incident radiation. In this voltage range, the device is called an **ion chamber.** The device may be used to count the number of radiation particles when the frequency is low; when the frequency is high, an external integrating circuit yields an output current proportional to the radiation intensity. Since the amplification factor of the ion chamber is low, high-gain electronic amplification of the current signal is necessary.

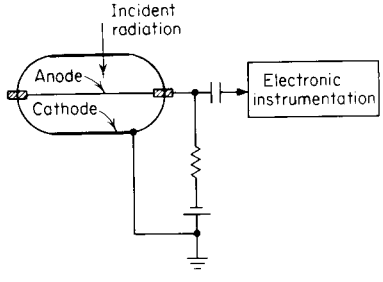


Fig. 16.1.41 Gas-ionization tube.

If the applied voltage is increased, a point is reached where the radiation-produced ions have enough energy to collide with other gas molecules and produce more ions which also enter into collisions so that an ''avalanche'' of electrons is collected at the anode. Thus, there is a very considerable amplification of the output signal. In this region, the device is called a **proportional counter** and is characterized by the voltage or current pulse being proportional to the energy content of the incident radiation signal.

With still further increase in the applied voltage, a point of saturation is reached wherein the output pulses have a constant amplitude independent of the incident radiation level. The resulting **Geiger counter** is capable of producing output pulses up to 10 V in amplitude, thus greatly reducing the requirements on the external circuitry and instrumentation. This advantage is offset somewhat by a lower maximum counting rate and more limited ability to differentiate among the various types of radiation as compared with the proportional counter.

The **scintillation counter** is based on the excitation of a phosphor by incident radiation to produce light radiation which is in turn detected by a photomultiplier tube to yield an output voltage. The signal output is greatly amplified and nearly proportional to the energy of the initial radiation. The device may be applied to a wide range of radiations, it has a very fast response, and, by choice of phosphor material, it offers a large degree of flexibility in applications.

Applications to the Measurement of Physical Variables The ready availability of radioactive isotopes of long half-life, such as cobalt 60, make possible a variety of industrial and laboratory measuring techniques based on radiation instruments of the type described above. Most applications are based on (1) radiation absorption, (2) tracer identification, and (3) other properties. These techniques often have the advantages of isolation of the measuring device from the system, access to a variable not observable by conventional means, or measurement without destruction or modification of the system.

In the utilization of **absorption** properties, a radioactive source is separated from the radiation-measuring device by that part of the system to be measured. The measured radiation intensity will depend on the fraction of radiation absorbed, which in turn will depend on the distance traveled through the absorbing medium and the density and nature of the material. Thus, the instrument can be adapted to measuring thickness (see Fig. 16.1.12), coating weight, density, liquid or solids level, or concentration (of certain components).

Tracer techniques are effectively used in measuring flow rates or velocities, residence time distributions, and flow patterns. In flow measurement, a sharp pulse of radioactive material may be injected into the flow stream; with two detectors placed downstream from the injection point and a known distance apart, the velocity of the pulse is readily measured. Alternatively, if a known constant flow rate of tracer is injected into the flow stream, a measure of the radiation downstream is easily converted into a measure of the desired flow rate. Other applications of tracer techniques involve the use of tagged molecules embedded in the process to provide measures of wear, chemical reactions, etc.

Other applications of radiation phenomena include level measurements based on a floating radioactive source, level measurements based on the back-scattering effect of the medium, pressure measurements in the high-vacuum region based on the amount of ionization caused by alpha rays, location of interface in pipeline transmission applications, and certain chemical analysis applications.

INDICATING, RECORDING, AND LOGGING

An important element of measurement is the display of the measured value in a form which the human operator can readily interpret. Two

basic types of display are employed: analog and digital. **Analog** refers to a reading obtained from the motion of a pointer on a scale or the record of a pen moving over a chart. Digital refers to the reading displayed as a number, a series of holes on a punched card, a sequence of pulses on magnetic tape, or dots on a heat sensitive paper surface forming a trace. Further classification relates to indicating and recording functions. The **indicator** consists merely of a pointer moving over a calibrated scale. The scale may be concentric, as in the Bourdon gage (Fig. 16.1.16) or eccentric, as in the flowmeter. There are also digital indicators which directly display or illuminate the specific digits corresponding to the reading. Obviously, use of the indicator is limited to cases where the variable of interest is constant during the measuring period, or at most, changes slowly.

The **recorder** is used where long-term trends or detailed variations with time are of interest, or where the response is too rapid for the human eye to follow. In the common **circular-chart recorder** (Fig. 16.1.42), the pointer is replaced by a pen which writes on a chart rotated by a constant-speed electrical or spring-wound clock. Various chart

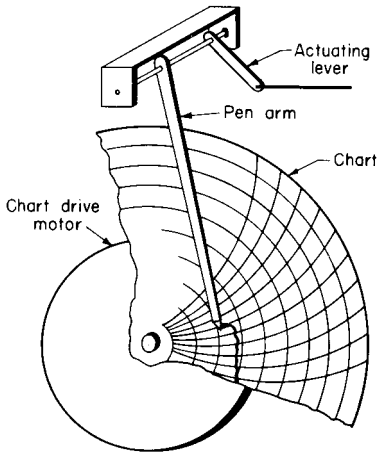


Fig. 16.1.42 Circular-chart recorder.

speeds are available from 1 r/min to 1 every 7 days. Up to four recording pens on a single chart are available (with a print-wheel mechanism, six color-identified records may be had). The **strip-chart recorder** shown in Fig. 16.1.43 is of the type used in electronic potentiometers, where the pen is positioned by a servomotor or a stepper motor as in the case of a digital recorder. A constant-speed motor drives the chart vertically past the pen, which deflects horizontally. **Multipoint recording** is achieved by replacing the pen with a print-wheel assembly. A selector switch

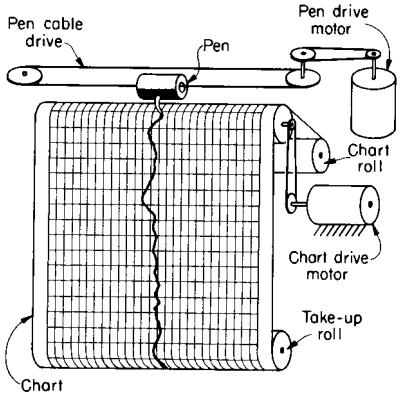


Fig. 16.1.43 Strip-chart recorder.

switches the input signal from one variable to another at the same time that the print wheel switches from one number (or symbol) to another. The record of each variable appears then as a sequence of dots with an identifying numeral. Up to 16 different records may be recorded on a chart (with external switching, as many as 144 records have been applied). Miniature recorders with 3- and 4-in strip charts are gaining favor in process industries because of their compactness and readability. The pen may be pneumatically or electrically actuated. Maximum number of records per chart is two.

For **direct-writing recording** of high-speed phenomena up to about 100 Hz, a pen or stylus can be driven by a galvanometer. The chart is in strip form and is driven at a speed suitable for the resolution needed. Recording may be done with ink and standard chart paper or heated stylus and special heat-sensitive paper. Mirror galvanometers projecting a spot of light onto a moving chart of light-sensitive paper can be used up to several kilohertz. A number of galvanometers can be used side by side to record several signals simultaneously on the same chart.

For higher frequencies a form of **magnetic recording** is common. Analog signals can be recorded by amplitude and frequency modulation. The latter is particularly convenient for playback at reduced speed.

Digital signals can be recorded in magnetic form. They can be recorded to any desired precision by using more bits to represent the data. Resolution is 1 part in 2^n, where n is the number of bits used in straight binary form, less in binary-coded decimal, where 4 bits are used to encode each decimal digit. Digital recording and data transmission have the advantage that in principle error rates can be made as small as desired in the presence of noise by adding more bits which serve as checks in error correcting codes (Raisbeck, ''Information Theory: An Introduction for Scientists and Engineers,'' M.I.T. Press).

Most physical variables are in analog form. Popular standards for the transmission of analog signals include 3- to 15-psig pneumatic signals, direct currents of 4 to 20 or 10 to 50 mA, 0 to 5 and 0 to 10 V. Suitable **signal conditioners** are needed to convert thermocouple outputs and the like to these levels. (Of course, if the instrument is specifically for the particular thermocouple, this conversion is not needed.) This standardization gives greater flexibility in interconnecting signal sources with indicators and recorders. Some transducers and signal conditioners are designed to receive their power over the same two wires used to transmit their output signals.

Often it is necessary to convert from analog to digital form (as for the input to a digital computer) and vice versa. The **analog-digital (A/D)** and **digital-analog (D/A)** converters provide these interfaces. They are available in various conversion speeds and resolutions. Resolution is specified in terms of the number of bits in the digital signal.

Data which have been stored in magnetic form can be recovered at any time by connecting the storage device to an electrically actuated typewriter, printer, or other readout device. Modern **logging systems** have the measurements from hundreds of different points in the process tabulated periodically. These systems may provide such additional features as the printing of deviations from the normal in red and the more frequent scanning of abnormal conditions. Computer elements are also used in conjunction with logging systems to compute derived variables (such as operating efficiency, system losses, etc.) and to apply corrections to measured variables, e.g., temperature and pressure compensation of gas-flow readings.

In quality control and time-motion studies, often a simple **on-off-type recorder** is sufficient for the purpose. Here, a pen is deflected when the machine or system is on and not deflected whenever the system is off. Pen actuation is usually by solenoid or other electromagnetic element.

INFORMATION TRANSMISSION

In the analog form of data representation, a transmission variable (e.g., pressure, current, voltage, or frequency) is chosen appropriate to the data receiving device, distance, response speed, and environmental considerations. The variable may be related to the data by a simple linear function, by linearization such as taking the square root of an orifice pressure drop, or some other monotonic function.

A 3 to 15 (or 3 to 27, 6 to 30) psig air pressure, a 4 to 20 (or 10 to 50) mA dc current, and a 1 to 5 (or 0 to 5, 0 to 10) V dc voltage are in use to represent a data range of 0 to 100 percent. Where the 0 percent data level is transmitted as a nonzero value (e.g., 3 psig, 4 mA), a loss of power or a line break is detectable as an out-of-range condition. A variety of two-wire transmitters are powered by the loop current and they vary this same loop current to transmit data values. They are available for inputs including pressure, differential pressure, thermocouples, RTDs (resistance temperature detectors), strain gages, and pneumatic signals. Other systems use three or four wires, permitting separation of power and output signal. Converters are available to change a signal from one form to another, e.g., a 3- to 15-psig signal is converted to a 4- to 20-mA signal.

In the digital form of data transmission, patterns of binary (two-level) signals are sent in an agreed-upon manner to represent data. Binary coded decimal (BCD) uses 4 information units (bits) to represent each of the digits 0 through 9 (0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001). The ASCII (American National Standard Code for Information Interchange) code uses 7 bits (128 different codes) to represent the alphabet, digits, punctuation, and control codes. The PC character set extends the ASCII 7-bit code set by an eighth bit to provide another 128 codes (codes 128 through 255) devoted to foreign characters, mathematical symbols, lines, box, and shading elements. Data are often sent in groups of 8 bits (1 byte). Commonly used 8-bit transmission permits the full ASCII/PC character set to be transmitted.

Two distinct signal levels are used in binary digital transmission. Some values (at the receiving device) are:

Name	Space	Mark
Binary name	0	1
TTL (5 V), V	0.0 to 0.8	2.0 to 5.0
CMOS (3 to 15 V), % of supply volts	0 to 30	70 to 100
RS-232-C, V	+3 to +15	−3 to −15
20-mA loop	Current off	Current on
Telephone modem (Bell System 103), Hz tone		
From originator	1,070	1,270
To originator	2,025	2,225

The band between the two levels provides some protection against noise. The 20-mA loop uses a pair of wires for each transmission direction; optoisolators convert the 20-mA current to appropriate voltages at each end while providing electrical isolation.

The EIA Standard RS-232-C specifies an ''Interface Between Data Terminal Equipment (DTE) and Data Communication Equipment Employing Serial Binary Data Interchange.'' Twenty lines are defined; a minimum for two-way systems uses: line 1, protective ground; line 2, transmitted data (DTE to DCE); line 3, received data (DCE to DTE); and line 7, signal ground (common return). When a 9-pin (rather than 25-pin) connector is used, pin 3 is transmitted data and pin 2 is received data. If both devices are the same type, a crossover between wires 2 and 3 is needed. A cable with a 25-pin connector on one end and a 9-pin connector at the other end does not require a crossover.

In asynchronous serial transmission, an example of which is shown in Fig. 16.1.44, the no-signal state is the MARK level. A change to the SPACE level indicates that an ASCII code will start 1-bit-timer later. The start bit is always SPACE. The ASCII code follows, least significant bit first. This example is the letter S, binary form (most significant bit, msb, written first): 1 010 011, or 124 in octal form.

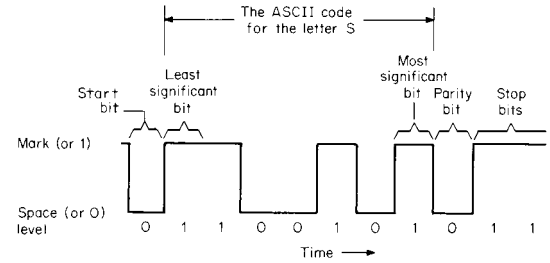


Fig. 16.1.44 ASCII transmission.

The parity bit is optional for error checking. This example uses even parity: the total number of 1s in the ASCII code and the parity bit is even. The bit sequence concludes with one or more stop bits at the MARK level. The minimum number required is set by the receiving device. Standard RS-232-C baud rates (bits per second) include 110, 150, 300, 600, 1,200, 2,400, 4,800, 9,600, 19,200, and 38,400. The EIA Standard RS-422 improves upon the RS-232-C by using transmission lines balanced to ground. This improves noise immunity and increases usable baud rates and transmission distance.

Transmissions are classed as **simplex** (one direction only), **half-duplex** (one direction at a time), and **full-duplex** (capable of simultaneous transmission in both directions). An agreed-upon protocol allows the receiver to signal the sender whether or not it is able to accept data. This may be done by a separate line(s) or by special (XON/XOFF; control Q/control S) ASCII signals on the return path of a full-duplex line.

For parallel transmission, multiple wires carry signals representing all the bits at once. Separate lines indicate when the receiver is ready for new data and when the sender has put new data on the lines. This exchange is called **handshaking**.

The above forms are used for communication between two devices. Where more than two devices are to be interconnected, a network, or bus system, is employed. The IEEE-488 General Purpose Interface Bus, GPIB (based on the Hewlett-Packard HP-1B), uses a parallel bus structure and can interconnect up to 15 devices, at a total connection path length of 20 m. One device acts as a controller at any time. Multiple instruments and control devices may be interconnected using 2 or 4 wire circuits and serial bus standard RS-485. Alternately, the ISA SP-50 protocol and other schemes still in development may be employed to achieve serial multidrop communications over distances of up to 2,500 m (8,200 ft). See also Sec. 2.2, ''Computers,'' and Sec. 15.2, ''Electronics.''

16.2 AUTOMATIC CONTROLS

by Gregory V. Murphy

REFERENCES: Thaler, ''Elements of Servomechanism Theory,'' McGraw-Hill. Shinskey, ''Process Control Systems: Application, Design and Tuning,'' McGraw-Hill. Kuo, ''Automatic Control Systems,'' Prentice-Hall. Phillips and Nagle, ''Digital Control System Analysis and Design,'' Prentice-Hall. Lewis, ''Applied Optimal Control and Estimation: Digital Design and Implementation,'' Prentice-Hall. Cochin and Plass, ''Analysis and Design of Dynamic Systems,'' HarperCollins. Astrom and Hagglund, ''Automatic Tuning of PID Controllers,'' Instrument Society of America. Maciejowski, ''Multivariable Feedback Design,'' Addison-Wesley. Murphy and Bailey, LQG/LTR Control System Design for a Low-Pressure Feedwater Heater Train with Time Delay, *Proc. IECON*, 1990. Murphy and Bailey, Evaluation of Time Delay Requirements for Closed-Loop Stability Using Classical and Modern Methods, *IEEE Southeastern Symp. on System Theory*, 1989. Murphy and Bailey, ''LQG/LTR Robust Control System Design for a Low-Pressure Feedwater Heater Train,'' *Proc. IEEE Southeastcon*,

1990. Kazerooni and Narayanan, ''Loop Shaping Design Related to LQG/LTR for SISO Minimum Phase Plants,'' *IEEE American Control Conf.*, Vol. 1, 1987. Murphy and Bailey, ''Robust Control Technique for Nuclear Power Plants,'' ORNL-10916, March 1989. Birdwell, Crockett, Bodenheimer, and Chang, The CASCADE Final Report: Vol. II, ''CASCADE Tools and Knowledge Base,'' University of Tennessee. Wang and Birdwell, A Nonlinear PID-Type Controller Utilizing Fuzzy Logic, *Proc. Joint IEEE/IFAC Symp. on Controller-Aided Control System Design*, 1994. Upadhyaya and Eryurek, Application of Neural Networks for Sensor Validation and Plant Monitoring, *Nuclear Technology*, **97,** no. 2, Feb. 1992. Vasadevan et al., Stabilization and Destabilization Slugging Behavior in a Laboratory Fluidized Bed, *International Conf. on Fluidized Bed Combustion*, 1995. Doyle and Stein, ''Robustness with Observers.'' *IEEE Trans. Automatic Control*, **AC-24,** 1979. Upadhaya et al. ''Development and Testing of an Integrated Validation System for Nuclear Power Plants,'' Report prepared for the U.S. Dept. of Energy. Vols. 1–3, DOE/NE/37959-34, 35, 36, Sept. 1989.

INTRODUCTION

The purpose of an **automatic control** on a system is to produce a desired output when inputs to the system are changed. Inputs are in the form of commands, which the output is expected to follow, and disturbances, which the automatic control is expected to minimize. The usual form of an automatic control is a **closed-loop feedback control** which Ahrendt defines as ''an operation which, in the presence of a disturbing influence, tends to reduce the difference between the actual state of a system and an arbitrarily varied desired state and which does so on the basis of this difference.'' The general theories and definitions of automatic control have been developed to aid the designer to meet primarily three basic specifications for the performance of the control system, namely, stability, accuracy, and speed of response.

The **terminology** of automatic control is being constantly updated by the ASME, IEEE, and ISA. Redundant terms, such as *rate, preact,* and *derivative,* for the same controller action are being standardized. Common terminology is still evolving. The introduction of the digital computer as a control device has necessitated the introduction of a whole new subset of terminology. The following terms and definitions have been selected to serve as a reference to a complex area of technology whose breadth crosses several professional disciplines.

Adaptive control system: A control system within which automatic means are used to change the system parameters in a way intended to improve the performance of the system.

Amplification: The ratio of output to input, in a device intended to increase this ratio. A gain greater than 1.

Attenuation: A decrease in signal magnitude between two points, or a gain of less than 1.

Automatic-control system: A system in which deliberate guidance or manipulation is used to achieve a prescribed value of a variable and which operates without human intervention.

Automatic controller: A device, or combination of devices, which measures the value of a variable, quantity, or condition and operates to correct or limit deviation of this measured value from a selected command (set-point) reference.

Bode diagram: A plot of log-gain and phase-angle values on a log-frequency base, for an element, loop, or output transfer function.

Capacitance: A property expressible by the ratio of the time integral of the flow rate of a quantity (heat, electric charge) to or from a storage, divided by the related potential charge.

Command: An input variable established by means external to, and independent of, the automatic-control system, which sets the ideal value of the controlled variable. See *set point.*

Control action: Of a control element or controlling system, the nature of the change of the output affected by the input.

Control action, derivative: That component of control action for which the output is proportional to the rate of change of input.

Control action, floating: A control system in which the rate of change of the manipulated variable is a continuous function of the actuating signal.

Control action, integral (reset): Control action in which the output is proportional to the time integral of the input.

Control action, proportional: Control action in which there is a continuous linear relationship between the output and the input.

Control system, sampling: Control using intermittently observed values of signals such as the feedback signal or the actuating signal.

Damping: The progressive reduction or suppression of the oscillation of a system.

Deviation: Any departure from a desired or expected value or pattern. Steady-state deviation is known as *offset.*

Disturbance: An undesired variable applied to a system which tends to affect adversely the value of the controlled variable.

Error: The difference between the indicated value and the accepted standard value.

Gain: For a linear system or element, the ratio of the change in output to the causal change in input.

Load: The material, force, torque, energy, or power applied to or removed from a system or element.

Nyquist diagram: A polar plot of the loop transfer function.

Nichols diagram: A plot of magnitude and phase contours using ordinates of logarithmic loop gain and abscissas of loop phase angle.

Offset: The steady-state deviation when the command is fixed.

Peak time: The time for the system output to reach its first maximum in responding to a disturbance.

Proportional band: The reciprocal of gain expressed as a percentage.

Resistance: An opposition to flow that results in dissipation of energy and limitation of flow.

Response time: The time for the output of an element or system to change from an initial value to a specified percentage of the steady state.

Rise time: The time for the output of a system to increase from a small specified percentage of the steady-state increment to a large specified percentage of the increment.

Self-regulation: The property of a process or a machine to settle out at equilibrium at a disturbance, without the intervention of a controller.

Set point: A fixed or constant command given to the controller designating the desired value of the controlled variable.

Settling time: The time required, after a disturbance, for the output to enter and remain within a specified narrow band centered on the steady-state value.

Time constant: The time required for the response of a first-order system to reach 63.2 percent of the total change when disturbed by a step function.

Transfer function: A mathematical statement of the influence which a system or element has on a signal or action compared at input and output terminals.

BASIC AUTOMATIC-CONTROL SYSTEM

A **closed-loop control system** consists of a process, a measurement of the controlled variable, and a controller which compares the actual measurement with the desired value and uses the difference between them to automatically adjust one of the inputs to the process. The **physical system** to be controlled can be electrical, thermal, hydraulic, pneumatic, gaseous, mechanical, or described by any other physical or chemical process. Generally, the control system will be designed to meet one of two **objectives.** A **servomechanism** is designed to follow changes in set point as closely as possible. Many electrical and mechanical control systems

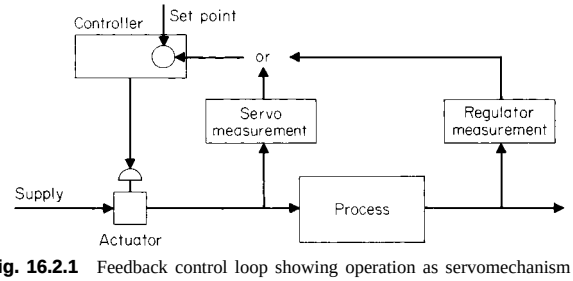


Fig. 16.2.1 Feedback control loop showing operation as servomechanism or regulator.

are servomechanisms. A **regulator** is designed to keep output constant despite changes in load or disturbances. Regulatory controls are widely used on chemical processes. Both objectives are shown in Fig. 16.2.1.

The **control components** can be actuated pneumatically, hydraulically, electronically, or digitally. Only in very few applications does actuation affect controllability. **Actuation** is chosen on the basis of economics.

The **purpose** of the control system must be defined. A large capacity or inertia will make the system sluggish for servo operation but will help to minimize the error for regulator operation.

PROCESS AS PART OF THE SYSTEM

Figure 16.2.1 shows the **process** to be part of the control system either as load on the servo or process to be controlled. Thus the process must be designed as part of the system. The process is **modeled** in terms of its static and dynamic gains in order that it be incorporated into the system diagram. Modeling uses Ohm's and Kirchhoff's laws for electrical systems, Newton's laws for mechanical systems, mass balances for fluid-flow systems, and energy balances for thermal systems.

Consider the **electrical system** in Fig. 16.2.2

$$i = \frac{E_1 - E_2}{R} \qquad i = C\frac{dE_2}{dt} \tag{16.2.1}$$

Combining

$$RC\frac{dE_2}{dt} + E_2 = E_1 \tag{16.2.2}$$

where $\left(R\,\frac{\text{volt-second}}{\text{coulomb}}\right)\left(C\,\frac{\text{coulomb}}{\text{volt}}\right) = \tau\,\text{s} = \text{time constant}$

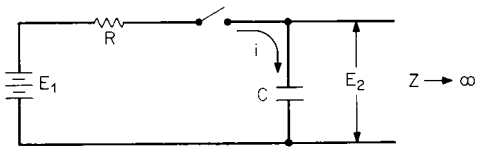


Fig. 16.2.2 Electrical system where current flows upon closing switch.

Consider the **mass balance** of the vessels shown in Figs. 16.2.3 and 16.2.4:

$$\text{Accumulation} = \text{input} - \text{output}$$
$$\frac{d(\rho V)}{dt} = w_1 - w_2 \quad \text{lb/min} \tag{16.23}$$

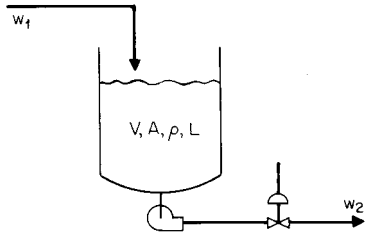


Fig. 16.2.3 Liquid level process.

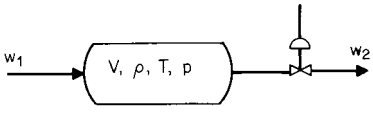


Fig. 16.2.4 Gas pressure process.

For the liquid level process (Fig. 16.2.3):

$$V = A\,L \tag{16.2.4}$$

Linearizing

$$w_2 = \frac{w_2}{x}\bigg|_{ss} x$$

Substituting

$$\rho A\frac{dL}{dt} = w_1 - \frac{w_2}{x}\bigg|_{ss} x \tag{16.25}$$

where ρA = analogous capacitance.

For the **gas pressure process** (Fig. 16.2.4):

$$V\frac{d\rho}{dt} = w_1 - w_2 \tag{16.2.6}$$

From thermodynamics

$$\left(\frac{\rho}{p}\right)^n = \text{constant} \tag{16.2.7}$$

Linearizing

$$w_2 = \frac{w_2}{x}\bigg|_{ss} x + \frac{w_2}{2(p - p_2)}\bigg|_{ss} p - p_2 \tag{16.2.8}$$

Substituting

$$RC\frac{dp}{dt} + p = \frac{2(p - p_2)}{x}\bigg|_{ss} x + p_2 \tag{16.2.9}$$

where $C = \dfrac{V}{nRT}$ $R = \dfrac{2(p - p_2)}{w_2}$ $RC = \tau\,\text{min}$

and the terms are defined: V = volume, ft^3 (m^3); A = cross-sectional area, ft^2 (m^2); L = level, ft (m); ρ = density, lb/ft^3 (g/ml); x = valve stem position (normalized 0 to 1); T = temperature, °F (°C); p = pressure, lb/in^2 (kPa); and w = mass flow, lb/min (kg/min).

The **thermal process** of Fig. 16.2.5 is modeled by a heat balance (Shinskey, ''Process Control Systems,'' McGraw-Hill):

$$Mc\frac{dT}{dt} = wc(T_0 - T) - UA(T - T_w) \tag{16.2.10}$$

$$\frac{Mc}{wc + UA}\frac{dT}{dt} + T = \frac{wc}{wc + UA}T_0 + \frac{UA}{wc + UA}T_w \tag{16.2.11}$$

where $C = Mc$ $R = \dfrac{1}{wc + UA}$ $RC = \tau\,\text{min}$

$$RC\frac{dT}{dt} + T = \frac{wc}{wc + UA}T_0 + \frac{UA}{wc + UA}T_w \tag{16.2.12}$$

The terms are defined: M = weight of process fluid in vessel, lb (kg); c = specific heat, Btu/lb · °F (J/kg · °C); U = overall heat-transfer coefficient, Btu/ft^2 · min · °F (W/m^2 · °C); and A = heat-transfer area, ft^2 (m^2).

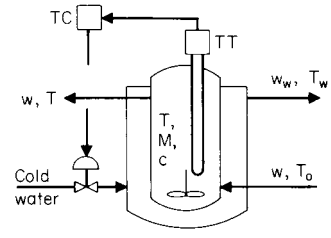


Fig. 16.2.5 Thermal process with heat from vessel being removed by cold water in jacket.

Newton's laws can be applied to the manometer shown in Fig. 16.2.6.

$$\text{Inertia force} = \text{restoring force} - \text{flow resistance}$$

$$\frac{Al\rho}{g_c}\frac{d^2h}{dt^2} = A\left(p - 2h\rho\frac{g}{g_c}\right) - RA\frac{dh}{dt} \qquad (16.2.13)$$

Flow resistance for laminar flow is given by the Hagen-Poiseuille equation:

$$R = \frac{\text{driving force}}{\text{rate of transfer}} = \frac{321\mu}{d^2 g_c} \qquad (16.2.14)$$

Substituting

$$\frac{1}{2g}\frac{d^2h}{dt^2} + \frac{161\mu}{\rho d g^2}\frac{dh}{dt} + h = h_i \qquad (16.2.15)$$

In standard form

$$\tau_c^2\frac{d^2h}{dt^2} + 2\tau_c\zeta\frac{dh}{dt} + h = h_i \qquad (16.2.16)$$

where τ_c = characteristic response time, min, = $1/[(60)(2\pi)\omega_n]$ where ω_n = natural frequency, Hz; and ζ = damping coefficient (ratio), dimensionless.

The variables τ_c and ζ are very valuable design aids since they define **system response** and **stability** in terms of system parameters.

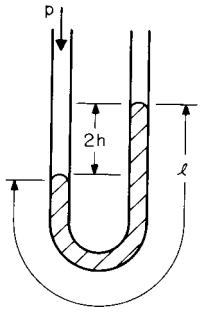


Fig. 16.2.6 Filled manometer measuring pressure *P*.

TRANSIENT ANALYSIS OF A CONTROL SYSTEM

The stability, accuracy, and speed of response of a control system are determined by analyzing the **steady-state** and the **transient** performance. It is desirable to achieve the steady state in the shortest possible time, while maintaining the output within specified limits. Steady-state performance is evaluated in terms of the accuracy with which the output is controlled for a specified input. The transient performance, i.e., the behavior of the output variable as the system changes from one steady-state condition to another, is evaluated in terms of such quantities as maximum overshoot, rise time, and response time (Fig. 16.2.7).

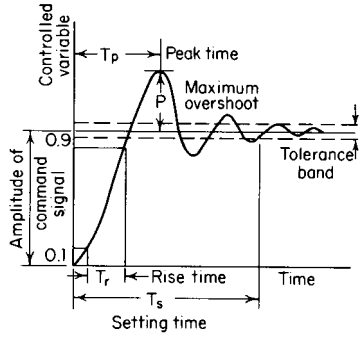


Fig. 16.2.7 System response to a unit step-function command.

Transient-Producing Disturbances A number of factors affect the **quality of control,** among them **disturbances** caused by set-point changes and process-load changes. Both set point and process load may be defined in terms of the setting of the final control element to maintain the controlled variable at the set point. Thus both cause the final control element to reposition. For a purely mechanical system the disturbance may take the form of a vibration, a displacement, a velocity, or an acceleration. A process-load change can be caused by variations in the rate of energy supply or variations in the rate at which work flows through the process. Reference to Fig. 16.2.5 and Eq. (16.2.12) shows disturbances to be variations in inlet process fluid temperature and cooling-water temperature. Further linearization would show variations in process flow and the overall heat-transfer coefficient to also be disturbances.

The Basic Closed-Loop Control To illustrate some characteristics of a basic closed-loop control, consider a mechanical, rotational system composed of a prime mover or motor, a total system inertia *J*, and a viscous friction *f*. To control the system's output variable θ_o, a command signal θ_i must be supplied, the output variable measured and compared to the input, and the resulting signal difference used to control the flow of energy to the load. The basic control system is represented schematically in Fig. 16.2.8.

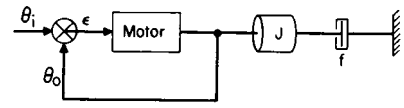


Fig. 16.2.8 A basic closed-loop control system.

The differential equation of this basic system is readily obtained from the idealized equations

$$\text{Load torque } T_L = J\frac{d^2\theta_o}{dt^2} + f\frac{d\theta_o}{dt} \qquad (16.2.17)$$

$$\text{Developed torque } T_D = K\varepsilon \qquad (16.2.18)$$

$$\text{Error } \varepsilon = \theta_i - \theta_o \qquad (16.2.19)$$

The above equations combine to yield the system differential equation:

$$J\frac{d^2\theta_o}{dt^2} + f\frac{d\theta_o}{dt} + K\theta_o = K\theta_i \qquad (16.2.20)$$

Step-Input Response of a Viscous-Damped Control If the control system described in Fig. 16.2.8 by Eq. (16.2.20) is subjected to a step change in the input variable θ_i, a solution $\theta_o = \theta_o(t)$ can be obtained as follows. (1) Let the ratio $\sqrt{K/J}$ be designated by the symbol ω_n and be called the **natural frequency.** (2) Let the quantity $2\sqrt{JK}$ be designated by the symbol f_c and be called the **friction coefficient** required for critical damping. (3) Let f/f_c be designated by the symbol ζ and be called the **damping ratio.** Equation (16.2.20) can then be written as

$$\frac{d^2\theta_o}{dt^2} + 2\zeta\omega_n\frac{d\theta_o}{dt} + \omega_n^2\theta_o = \omega_n^2\theta_i \qquad (16.2.21)$$

For $\theta_i = 1$:

$$\theta_o = 0 \qquad \text{and} \qquad \frac{d\theta_o}{dt} = 0 \text{ at } t = 0$$

The complete solution of Eq. (16.2.21) is

$$\theta_o = 1 - \frac{\exp(-\zeta\omega_n t)}{\sqrt{1-\zeta^2}} \sin\left(\sqrt{1-\zeta^2}\,\omega_n t + \tan^{-1}\frac{\sqrt{1-\zeta^2}}{\zeta}\right) \qquad (16.2.22)$$

Equation (16.2.22) is plotted in dimensionless form for various values of damping ratio in Fig. 16.2.9. The curves for $\zeta = 0.1$, 2, and 1 illustrate the underdamped, overdamped, and critically damped case, where any further decrease in system damping would result in overshoot.

Damping is a property of the system which opposes a change in the output variable.

The immediately apparent features of an observed transient performance are (1) the existence and magnitude of the maximum overshoot, (2) the frequency of the transient oscillation, and (3) the response time.

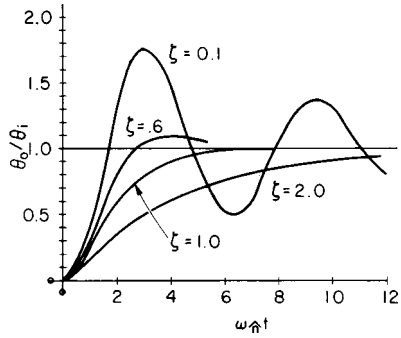


Fig. 16.2.9 Transient response of a second-order viscous-damped control to unit-step input displacement.

Maximum Overshoot When an automatic-control system is underdamped, the output variable overshoots its desired steady-state condition and a transient oscillation occurs. The first overshoot is the greatest, and it is the effect of its amplitude which must concern the control designer. The primary considerations for limiting this maximum overshoot are (1) to avoid damage to the process or machine due to excessive excursions of the controlled variable beyond that specified by the command signal, and (2) to avoid the excessive settling time associated with highly underdamped systems. Obviously, exact quantitative limits cannot generally be specified for the magnitude of this overshoot. However, experience indicates that satisfactory performance can generally be obtained if the overshoot is limited to 30 percent or less.

Transient Frequency An undamped system oscillates about the final steady-state condition with a frequency of oscillation which should be as high as possible in order to minimize the response time. The designer must, however, avoid resonance conditions where the frequency of the transient oscillation is near the natural frequency of the system or its component parts.

Rise Time T_n, Peak Overshoot P, Peak Time T_P These quantities are related to ζ and ω_n in Figs. 16.2.10 and 16.2.11. Some useful formulas are listed below:

$$\omega_n T_n \approx 1.02 + 0.48\zeta + 1.15\zeta^2 + 0.76\zeta^3 \qquad 0 \le \zeta \le 1$$

$$\omega_n T_s = \left.\begin{array}{ll} 17.6 - 19.2\zeta & 0.2 \le \zeta \le 0.75 \\ -3.8 + 9.4\zeta & 0.75 \le \zeta \le 1 \end{array}\right\} \text{2\% tolerance band}$$

$$P = \exp\left(\frac{-\pi\zeta}{\sqrt{1-\zeta^2}}\right) \qquad T_p = \frac{\pi}{\omega_n\sqrt{1-\zeta^2}}$$

Although these quantities are defined for a second-order system, they may be useful in the early design states of higher-order systems if the response of the higher-order system is dominated by roots of the characteristic equation near the imaginary axis.

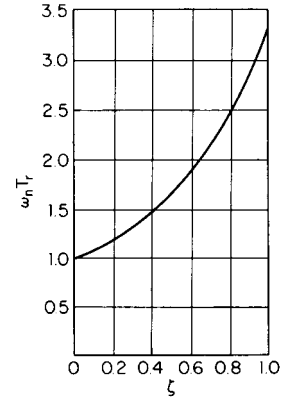


Fig. 16.2.10 Rise time T_r as a function of ζ and ω_n.

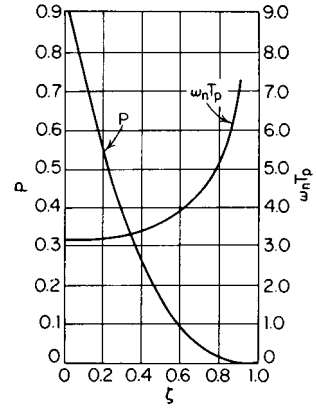


Fig. 16.2.11 Peak overshoot P and peak time T_p as functions of ζ and ω_n.

Derivative and Integral Compensation (Thaler) Four common compensation methods for improving the steady-state performance of a proportional-error control without damaging its transient response are shown in Fig. 16.2.12. They are (1) error derivative compensation, (2) input derivative compensation, (3) output derivative compensation, (4) error integral compensation.

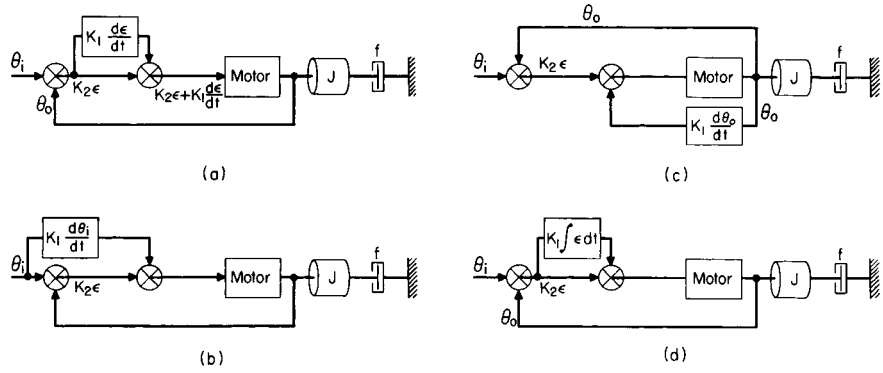


Fig. 16.2.12 Derivative and integral compensation of a basic closed-loop system. (*a*) Error derivative compensation; (*b*) input derivative compensation; (*c*) output derivative compensation; (*d*) error integral compensation. *(Thaler.)*

Error Derivative Compensation The torque equilibrium equation is

$$J\frac{d^2\theta_o}{dt^2} + f\frac{d\theta_o}{dt} = K_2\varepsilon + K_1\frac{d\varepsilon}{dt} \qquad (16.2.23)$$

Writing Eq. (16.2.23) in terms of the input and output variables yields

$$J\frac{d^2\theta_o}{dt^2} + (f + K_1)\frac{d\theta_o}{dt} + K_2\theta_o = K_2\theta_i + K_1\frac{d\theta_i}{dt} \qquad (16.2.24)$$

By adjusting K_1 and reducing f so that the quantity $f + K_1$ is equal to f in the uncompensated system, the system performance is affected as follows: (1) ε resulting from a constant-first-derivative input is reduced because of the reduction in viscous friction; (2) the transient performance of the uncompensated system is preserved unchanged.

Derivative Input Compensation The torque equilibrium equation is

$$J\frac{d^2\theta_o}{dt^2} + f\frac{d\theta_o}{dt} = K_2\varepsilon + K_1\frac{d\theta_i}{dt} \qquad (16.2.25)$$

Writing Eq. (16.2.25) in terms of the input and output variables yields

$$J\frac{d^2\theta_o}{dt^2} + f\frac{d^2\theta_o}{dt} + K_2\theta_o = \theta_i + K_1\frac{d\theta_i}{dt} \qquad (16.2.26)$$

Examination of Eq. (16.2.26) yields the following information about the compensated system's performance: (1) since the characteristic equation is unchanged from that of the uncompensated system, the transient performance is unaltered; (2) the steady-state solution to Eq. (16.2.26) is

$$\theta_o = \theta_i - \frac{f}{K_2}\left(1 - \frac{K_1}{K_2}\right)\frac{d\theta_i}{dt} \qquad (16.2.27)$$

Therefore the input derivative signal can reduce the steady-state error by adjusting K_1 to equal K_2.

Derivative Output Compensation The torque equilibrium equation is

$$J\frac{d^2\theta_o}{dt^2} + f\frac{d^2\theta_o}{dt} = K_2\varepsilon \pm K_1\frac{d\theta_o}{dt} \qquad (16.2.28)$$

Writing Eq. (16.2.28) in terms of the input and output variables yields

$$J\frac{d^2\theta_o}{dt^2} + (f \pm K_1)\frac{d\theta_o}{dt} + K_2\theta_o = K_2\theta_i \qquad (16.2.29)$$

Examination of Eq. (16.2.29) yields the following information about the compensated system's performance. (1) Output derivative feedback produces the same system effect as the viscous friction does. This compensation therefore damps the transient performance. (2) Under conditions where $\theta_i = ct$, the steady-state error is increased.

Error Integral Compensation Error integral compensation is used where it is necessary to eliminate steady-state errors resulting from input signals with constant first derivatives or under conditions of externally applied loads. The torque equilibrium equation is

$$J\frac{d^2\theta_o}{dt^2} + f\frac{d\theta_o}{dt} \pm \text{external load torque} = K_2\varepsilon + K_1\int_0^t \varepsilon\, dt \qquad (16.2.30)$$

Writing Eq. (16.2.30) in terms of the input variable and the error yields

$$\text{External load torque} + J\frac{d^2\theta_i}{dt^2} + f\frac{d\theta_i}{dt} = J\frac{d^2\varepsilon}{dt^2} + f\frac{d\varepsilon}{dt} + K_2\varepsilon + K_1\int_0^t \varepsilon\, dt \qquad (16.2.31)$$

At steady state for $t \gg 0$ and with a step change in the input derivative

$$\frac{d^2\theta_i}{dt^2} = \frac{d\varepsilon}{dt} = \frac{d^2\varepsilon}{dt^2} = 0 \qquad \frac{d\theta_i}{dt} = \text{const} \qquad (16.2.32)$$

Eq. (16.2.31) assumes the form

$$\pm\,\text{External load torque} + f\frac{d\theta_i}{dt} = K_2\varepsilon + K_1\int_0^t \varepsilon\, dt \qquad (16.2.33)$$

Since the sum of the load torque and the term $f(d\theta_i/dt)$ is finite, $\varepsilon = 0$ for

$$\int_0^\infty \varepsilon\, dt \to \infty \qquad (16.2.34)$$

If the integrating coefficient is a small number, additional torque produced by the integration action is developed very slowly, and, although the steady-state error is eventually eliminated, the transient performance is essentially unchanged. However, if K_1 is a large value, a large torque is produced in a short period of time, increasing the effect T/J ratio, and thereby decreasing the damping. The general effects of error integral compensation within its useful range are (1) steady-state error is eliminated; and (2) transient response is adversely effected, resulting in increased overshoot and the attendant increase in response time.

TIME CONSTANTS

The **time constant** τ, the **characteristic response time** (τ_c), and the **damping coefficient** are combined with operational calculus to design a control system without solving the classical differential equations. Note that the electrical, Eq. (16.2.2), mass, Eq. (16.2.9), and energy, Eq. (16.2.12) processes are all described by analogous first-order differential equations of the form

$$\tau\frac{dc}{dt} + c = Ka \qquad (16.2.35)$$

Solving with initial conditions equal to zero, the **time-domain response** to a step disturbance is

$$c = Ka(1 - e^{t/\tau}) \qquad (16.2.36)$$

Plotting is shown in Fig. 16.2.13.

Figure 16.2.13 shows the **time constant** to be defined by the point at 63.2 percent of the response, while response is essentially complete at three time constants (95 percent).

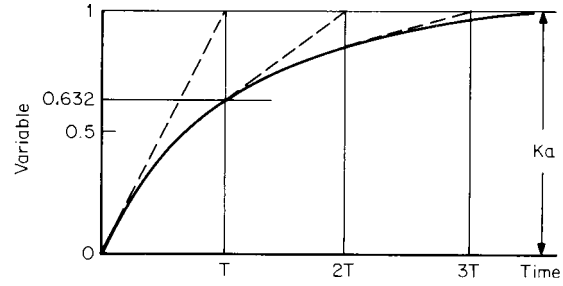


Fig. 16.2.13 Response of first-order system showing time constant relationship.

Operational calculus provides a systematic and simple method for handling linear differential equations with constant coefficients. The **Laplace operator** is used in control system analysis because of the straightforward transformation among the domains of interest:

Time domain	Complex domain	Frequency domain
$\frac{d}{dt}$	s	$j\omega$

Extensive tables of transforms are available for converting between the time and complex domains and back again. Transformation from the **complex** or **frequency domains** to **time** is difficult, and it is seldom attempted. Time-domain solutions are obtained only from computer simulations, which can solve the finite difference equations fast enough to serve as a design tool. Control system analysis and design proceed with a number of analytical and graphical techniques which do not require a time-domain solution.

The **root locus method** (Evans), is a graphical technique used in the complex domain which provides substantial insight into the system. It has the weaknesses of handling dead time poorly and of the graph's being very tedious to plot. It is used only when computerized plotting is available.

BLOCK DIAGRAMS

The **physical diagram** of the system is converted to a **block diagram** in order that the different components of the system (all the way from a steam boiler to a thermocouple) can be placed on a common mathematical footing for analysis as a system. The block diagram shows the functional relationship among the parts of the system by depicting the action of the variables in the system. The circle represents an algebraic function of addition. Each system component is represented by a **block** which has one input and one output. The block represents a dynamic function such that the output is a function of time and also of the input. The dynamic function is called a **transfer function**—the ratio of the Laplace transform of the output variable to the input variable with all initial conditions equal to zero. The input and output variables are considered as signals, and the blocks are connected by arrows to show the flow of information in the system.

Rewriting Eq. (16.2.12) for the thermal system,

$$RC\frac{dT}{dt} + T = K_1T_0 + K_2T_w \tag{16.2.37}$$

Transforming

$$(\tau s + 1)T = K_1T_0 + K_2T_w \tag{16.2.38}$$

for which the block diagram is shown in Fig. 16.2.14. Two conditions are specified: (1) the components must be described by linear differential equations (or nonlinear equations linearized by suitable approximations), and (2) each block is unilateral. What occurs in one component may not affect the components preceding.

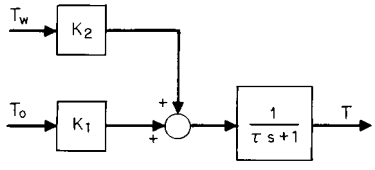


Fig. 16.2.14 Block diagram of thermal process.

Block-Diagram Algebra

The block diagram of a single-loop feedback-control system subjected to a command input $R(s)$ and a disturbance $U(s)$ is shown in Fig. 16.2.15.

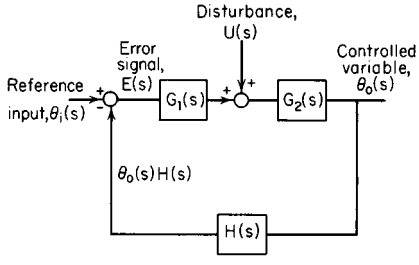


Fig. 16.2.15 Single-loop feedback control system.

When $U(s) = 0$ and the input is a reference change, the system may be reduced as follows:

$$E(s) = -\theta_o(s)H(s)$$
$$\theta_o(s) = E(s)[G_1(s)G_2(s)] \tag{16.2.39}$$

Therefore

$$\frac{\theta_o(s)}{\theta_i(s)} = \frac{G_1(s)G(s)}{1 + G_1(s)G_2(s)H(s)}$$

When $\theta_i(s) = 0$ and the input is a disturbance, the system may be reduced as follows:

$$E(s) = -\theta_o(s)H(s)$$
$$[E(s)\,G_1(s) + U(s)]G_2(s) = \theta_o(s)$$
$$\frac{\theta_o(s)}{U(s)} = \frac{G_2(s)}{1 + G_1(s)G_2(s)H(s)} \tag{16.2.40}$$

Closed-loop transfer function Eq. (16.2.40) can be determined by observation from

$$\frac{\theta_o(s)}{U(s)} = \frac{\text{feedforward functions}}{1 + \text{complete-loop functions}}$$

The denominator of Eq. (16.2.40) is the **characteristic equation** which determines **stability.** Most systems are designed on the basis of the characteristic equation since it sets both response time τ_c and damping ζ. The equation is undefined (unstable) if $G_1(s)G_2(s)H(s)$ equals -1. But -1 is a vector of magnitude 1 and a phase of $-180°$. This fact is used to determine stable parameter adjustments in the graphical techniques to be discussed later.

The input disturbance $[U(s)]$ can be any time function in actual operation. The **step input** is widely used for analysis and testing since it is easily implemented; it results in a simple Laplace transform; it is the most severe type of disturbance; and the response to a step change shows the maximum error that could occur.

In many complex control systems, especially in the nonmechanical process-control field, auxiliary feedback paths are provided in order to adjust the system's performance. Figure 16.2.16*a* illustrates such a condition. In analyzing such a system it is usually best to combine secondary loops into the main control loop to form an equivalent series block and transfer function. The system of Fig. 16.2.16*a* might be reduced in the following sequence.

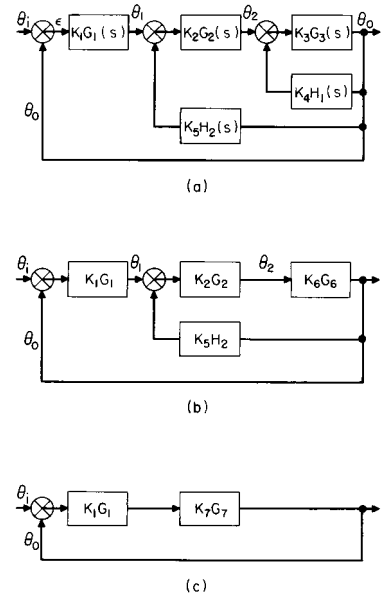


Fig. 16.2.16 Reduction of a closed-loop control system with multiple secondary loops.

1. Replace $K_3G_3(s)$ and $K_4H_1(s)$ with a single equivalent element

$$\frac{\theta_o}{\theta_2} = \frac{K_3G_3(s)}{1 + K_4H_1(s)K_3G_3(s)} = K_6G_6 \tag{16.2.41}$$

The result of this first reduction is shown in Fig. 16.2.16*b*:

2. Figure 16.2.16*b* can be treated in a similar fashion and a single block used to replace K_2G_2, K_6G_6, and K_5H_2

$$\frac{\theta_o}{\theta_1} = \frac{K_2G_2K_6G_6(s)}{1 + K_5H_2K_2G_2K_6G_6(s)} = K_7G_7 \tag{16.2.42}$$

The result of this second reduction is shown in Fig. 16.2.16*c*. The resulting open-loop transfer function is

$$\theta_o/\varepsilon = K_1G_1K_7G_7(s) \tag{16.2.43}$$

The closed-loop or frequency response function is

$$\frac{\theta_o}{\theta_i} = \frac{K_1G_1K_7G_7(s)}{1 + K_1G_1K_7G_7(s)} \tag{16.2.44}$$

Equation (16.2.44) can, of course, be expanded to include the terms of the system's secondary loops.

SIGNAL-FLOW REPRESENTATION

An alternate graphical representation of the mathematical relationships is the signal-flow graph. For complicated systems it allows a more compact representation and more rapid reduction techniques than the block diagram.

In Fig. 16.2.17, the nodes represent the variables θ_i, ε, θ_1, . . . , θ_o, and the branches the relationships between the nodes, of the system shown in Fig. 16.2.16. For example,

$$\theta_1(s) = \varepsilon K_1G_1(s) - K_5H_2(s)\theta_o(s) \tag{16.2.45a}$$

and

$$\varepsilon = \theta_i - \theta_o \tag{16.2.45b}$$

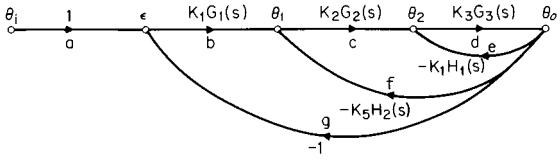


Fig. 16.2.17 Signal flow graph of the closed-loop control system shown in Fig. 16.2.16.

Signal-flow terminology follows:

Source: node having only outgoing branches, for example, θ_i
Sink: node having only incoming branches, θ_o
Path: series of branches with the same sense of direction, for example, *abcd, cdf*
Forward path: path originating at a source and ending at a sink, with no node encountered more than once, for example, *abcd*
Path gain: product of the coefficients along a path, for example, 1 $[K_1G_2(s)][K_2G_2(s)][K_3G_3(s)]$
Feedback loop: path starting at a node and ending at the same node, for example, *bcdg*
Loop gain: product of coefficients along a feedback loop, for example, $[K_1G_2(s)][K_2G_2(s)][K_3G_3(s)]$ (-1)

The overall gain of the system can be calculated from

$$G = \frac{\Sigma_i G_i\Delta_i}{\Delta} \tag{16.2.46}$$

where G_i = gain of *i*th forward path and

$$\Delta = 1 - \Sigma L_1 + \Sigma L_2 - \Sigma L_3 + \cdots + (-1)^k\Sigma L_k$$

where ΣL_1 = sum of the gains of each forward loop; ΣL_2 = sum of products of loop gains for nontouching loops (no node is common), taken two at a time; ΣL_3 = sum of products of loop gains for nontouching loops taken three at a time; Δ_i = value of Δ for signal flow graph resulting when *i*th path is removed.

From Fig. 16.2.17 there is only one forward path, *abcd.*

$$\therefore G_1 = K_1G_1(s)K_2G_2(s)K_3G_3(s)$$

Closed loops are *de, cdf,* and *bcdg,* with gains $-K_1H_1(s)K_3G_3(s)$, $-K_2G_2(s)K_3G_3(s)K_5H_2(s)$, and $-K_1G_1(s)K_2G_2(s)K_3G_3(s)$.

There are no nontouching closed loops;

$$\therefore \Delta = 1 + K_1H_1(s)K_3G_3(s) + K_2G_2(s)K_3G_3(s)K_5H_2(s) + K_1G_1(s)K_2G_2(s)K_3G_3(s)$$

There are no loops remaining if the forward path *abcd* is removed;

$$\therefore \Delta_1 = 1$$

Thus

$$\theta_o/\theta_1 = K_1G_1(s)K_2G_2(s)K_3G_3/[1 + K_1H_1(s)K_3G_3(s) + K_2G_2(s)K_3G_3(s)K_5H_2(s) + K_1G_1(s)K_2G_2(s)K_3G_3(s)] \tag{16.2.47}$$

which is identical with Eq. (16.2.44).

CONTROLLER MECHANISMS

The **controller** modifies the error signal in a desired manner to produce an output pressure which is used to actuate the valve motor. The several controller modes used singly or in combination are (1) the proportional mode in which $P_{out}(t) = K_cE(t)$, (2) the integral mode, in which $P_{out}(t) = 1/T_1\int E(t)\,dt$, and (3) the rate mode, in which $P_{out}(t) = T_2dE(t)/dt$. In these expressions $P_{out}(t)$ = controller output pressure, $E(t)$ = input error signal, K_c = proportional gain, $1/T_1$ = reset rate, and T_2 = rate time.

A **pneumatic controller** consists of an **error-detecting mechanism, control modes** made up of proportional (*P*), integral (*I*), and derivative (*D*) actions in almost any combination, and a **pneumatic amplifier** to provide output capacity. The error-detecting mechanism is a differential link, one end of which is positioned by the signal proportional to the controlled variable, and the other end of which is positioned to correspond to the command set point. The proportional action is provided by a flapper nozzle (Fig. 16.2.18), where the flapper is positioned by the error signal. A motion of 0.0015 in by the flapper is sufficient for nearly full output range. Nozzle back pressure is inversely proportional to the distance between nozzle opening and flapper.

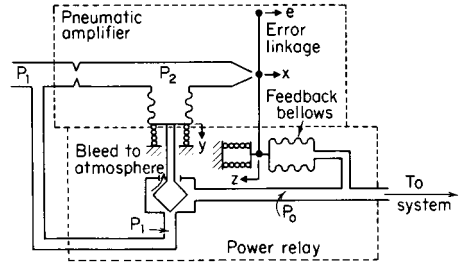


Fig. 16.2.18 Gain reduction of pneumatic amplifier by means of feedback bellows. *(Raven, ''Automatic Control Engineering,'' McGraw-Hill.)*

The controller employs a power-amplifying pilot for providing a larger quantity of air than could be provided through the small restriction shown in Fig. 16.2.18. The nozzle back pressure, instead of operating the final control element directly, is transmitted to a bellows chamber where it positions the pilot valve.

The combination of flapper-nozzle amplifier and power relay shown in Fig. 16.2.18 has a very high gain since small flapper displacements can result in the output traversing the full range of output pressure. Negative feedback is employed to reduce the gain. Controller output is connected to a feedback bellows which operates to reposition the flapper. With the feedback bellows, a movement of the flapper toward the nozzle increases back pressure, causing output pressure to decrease and the feedback bellows to move the flapper away from the nozzle. Thus the mechanism is stabilized. Fig. 16.2.19 shows controller gain being varied by adjusting the point at which the feedback bellows bears on the flapper.

The **rate mode,** called ''anticipatory,'' can take large corrective action when errors are small but have a high rate of change. The mode resists not only departures from the set point but also returns and so provides a stabilizing action. Since the rate mode cannot control to a set point, it is not used alone. When used with the proportional mode, its stabilizing influence (90° phase lead; see Table 16.2.3) may allow an increase in gain K_c and a consequent decrease in steady-state error.

Proportional-derivative (PD) control is obtained by adding an adjustable feedback restriction as shown in Fig. 16.2.20. This results in delayed negative feedback. The restriction delays and reduces the feedback, and, since the feedback is negative, the output pressure is momentarily higher and leads, instead of lags, the error signal.

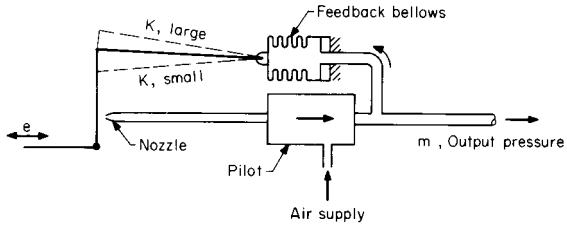


Fig. 16.2.19 Proportional controller with negative feedback. *(Reproduced by permission. Copyright© John Wiley & Sons, Inc. Publishers, 1958. From D. P. Eckman, ''Automatic Process Control.'')*

Since the proportional mode requires an error signal to change output pressure, set point and load changes in a proportionally controlled system are accompanied by a steady-state error inversely proportional to the gain. For systems which, because of stability considerations, cannot tolerate high gains, the integral mode added to the proportional will

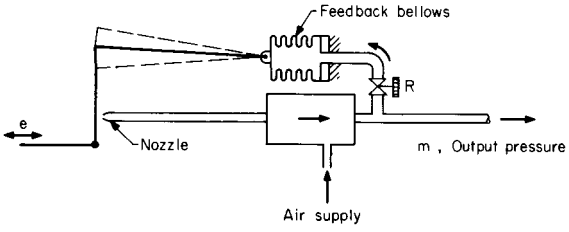


Fig. 16.2.20 Proportional-derivative controller. *(Reproduced by permission. Copyright© John Wiley & Sons, Inc. Publishers, 1958. From D. P. Eckman, ''Automatic Process Control.'')*

eliminate the steady-state error since the output from this mode is continually varying so long as an error exists. The addition of the integral mode to a proportional controller has an adverse effect on the relative stability of the control because of the 90° phase lag introduced.

Proportional-integral (PI) control is obtained by adding a positive feedback bellows and an adjustable restriction (Fig. 16.2.21). The addition of the positive feedback bellows cancels the gain reduction brought about by the negative feedback bellows at a rate determined by the adjustable restriction.

Electronic controllers which are analogous to the pneumatic controller have been developed. They have the advantages of elimination of time lags; compatability with computers; being less expensive to install (although more expensive to purchase); being more energy efficient; and being immune to low temperatures (water in pneumatic lines freezes).

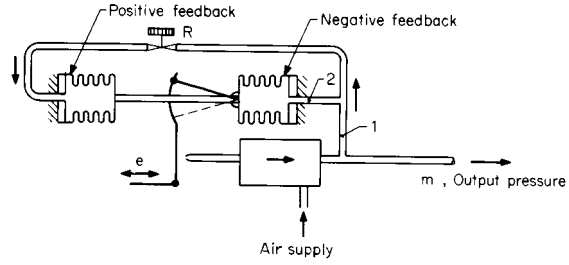


Fig. 16.2.21 Proportional-integral controller. *(Reproduced by permission. Copyright© John Wiley & Sons, Inc. Publishers, 1958. From D. P. Eckman, ''Automatic Process Control.'')*

The heart of the electronic controller is the **high-gain operational amplifier** with passive elements at the input and in the feedback. Figure 16.2.22 shows the classical control elements, though circuits in commercially available controllers are far more sophisticated. The current drawn by the dc amplifier is negligible, and the amplifier gain is around 1,000, so that the junction point is essentially at ground potential. The reset amplifier has delayed negative feedback, similar to the pneumatic circuit in Fig. 16.2.21. The derivative amplifier has advanced negative feedback, similar to Fig. 16.2.20. Gains are adjusted by changing the ratios of resistors or capacitors, while adjustable time constants are made up of a variable resistor and a capacitor (RC).

Regardless of actuation, the commonly applied **controller modes** are as follows:

Designation	*Transform*	*Symbol*
Proportional	K (gain)	P
Reset	$1 + \dfrac{1}{\tau_i s}$	I
Derivative	$\dfrac{\tau_d s + 1}{(t_d/\alpha)s + 1}$	D
Floating	$\dfrac{1}{\tau_F s}$	F

The modes are combined as needed into P, PI, PID, PD, and F. Choice depends on the application and can be made from the Bode diagram. The block diagram for a PID controller is shown in Fig. 16.2.23.

The **derivative mode** can be placed on the measured variable signal (as shown), on the error, or on the controller output. The arrangement of

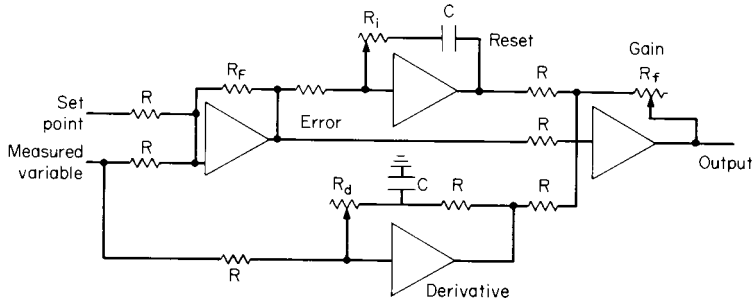


Fig. 16.2.22 Simplified circuit for a typical electronic controller.

Fig. 16.2.23 is preferred for regulation since the derivative does not affect set-point changes, and the derivative acts to reduce overshoot on start-up (Zoss, ''Applied Instrumentation in the Process Industries,'' Vol. IV, Gulf Publishing Co.). Regardless of a sequence, the block

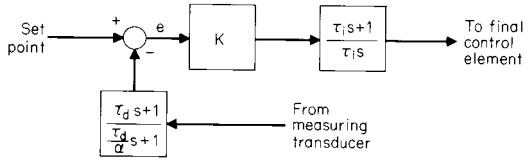


Fig. 16.2.23 Block diagram of PID controller.

diagram of a PID controller is commonly drawn as shown in Fig. 16.2.24.

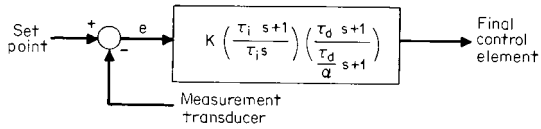


Fig. 16.2.24 Combined controller block diagram.

FINAL CONTROL ELEMENTS

The **final control element** is a mechanism which alters the value of the **manipulated variable** in response to an output signal from the automatic-control device. It will typically receive a signal from the controller and manipulate a flow of material or energy to the process. The final control element can be a control valve, an electrical motor, a servovalve, or a damper. Servovalves are discussed in the next section, ''Hydraulic-Control Systems.''

The final control element often consists of two parts: first, an **actuator** which translates the controller signal into a command for the manipulating device, and, second, a **mechanism** to adjust the manipulated variable. Figure 16.2.25 shows a control valve made up of an actuator and an adjustable orifice.

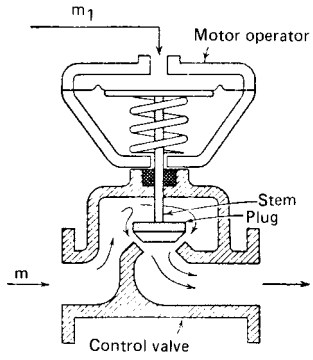


Fig. 16.2.25 Control valve and pneumatic actuator. *(Reproduced by permission. Copyright© John Wiley & Sons, Inc. Publishers, 1958. From D. P. Eckman, ''Automatic Process Control.'')*

The most commonly used **actuator** is a diaphragm motor in which the output pressure from the controller is counteracted not only by the spring but also by fluid forces in the valve body. The latter may cause serious deviation from linear static behavior with deleterious control effects. High friction at the valve stem or large unbalanced fluid forces at the plug can be overcome by valve positioners, which are essentially proportional controllers.

A **control valve** in liquid service is described by Eq. (16.2.48):

$$q = C_v \sqrt{\frac{\Delta p}{G}} \tag{16.2.48}$$

And C_v is described by its relationship to stem lift x, which is called the **valve-flow characteristic.**

The linear characteristic is

$$C_v = C_v|_{\max}\, x \tag{16.2.49}$$

and the equal-percentage characteristic is

$$C_v = C_v|_{\max}\, (r)^{x-1} \tag{16.2.50}$$

where q = flow, gpm (m^3/min); $C_v|_{\max}$ = valve sizing coefficient; Δp = pressure drop across valve, lb/in^2 (kPa); G = specific gravity (density, g/ml); r = valve rangeability; and x = stem position, normalized to 0 to 1.

The flow characteristic relationship is not necessarily the lift-flow characteristic of the valve when installed since the valve is but one component in a piping system in which pressure drops vary with the flow rate. The change in characteristic as less percentage of total system drop is taken across the control valve is shown in Fig. 16.2.26.

Valve characteristics are generally selected according to their ability to compensate for nonlinearities in the system. Nonlinearities typically represent a change in gain which in turn changes the character of the control response for a given controller setting when load- or set-point

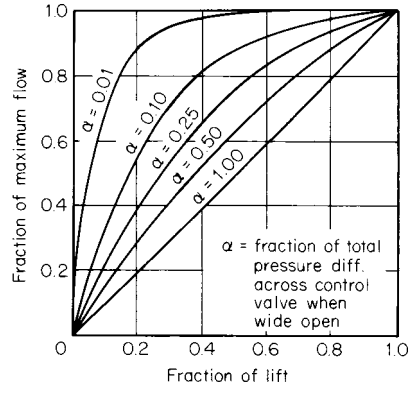


Fig. 16.2.26 Control valve linear flow characteristic. *(Reproduced by permission. Copyright© John Wiley & Sons, Inc. Publishers, 1958. From D. P. Eckman, ''Automatic Process Control.'')*

changes occur or when there is a variable overall pressure drop across the system. Equal-percentage valves, for example, tend to control over widely varying operating conditions, since the change in flow is always proportional to flow rate. The selection of the proper valve therefore depends on study of the particular system.

HYDRAULIC-CONTROL SYSTEMS

Hydraulic systems are used for rapid-response servomechanisms at high power levels. Operating system pressures are from 50 to 100 lb/in^2 for slower-acting systems and up to 5,000 lb/in^2 where lightweight and fast responses are required. Compared with **electrical systems** the major advantages are a rapid response in the large horsepower ranges and the capability of operating at high power-density levels since the fluid can transmit dissipated energy from the point of generation. Compared with **pneumatic systems,** hydraulic systems are faster because the fluid is essentially incompressible. Major disadvantages are vulnerability to dirt, since the components generally require close machining tolerances, and the danger of fire and explosion resulting from the flammability of the hydraulic fluids used (Blackburn, Reethof, and Shearer, ''Fluid Power Control,'' Wiley).

The direction and volume of flow are controlled by **servo valves** in the

system. They may be single-stage (pilot-operated) and mechanically or electrically actuated. A schematic of a spool-type four-way single-stage control piston and inertia load is shown in Fig. 16.2.27. Hydraulic fluid at constant pressure enters at the supply port. With displacement of the spool valve downward, for example, inflow to the top side of the piston

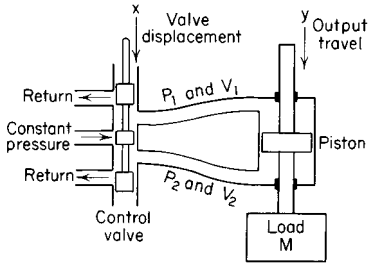


Fig. 16.2.27 Four-way valve-piston circuit. *(Truxal, ''Control Engineer's Handbook,'' McGraw-Hill.)*

moves the piston downward. Because of machining tolerances the spool dimensions are either large (overlapped) or smaller (underlapped) than the port dimension. Underlapped valves permit leakage to the piston in the centered position; overlapped valves result in a dead zone, where motion x results in no flow until a port is opened.

The transfer function of this circuit is given as (Truxal, ''Control Engineers' Handbook'')

$$\frac{y}{x} = \frac{C_1 \dfrac{1}{1 + \alpha(C_1/C_2)}}{s\left[\dfrac{VM}{2BA^2}\dfrac{1}{1 + \alpha(C_1/C_2)}s^2 + \dfrac{C_1 m/C_2 + V\alpha/2BA^2}{1 + \alpha(C_1/C_2)}s + 1\right]}$$

where C_1 = servo velocity gradient, in/(s)(in), C_2 = servo force gradient, lb/in, α = viscous friction of load and piston, lb/(in)(s), B = bulk modulus of fluid, lb/in^2 M = mass of load and piston, lb/(in)(s^2), A = piston area, in^2, and V = effective entrained fluid volume, in^3 (one-half of total entrained volume between valve and piston).

The velocity of the output is proportional to the input resulting in a **velocity-control** servo. To convert this system to a **position-control** servo, mechanical, hydraulic, or electrical feedback may be employed. A valve-piston position servo with **mechanical feedback** is shown in Fig. 16.2.28. Any difference between the input D and the piston position y causes a motion x, which causes the piston to move in a direction opposite to D, that is, in a direction to reduce x. The lever ratio establishes the relationship between y and D.

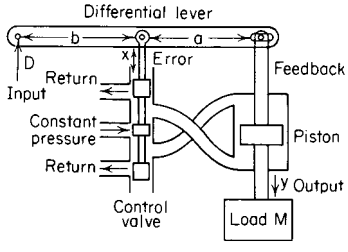


Fig. 16.2.28 Valve-piston position servomechanical feedback. *(Truxal, ''Control Engineer's Handbook,'' McGraw-Hill.)*

Most commercially available servo valves are two stages, permitting electrohydraulic action. The pilot stage can be operated by a low-power, short-travel electrical device, with a concomitant increase in flexibility. A typical pilot-operated servo valve is shown in Fig. 16.2.29. In this case the pilot is a double-flapper valve rather than a spool valve. (In general, small, accurate low-leakage spool valves are costly.) Upward movement of the flapper by the actuating motor results in increased pressure to the right end of the power spool. Hydraulic feedback occurs because of the increased flow across restrictor *a*. The power spool moves to the left until the unbalanced pressure is matched by the spring resistance. The disadvantage of this valve is the continual leakage flow through the flapper nozzle, but the torque motor has a low-power requirement and is inexpensive.

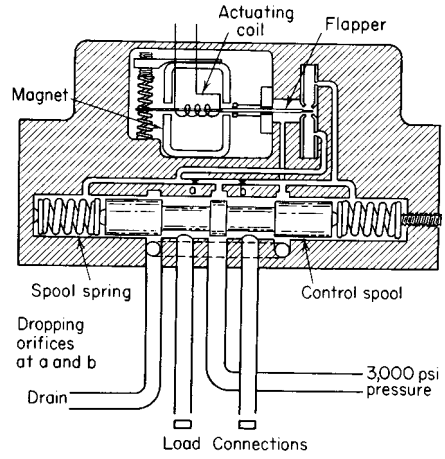


Fig. 16.2.29 Two-stage electrohydraulic servo valve. The first-stage is a four-way flapper valve with a calibrated pressure output driving a second-stage, spring-loaded, four-way spool valve. *(Moog Servocontrols, Inc.)*

The major disadvantages of spool-type valves are (1) high cost, because of high-tolerance requirements between the valve lands, (2) high static friction and inertia, and (3) susceptibility to dirt. The **flapper** valve is less expensive to manufacture than a spool valve of equivalent characteristics and is not so susceptible to damage by dirt particles. In Fig. 16.2.30, P_1, the supply pressure, is constant. Input motion of the flapper toward the nozzle increases P_2 and drives the piston toward the right. The steady-state characteristic P_2 vs. x is shown in Fig. 16.2.31.

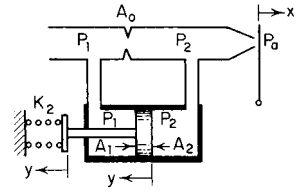


Fig. 16.2.30 Flapper valve. *(Raven.)*

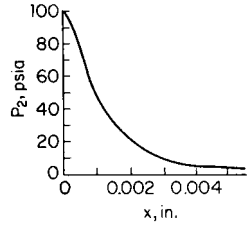


Fig. 16.2.31 Equilibrium curve of P_2 versus x for a flapper valve. *(Raven, ''Automatic Control Engineering,'' McGraw-Hill.)*

STEADY-STATE PERFORMANCE

The steady-state error of a control system can be determined by using the final value theorem (see Laplace Transforms, Sec. 2) in which

$$\lim_{t\to\infty} \theta_o(t) = \lim_{s\to 0} s\theta_o(s) \quad \text{provided } \theta_o(t) \text{ is stable}$$

The classification of control systems according to the form of the open-loop transfer function facilitates the determination of the steady-state errors when the system is subjected to various inputs.

The open-loop transfer function $\theta_o(s)/\varepsilon(s) = KG(s)$ may be written

$$KG(s) = \frac{\theta_o(s)}{\varepsilon(s)} = \frac{K(1+\tau_a s)(1+\tau_a s)(1+\tau_c s)\cdots}{s^N(1+\tau_1 s)(1+\tau_2 s)(1+\tau_3 s)\cdots} \quad (16.2.51a)$$

The **system type** is given according to the value of N as

Type 0 system	$N = 0$
Type 1 system	$N = 1$
Type 2 system	$N = 2$
Type 3 system	$N = 3$

(16.2.51b)

Error coefficients, based on a system with unity feedback [$H(s) = 1$], are defined as

$$\text{Positional error constant} = K_o = \lim_{s\to 0} KG(s) \quad (16.2.52a)$$

$$\text{Velocity error constant} = K_v = \lim_{s\to 0} sKG(s) \quad (16.2.52b)$$

$$\text{Acceleration error constant} = K_a = \lim_{s\to 0} s^2KG(s) \quad (16.2.52c)$$

A summary of error coefficients for systems of different types is given in Table 16.2.1.

Table 16.2.1 Summary of Error Coefficients

N	K_o	K_v	K_a
0	const	0	0
1	∞	const	0
2	∞	∞	const

A summary of the errors for types 0, 1, and 2 systems, when subjected to various inputs, is given in Table 16.2.2.

Table 16.2.2 Summary of Errors

Input error	$\theta_i(t) = A$ ε_0/A	$\theta_i(t) = vt$ ε_v/v	$\theta_i(t) = at^2$ ε_a/a
$N = 0$	$1/(1 + K_0)$	∞	∞
1	0	$1/K_v$	∞
2	0	0	$1/K_a$

The higher the system type, the better is the output able to follow the higher degrees of input. Higher-type systems, however, are more difficult to stabilize, and a compromise must be made between the steady-state error and the settling time of the response.

CLOSED-LOOP BLOCK DIAGRAM

The **transfer functions** of the process, controller, and final control element can be combined with the measuring transducer into a **block diagram** of the complete control loop. Consider the thermal system of Fig. 16.2.5 and the heat balance on the vessel jacket:

$$M_j c \frac{dT_w}{dt} = w_w c(T_{w1} - T_w) \quad (16.2.53)$$

Linearizing and rearranging:

$$\frac{M_j}{w_w} T_w s + T_w = \left[\frac{(T_{w1} - T_w)}{w_w}\right]\left[\frac{w_w}{x}\right] x \quad (16.2.54)$$

where M_j = weight of cooling water in jacket, lb (kg); c = specific heat of water, Btu/lb·°F (J/kg·°C); T = temperature, °F; x = valve stem position (normalized 0 to 1); and w_w = water flow, lb/min (kg/min).

For simplicity, let the jacket time constant M_j/w_w be so small as to be negligible (not typically the case). The block diagram of Fig. 16.2.32 then represents process diagram Fig. 16.2.5. The **gain** K in each case is an incremental change in output divided by an incremental change in input. For example, if the temperature transmitter has a temperature span of 100°F and an output of 4 to 20 mA, the gain would be

$$K_\tau = \frac{20 - 4}{100} = \frac{16}{100} = 0.16 \text{ mA/°F}$$

The **open-loop gain** is defined as the product of all the gains in the control loop:

$$K_o = K_c K_v K_T \left(\frac{T_{w1} - T_w}{x}\right)\left(\frac{UA}{wc + UA}\right)$$

Combining the **closed-loop system** of Fig. 16.2.32 results in the closed-loop equation

$$T = \frac{\left(\frac{wc}{wc + UA}\right)\left(\frac{1}{\tau s + 1}\right) T_0}{1 + \frac{K_o(\tau_i s + 1)(\tau_d s + 1)}{\tau_i s(\tau s + 1)\left(\frac{\tau_d}{\alpha} s + 1\right)(\tau_T s + 1)}} \quad (16.2.55)$$

The time-dependent modes on the controller have been designed to compensate for the dynamics of the rest of the loop. Let

$$\tau_i = \tau \quad \text{and} \quad \tau_d = \tau_T$$

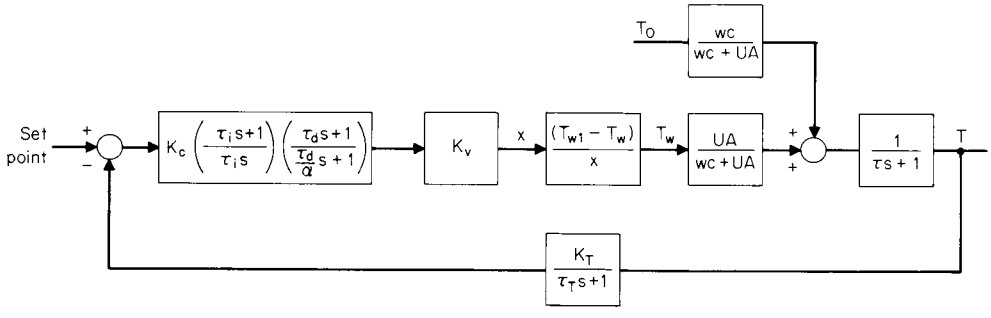


Fig. 16.2.32 Block diagram of a thermal system.

Table 16.2.3 Process Characteristics versus Mode of Control

Number of process capacities	Process reaction rate	Process time lags		Load changes		Suitable mode of control
		Resistance-capacitance (RC)	Dead time (transportation)	Size	Speed	
Single	Slow	Moderate to large	Small	Any	Any	Two-position. Two-position with differential gap
				Moderate	Slow	Multiposition. Proportional input
Single (self-regulating)	Fast	Small	Small	Any	Slow	Floating modes: Single-speed Multispeed
					Moderate	Proportional-speed floating
Multiple	Slow to moderate	Moderate	Small	Small	Moderate	Proportional position
Multiple	Moderate	Any	Small	Small	Any	Proportional plus rate
Multiple	Any	Any	Small to moderate	Large	Slow to moderate	Proportional plus reset
Multiple	Any	Any	Small	Large	Fast	Proportional plus reset plus rate
Any	Faster than that of the control system	Small or nearly zero	Small to moderate	Any	Any	Wideband proportional plus fast reset

SOURCE: Considine, ''Process Instruments and Controls Handbook,'' McGraw-Hill.

After cancellation, Eq. (16.2.55) becomes

$$T = \frac{\left(\frac{wc}{wc + UA}\right) T_0}{(\tau s + 1)\left[1 + \frac{K_c}{\tau_i s\left(\frac{\tau_d}{\alpha} s + 1\right)}\right]} \tag{16.2.56}$$

Equation (16.2.56) can be multiplied out to become

$$T = \frac{\left(\frac{wc}{wc + UA}\right)(\tau_i s)\left(\frac{\tau_d}{\alpha} s + 1\right) T_0}{(\tau s + 1)\left(\frac{\tau_i \tau_d}{\alpha} s^2 + \tau_i s + K_c\right)} \tag{16.2.57}$$

Equation (16.2.57) can be written in the standard form of Eqs. (16.2.16) and (16.2.21):

$$T = \frac{\left(\frac{wc}{wc + UA}\right)(\tau_i s)\left(\frac{\tau_d}{\alpha} s + 1\right) T_0}{K_c(\tau s + 1)(\tau_c^2 s^2 + 2\tau_c \zeta s + 1)} \tag{16.2.58}$$

Controller gain K_c can then be chosen to give the best response τ_c with stability ζ. More rigorous techniques for determining controller parameter values are shown under Bode diagrams and Routh's criterion. Table 16.2.3 provides preliminary guidance for the selection of control modes.

FREQUENCY RESPONSE

Although it is the time response of the control system that is of major importance, study of the effect on transient response of changes in system parameters, either in the process or controller, is more conveniently made from a **frequency-response** analysis of the system. The **frequency response** of a system is the steady-state output of the system to input sinusoids of varying frequency. The output for a linear system can be completely described in terms of the amplitude ratio of the output sinusoid to the input sinusoid. The amplitude ratio or gain, and phase, are functions of the frequency of the input sinusoid.

The use of **sinusoidal methods** to analyze and test dynamic systems has gained widespread popularity because system response is obtained easily from the response of the individual elements, no matter how many elements are included. By contrast, transient analysis is quite tedious, with only three dynamic components in the system, and is too difficult to be worthwhile for four or more components.

Frequency-response analysis provides quantitative information concerning the system on maximum gain K_o for stability, time response τ_c, ω_n, and the controller parameter adjustments τ_i, τ_d. In most cases, this information is adequate for assurance of stability, for comparing alternatives, or for judging the merits of a proposed control system. System parameters are generally obtained from open-loop frequency response, which is the response with the loop broken between the controller and the final control element (Fig. 16.2.32).

Frequency response also lends itself to system identification by testing for systems not readily amenable to mathematical analysis by subjecting the system to input sinusoids of varying frequency.

The shift from the **complex domain** to the **frequency domain** (linear systems, zero initial conditions) is accomplished by simply substituting $j\omega$ for s. Consider Eq. (16.2.38).

$$(\tau s + 1)T = K_1 T_0 \tag{16.2.59}$$

Substituting

$$\frac{T}{T_0} = \frac{K_1}{\tau j\omega + 1}$$

Multiply by the complex conjugate

$$\frac{T}{T_0} = \frac{K_1}{\tau j\omega + 1} \frac{\tau j\omega - 1}{\tau j\omega - 1}$$

Collecting terms

$$\frac{T}{T_0} = K_1 \left[\frac{1}{\tau^2\omega^2 + 1} - j \frac{\tau\omega}{\tau^2\omega^2 + 1}\right] \tag{16.2.60}$$

The first term is the real part of the solution, and the second term, the imaginary part. These terms define a vector, shown in Fig. 16.2.33, which has magnitude M and phase angle θ as determined either graphically or by the Pythagorean theorem.

$$M = \left[\frac{1}{\tau^2\omega^2 + 1}\right]^{1/2} \tag{16.2.61}$$

$$\theta = \tan^{-1}\left(\frac{-\zeta\omega}{1}\right) \tag{16.2.62}$$

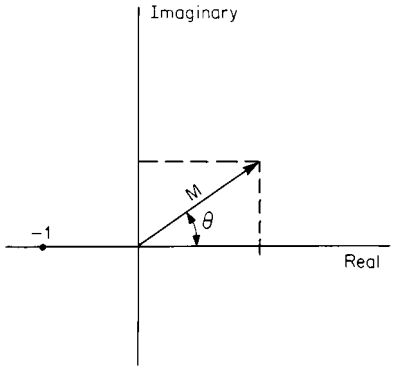


Fig. 16.2.33 Polar plot showing vector.

GRAPHICAL DISPLAY OF FREQUENCY RESPONSE

Sinusoidal response is **plotted** in three different ways: (1) the rectangular plot with the log of the amplitude versus the log of the frequency, called a **Bode plot;** (2) a phase-margin plot with magnitude shown versus a function of phase with frequency as a parameter, called a **Nichols plot;** and (3) a polar plot with magnitude and phase shown in vector form with frequency as a parameter, called a **Nyquist plot.** The Nichols plot is actually a plot of phase margin ($180° - \theta$) versus frequency and is used to define system performance. The polar plot enables the absolute stability of the system to be determined without the need for obtaining the roots of the characteristic equation. The logarithmic (Bode) plot has the advantage of ease in plotting, especially in design, since the individual effects of cascaded elements can be gaged by superposition. All the graphical procedures use the ''minus 1'' point [discussed with Eq. (16.2.38)] as the criterion for stability (magnitude = 1, phase = −180°).

NYQUIST PLOT

The **Nyquist diagram** is a graphical method of determining stability. The diagram is essentially a mapping into the $G(s)$ plane of a contour enclosing the major portion of the right half of the s plane. Figure 16.2.34 shows the Nyquist plot of a typical system.

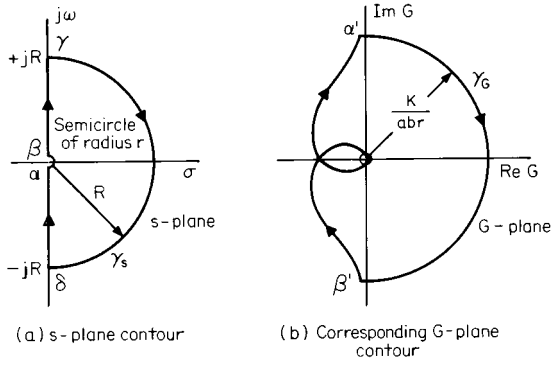


Fig. 16.2.34 Nyquist diagram for $G(s) = 1/[s(s + a)(s + b)]$. *(Truxal, ''Control System Synthesis,'' McGraw-Hill.)*

The frequency response of a closed-loop system can also be derived from the **polar plot** of the direct transfer function $KG(s)$. Systems with dynamic elements in the feedback can be transformed into direct systems by block-diagram manipulation. In Fig. 16.2.35, the amplitude ratio θ_o/θ_i at any frequency ω is the ratio of the lengths of the vectors $O\beta$ and $\alpha\beta$. The angle formed by the vectors $\alpha\beta$ and $O\beta$ is the phase angle of the frequency response $\angle\,\theta_o/\theta_i(j\omega)$. The numerator of the transfer function is a constant representing the closed-loop gain. Changing the gain proportionately changes the length of the vector $O\beta$. Figure 16.2.36 shows another typical Nyquist plot.

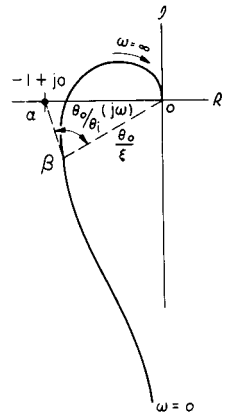


Fig. 16.2.35 Typical $(\theta_o/\varepsilon)/(j\omega)$ plot.

The Nyquist diagram is widely used in the design of electrical systems, but plots are very tedious to make and are generated via computer programs.

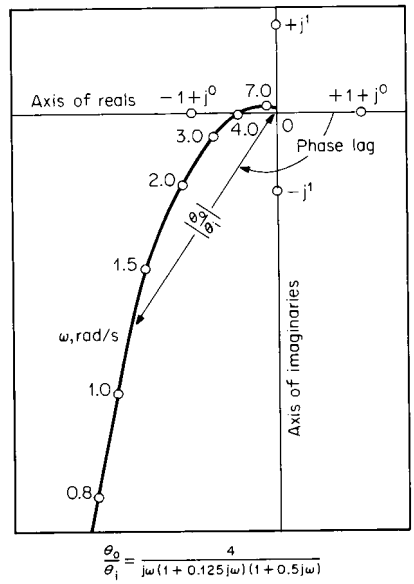


Fig. 16.2.36 Typical Nyquist diagram showing loop transfer function $|\theta_o/\theta_i|$. *(ASME, ''Terminology for Automatic Control,'' Pub. ASA C85.1-1963.)*

BODE DIAGRAM

The most common method of presenting the **response data at various frequencies** is to use the log-log plot for amplitude ratios, accompanied by a semilog plot for phase angles. These rectangular plots are called **Bode diagrams,** after H. W. Bode, who did basic work on the theory of feedback amplifiers. A major advantage of this method is that the numerical computation of points on the curve is simplified by the fact that log magnitude versus log frequency can be approximated to engineering accuracy with straight-line approximations. In the **sinusoidal testing** of a

system, the system dynamics will be faster than a very low frequency test signal, and the output amplitude has time to recover fully to the input amplitude. Thus the magnitude ratio is 1. Conversely, at high frequency, the system does not have time to reach equilibrium, and output amplitude will decrease. Magnitude ratio goes to a very small value under this condition. Thus the system will show a decrease in magnitude ratio as frequency increases. By similar reasoning, the difference between input and output phase angles becomes progressively more negative as frequency increases.

Consider the equations for magnitude, Eq. (16.2.61), and phase, Eq. (16.2.62). If $\tau_w \ll 1$,

$$M = \left(\frac{1}{1}\right)^{1/2} = 1 \qquad \theta = \tan^{-1}(0) = 0$$

and if $\tau_w \gg 1$,

$$M = \left(\frac{1}{\tau^2\omega^2}\right)^{1/2} = \frac{1}{\tau\omega} \approx 0 \qquad \theta = \tan^{-1}\left(\frac{-\tau_w}{1}\right) \approx -90°$$

The two **straight-line asymptotes** of magnitude, if extended, would intersect at $\tau\omega = 1$. Thus a relationship between frequency and time constant is established. The point at which $\tau\omega = 1$ is called the *corner frequency*, or the *break frequency*. The asymptote past the break frequency is easily plotted from a point at the break frequency $\omega = 1/\tau$ and a point at 10 times the frequency and 0.1 times the magnitude ratio.

The frequency response of systems containing different dynamic elements may be calculated by suitably employing the magnitude and phase characteristics of each element separately. Combined magnitudes are calculated

$$|M_1M_2| = |M_1| \times |M_2|$$

Or, on the log-log Bode diagram,

$$\log|M_1M_2| = \log|M_1| + \log|M_2|$$

The logarithmic scale permits **multiplication** by adding vertical distances. For phase

$$\angle\theta_1\theta_2 = \angle\theta_1 + \angle\theta_2$$

Thus **phase addition** is also made by adding vertical distances. Figure 16.2.37 shows the combination of the straight-line asymptotes for the following transfer function:

$$G(s) = \frac{1}{\tau s(\tau s + 1)} \tag{16.2.63}$$

Table 16.2.4 lists additional frequency-response relationships for system elements. The noninteracting controller, item 8, has the advantage of no interaction among tuning parameters but has the disadvantages of being more difficult to plot on the Bode diagram and of not offering the direct compensation of the controller shown in Fig. 16.2.24.

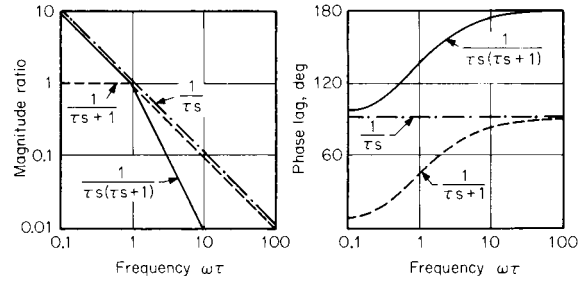


Fig. 16.2.37 Combined transfer functions. *(Reproduced by permission. Copyright© John Wiley & Sons, Inc. Publishers, 1958. From D. P. Eckman, ''Automatic Process Control.'')*

Figure 16.2.38 shows a typical combined Bode diagram. The ordinate has a scale in decibels in addition to the magnitude ratio scale. Decibels (dB) are sometimes used to provide a linear scale. They are defined

$$\text{dB} = -20 \log M$$

The discontinuity at the break frequency due to the meeting of the straight-line asymptotes will not be found in actual test data. Instead, the corner will be rounded, as shown dotted in Fig. 16.2.39. The error due to the approximation is some 3 dB at the break frequency, as shown in the figure.

Quadratic terms (Table 16.2.4) also lend themselves to asymptotic approximation, but the form of the magnitude and phase plots around the break frequency depends on the damping coefficient ζ. The slope of the magnitude ratio attenuation past the break point is twice that of the first-order system.

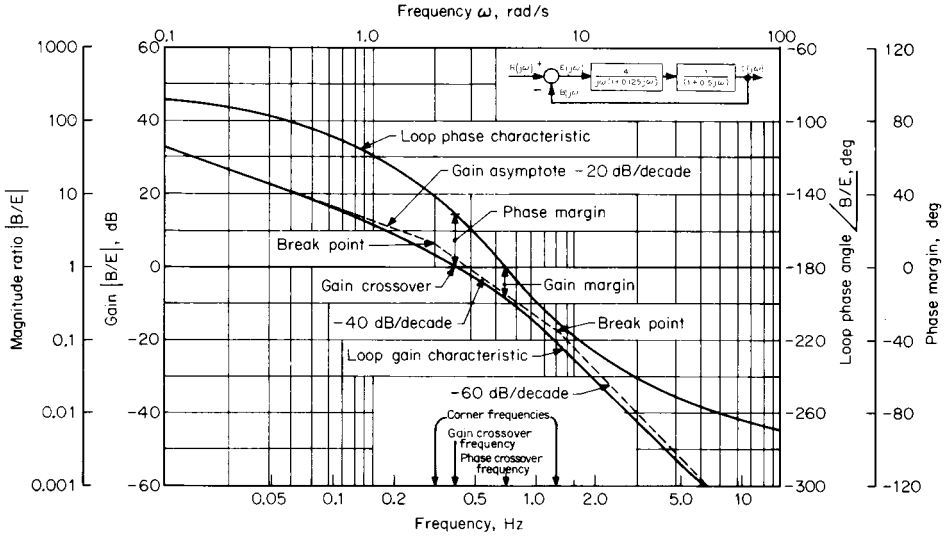


Fig. 16.2.38 Typical Bode diagram showing loop-transfer function *B/E. (ASME, ''Terminology for Automatic Control,'' Pub. ASA C85.1-1963.)*

Table 16.2.4 Frequency-Response Equations for Some Common Control-System Elements

Description	Transfer function $G(s)$	Frequency response $G(j\omega)$	Magnitude ratio	Phase angle
1. Dead time	$\varepsilon^{-T_L s}$	$\varepsilon^{-j\omega T_L}$	1	$-\omega T_L$ radians
2. First-order lag	$\dfrac{1}{Ts+1}$	$\dfrac{1}{j\omega T+1}$	$\dfrac{1}{\sqrt{\omega^2T^2+1}}$	$-\tan^{-1}(\omega T)$
3. Second-order lag	$\dfrac{1}{(Ts+1)(aTs+1)}$	$\dfrac{1}{-a\omega^2T^2+j(1+a)\omega T+1}$	$\dfrac{1}{\sqrt{(1-a\omega^2T^2)^2+(1+a)^2\omega^2T^2}}$	$-\tan^{-1}\left[\dfrac{(1+a)\omega T}{1-aT^2\omega^2}\right]$
4. Quadratic (underdamped)	$\dfrac{1}{\left(\dfrac{s}{\omega_n}\right)^2+\dfrac{2\zeta}{\omega_n}s+1}$	$\dfrac{1}{-\left(\dfrac{\omega}{\omega_n}\right)^2+j2\zeta\dfrac{\omega}{\omega_n}+1}$	$\dfrac{1}{\sqrt{\left(1-\dfrac{\omega^2}{\omega_n^2}\right)^2+4\zeta^2\left(\dfrac{\omega}{\omega_n}\right)^2}}$	$-\tan^{-1}\left[\dfrac{2\zeta\dfrac{\omega}{\omega_n}}{1-\left(\dfrac{\omega}{\omega_n}\right)^2}\right]$
5. Ideal proportional controller	K	K	K	0
6. Ideal proportional-plus-reset controller $T_i=\dfrac{1}{r}$ r = reset rate	$K\left(1+\dfrac{1}{T_i s}\right)$ or $K\dfrac{T_i s+1}{T_i s}$	$K\left(1+\dfrac{1}{j\omega T_i}\right)$ or $K\dfrac{j\omega T_i+1}{j\omega T_i}$	$K\sqrt{1+\left(\dfrac{1}{\omega T_i}\right)^2}$	$-\tan^{-1}\left(\dfrac{1}{\omega T_i}\right)$
7. Ideal proportional-plus-rate controller	$K(1+T_d s)$	$K(1+j\omega T_d)$	$K\sqrt{1+\omega^2T_d^2}$	$\tan^{-1}(\omega T_d)$
8. Ideal proportional-plus-reset-plus-rate controller	$K\left(1+T_d s+\dfrac{1}{T_i s}\right)$	$K\left(1+j\omega T_d+\dfrac{1}{j\omega T_i}\right)$ or $K\dfrac{j\omega T_i-\omega^2T_dT_i+1}{j\omega T_i}$	$K\sqrt{(\omega T_i)^2+(1-\omega^2T_dT_i)^2}$	$\tan^{-1}\left(\omega T_d-\dfrac{1}{\omega T_i}\right)$

SOURCE: Considine, ''Process Instruments and Controls Handbook,'' McGraw-Hill.

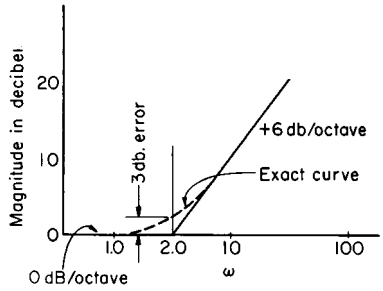


Fig. 16.2.39 Bode plot of term $(j\omega\tau_a + 1)$.

CONTROLLERS ON THE BODE PLOT

Controller modes, gain, reset, and derivative are plotted individually (in the form of Fig. 16.2.24) on the Bode diagram and summed as required. The functions for reset (integral) and derivative modes are moved horizontally on the diagram to determine proper parameter adjustment. The combined magnitude ratio curve for the whole system is moved vertically to determine open-loop gain.

After all the dynamic elements of the system are plotted and combined, reset and derivative are set in accordance with the following **criteria:**

Reset: The reset phase curve is positioned horizontally so that it contributes 10° of phase lag where the system phase lag is 170°.

Derivative: The derivative phase curve is positioned horizontally so that it contributes 60° of phase lead where the system phase lag is 180°.

The magnitude ratio curves are positioned with the phase-angle curves, and parameter values for reset and derivative can be read at the break points.

STABILITY AND PERFORMANCE OF AN AUTOMATIC CONTROL

An automatic-control system is stable if the amplitude of transient oscillations decreases with time and the system reaches a steady state. The stability of a system can be evaluated by examining the roots of the differential equation describing the system. The presence of positive real roots or complex roots with positive real parts dictates an unstable system. Any stability test utilizing the open-loop transfer function or its plot must utilize this fact as the basis of the test.

The Nyquist Stability Criterion The $KG(j\omega)$ locus for a typical single-loop automatic-control system plotted for all positive and negative frequencies is shown in Fig. 16.2.40. The locus for negative values of ω is the mirror image of the positive ω locus in the real axis. To complete the diagram, a semicircle (or full circle if the locus approaches $-\infty$ on the real axis) of infinite radius is assumed to connect in a positive sense, the + locus at $\omega \to 0$ with the negative locus at $\omega \to -0$. If this locus is traced in a positive sense from $\omega \to \infty$ to $\omega \to 0$, around the circle at ∞, and then along the negative-frequency locus the following may be concluded: (1) if the locus **does not** enclose the $-1 + j0$ point, the system is stable; (2) if the locus **does** enclose the $-1 + j0$ point, the system is unstable. The Nyquist criterion can also be applied to the log magnitude of $KG(j\omega)$ and phase-vs.-log ω diagrams. In this method of display, the criterion for stability reduces to the requirement that the log magnitude of $KG(j\omega)$ must cross the 0-dB axis at a frequency less than the frequency at which the phase curve crosses the $-180°$ line. Two stability conditions are illustrated in Fig. 16.2.41. The Nyquist criterion not only provides a simple test for the stability of an automatic-control system but also indicates the degree of stability of the system by indicating the degree to which $KG(j\omega)$ locus avoids the $-1 + j0$ point.

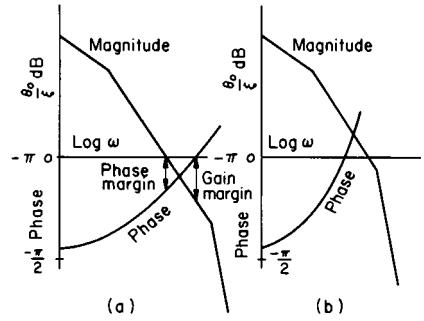


Fig. 16.2.41 Nyquist stability criterion in terms of log magnitude $KG(j\omega)$ diagrams. (*a*) Stable; (*b*) unstable. *(Porter, ''An Introduction to Servomechanism,'' Wiley.)*

The concepts of **phase margin** and **gain margin** are employed to give this quantitative indication of the degree of stability of an automatic-control system. **Phase margin** is defined as the additional negative phase shift necessary to make the phase angle of the transfer function $-180°$ at the frequency where the magnitude of the $KG(j\omega)$ vector is unity. Physically, phase margin can be interpreted as the amount by which the unity KG vector has to be shifted to make a stable system unstable.

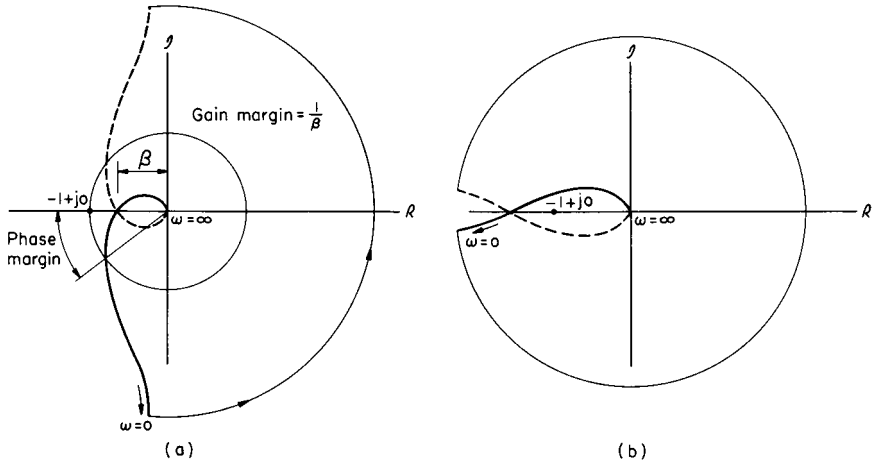


Fig. 16.2.40 Typical $KG(j\omega)$ loci illustrating application of Nyquist's stability criterion. (*a*) Stable; (*b*) unstable.

In a similar manner, **gain margin** is defined as the reciprocal of the magnitude of the KG vector at $-180°$. Physically, gain margin is the number by which the gain must be multiplied to put the system to the limit of stability. Thaler suggests satisfactory results can be obtained in most control applications if the phase margin is between 40 and 60° while the gain margin is between 3 and 10 (10 to 20 dB). These values will ensure a small transient overshoot with a single cycle in the transient. The margin concepts are qualitatively illustrated in Figs. 16.2.40 and 16.2.41.

Routh's Stability Criterion The frequency-response equation of a closed-loop automatic control is

$$\frac{\theta_o}{\theta_i} = \frac{KG(j\omega)}{1 + KG(j\omega)} \tag{16.2.64}$$

The characteristic equation obtained therefrom has the algebraic form

$$A(j\omega)^n + B(j\omega)^{n-1} + C(j\omega)^{n-2} + \cdots = 0 \tag{16.2.65}$$

The purpose of Routh's method is to determine the existence of roots of this equation which are positive or which are complex with positive real parts and thus identify the resulting instability. To apply the criterion the coefficients are written alternately in two rows as

$$\begin{array}{cccc} A & C & E & G \\ B & D & F & H \end{array}$$

This array is then expanded to

$$\begin{array}{cccc} A & C & E & G \\ B & D & F & H \\ \alpha_1 & \alpha_2 & \alpha_3 & \\ \beta_1 & \beta_2 & \beta_3 & \\ \gamma_1 & \gamma_2 & & \end{array}$$

where α_1, α_2, α_3, β_1, β_2, β_3, γ_1 and γ_2 are computed as

$$\alpha_1 = \frac{BC - AD}{B} \qquad \beta_1 = \frac{D\alpha_1 - B\alpha_2}{\alpha_1} \qquad \gamma_1 = \frac{\beta_1\alpha_2 - \alpha_1\beta_2}{\beta_1}$$

$$\alpha_2 = \frac{BE - AF}{B} \qquad \beta_2 = \frac{F\alpha_1 - B\alpha_3}{\alpha_1} \qquad \gamma_2 = \frac{\beta_1\alpha_3 - \alpha_1\beta_3}{\beta_1}$$

$$\alpha_3 = \frac{BG - AH}{B} \qquad \beta_3 = \frac{H\alpha_1 - B_o}{\alpha_1}$$

When the array has been computed, the left-hand column (A, B, α_1, β_1, γ_1) is examined. If the signs of all the numbers in the left-hand column are the same, there are no positive real roots. If there are changes in sign, the number of positive real roots is equal to the number of changes in sign. It should be recognized that this is a test for instability; the absence of sign changes does not guarantee stability.

SAMPLED-DATA CONTROL SYSTEMS

Definition Sampled-data control systems are those in which continuous information is transformed at one or more points of the control system into a series of pulses. This transformation may be performed intentionally, e.g., the flow of information over long distances to preserve the accuracy of the data during the transmission, or it may be inherent in the generation of the information flow, e.g., radiating energy from a radar antenna which is in the form of a train of pulses, or the signals developed by a digital computer during a direct digital control of machine-tool operation.

Methods of analysis analogous to those for continuous-data systems have been developed for the sampled-data systems. Discussed herein are (1) sampling, (2) the z transformation, (3) the z-transfer function, and (4) stability of sampled data systems.

Sampling The ideal sampler is a simple switch (Fig. 16.2.42) which is closed only instantaneously and opens and closes at a constant frequency. The switch, which may or may not be a physical component in a sampled-data feedback system, indicates a sampled signal. Such a sampled-data feedback system is shown in Fig. 16.2.43. The error signal in continuous form is $\varepsilon(t)$, and the sampled error signal is $\varepsilon^*(t)$. Figure 16.2.42 shows the relationship between these signals in a graphical form.

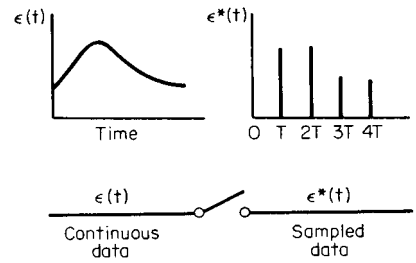


Fig. 16.2.42 Ideal sampler, showing continuous input and sampled output.

z Transformation In the analysis of continuous-data systems, it has been shown that the Laplace transformation can be used to reduce ordinary differential equations to algebraic equations. For sampled-data systems, an operational calculus, the z transform, can be used to simplify the analysis of such systems.

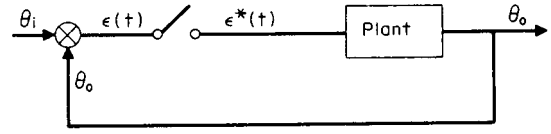


Fig. 16.2.43 Sampled data feedback system.

Consider the sampler as an impulse modulator; i.e., the sampling modulates an infinite train of unit impulses with the continuous-data variable. Then the Laplace transformation can be shown to be

$$F^*(s) = \sum_{n=0}^{\infty} f(nT)e^{-nTs} \tag{16.2.66}$$

(See A. W. Langill, ''Automatic Control Systems Engineering.'')

In terms of the z transform, this becomes

$$F^*(s) = F(z) = \sum_{n=0}^{\infty} f(nT)z^{-n} \tag{16.2.67}$$

where $z^{-n} = e^{-nTs}$.

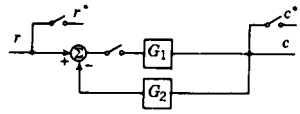
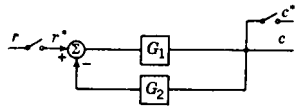
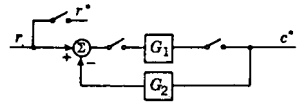
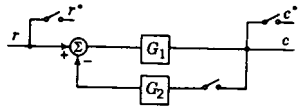
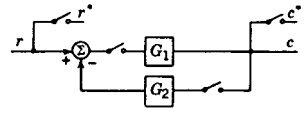
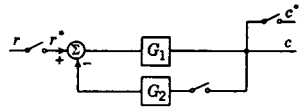
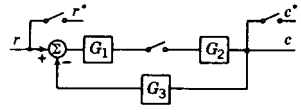
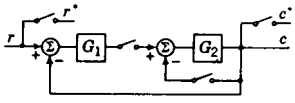
Tables in Sec. 2 list the Laplace transforms for a number of continuous time functions. Table 16.2.5 lists the z transforms for some of these functions.

It should be noted that unlike the Laplace transforms, the z-transform method imposes the restriction that the sampled-data-system response can be determined only at the sampling instant. The same z transform may apply to different time functions which may have the same value at

Table 16.2.5 Laplace and z Transforms

Time function	$f(t)$	$F(s) = \mathcal{L}[f(t)]$	$F^*(z) = F(z)$
Unit ramp function	t	$\frac{1}{s^2}$	$\frac{Tz}{(z-1)^2}$
Unit acceleration function	$\frac{t^2}{2}$	$\frac{1}{s^3}$	$\frac{1}{2}T^2\frac{z(z+1)}{(z-1)^3}$
Exponential function	$e^{-\alpha t}$	$\frac{1}{s+\alpha}$	$\frac{z}{z-e^{-\alpha T}}$
Sinusoidal function	$\sin \beta t$	$\frac{\beta}{s^2+\beta^2}$	$\frac{z \sin \beta T}{z^2 - 2z\cos \beta T + 1}$
Cosinusoidal function	$\cos \beta t$	$\frac{s}{s^2+\beta^2}$	$\frac{z(z-\cos \beta T)}{z^2 - 2z\cos \beta T + 1}$

Table 16.2.6 Typical Block Diagrams of Sampled-Data Control Systems and Their Transforms

System	Laplace transform of the output, $C(s)$	z-transform of the output, $C^*(z)$
1. 	$\frac{G_1(s)}{1 + G_1G_2^*(z)} R^*(z)$	$\frac{G_1^*(z)}{1 + G_1G_2^*(z)} R^*(z)$
2. 	$\frac{G_1(s)}{1 + G_1(s)G_2(s)} R^*(z)$	$\left[\frac{G_1(s)}{1 + G_1(s)G_2(s)}\right]^* R^*(z)$
3. 		$\frac{G_1^*(z)}{1 + G_1^*(z)G_2^*(z)} R^*(z)$
4. 	$G_1(s)\left[R(s) - G_2(s)\frac{RG_1^*(z)}{1 + G_1G_2^*(z)}\right]$	$\frac{RG_1^*(s)}{1 + G_1G_2^*(s)}$
5. 	$G_1(s)\left[R^*(z) - \frac{G_1^*(z)G_2^*(z)\,R^*(z)}{1 + G_1^*(z)G_2^*(z)}\right]$	$\frac{G_1^*(z)}{1 + G_1^*(z)G_2^*(z)} R^*(z)$
6. 	$G_1(s)\left[R^*(z) - G_2(s)\frac{R^*(z)G_1^*(z)}{1 + G_1G_2^*(z)}\right]$	$\frac{G_1^*(z)}{1 + G_1G_2^*(z)} R^*(z)$
7. 	$\frac{G_2(s)}{1 + G_1G_2G_3^*(z)} RG_1^*(z)$	$\frac{G_2^*(z)}{1 + G_1G_2G_3^*(z)} RG_1^*(z)$
8. 	$\frac{G_2(s)}{1 + G_2^*(z) + G_1G_2^*(z)} RG_1^*(z)$	$\frac{G_2^*(z)}{1 + G_2^*(z) + G_1G_2^*(z)} RG_1^*(z)$

SOURCE: John G. Truxal (ed.), ''Control Engineers' Handbook,'' McGraw-Hill.

the instant of sampling. The z-transform function is not defined, therefore, in a continuous sense, and the inverse z transform is not unique.

z Transfer Function The ratio of the sampled output function of a discrete network to the sampled input function is the z-transfer function. A discrete network is one which has both a sampled input and output. A table of block diagrams of a number of sampled-data control systems with their associated transfer functions is presented in Table 16.2.6.

Stability of Sampled-Data Systems The stability of sampled-data systems can be demonstrated utilizing frequency-response methods which have been discussed in this section. See ''Stability and Performance of an Automatic Control.'' The Nyquist stability criterion again applies with the same conclusions relative to the -1 point, and the methods of generalizing the open- and closed-loop frequency response plots remain the same.

MODERN CONTROL TECHNIQUES

Many engineers and researchers have been actively pursuing the application and development of advanced control strategies. In recent years this effort has included extensive work in the area of robust control theory. One of the primary attractions of this theory is that it generalizes the **single input–single system output (SISO)** concept of gain and phase margins, and its effect on system sensitivity to multivariable control system.

In this section the design technique for a **linear quadratic gaussian with loop transfer recovery (LQG/LTR)** robust controller design method will be summarized. The concepts required to understand this method will be reviewed. A linearized model of a simple process has been chosen to illustrate this technique. The simple process has three state variables, one input, and one output. Three control system design methods are compared: LQG, a LQG/LTR, and a proportional plus integral controller (PI).

Previous Work and Results

Most applications of modern control techniques to process control have appeared in the form of **optimal control strategies** which were in the form of full-state feedback control with and without state estimation [linear quadratic regulator (LQR)]. The regulator problem, coupled with a state estimation problem, was shown not to have desirable robustness properties. Doyle and Stein (Doyle and Stein, 1979) proposed the multivari-

able robust design philosophy known as LQG/LTR to eliminate the shortcomings of the LQR and LQG design techniques. The LQG/LTR design technique has been applied to control submersible vehicles, engine speed, helicopters, ship steering, and large, flexible space structures.

The LQG/LTR method requires the plant to be **minimum phase,** which cannot be guaranteed in a plant process. Typically **nonminimum phase** conditions in a plant process are due to time delay. The LQG/LTR design procedure offers no guarantees when a nonminimum-phase plant is used. However recent work develops a method to obtain a robust controller design for plants with time delay (Murphy and Bailey, IEEE 1989 and IECON 1990).

The LQG/LTR synthesis method for minimum phase plants achieves desired loop shapes and maximum robustness properties in the design of feedback control systems. The LQG/LTR design procedure uses **singular value analysis** and design procedures to obtain the desired performance and stability robustness. The use of singular values yields a frequency-domain design and analysis method generally expected by control engineers to determine system robustness. The LQG/LTR synthesis method is applicable to both **multiple-input/multiple-output (MIMO)** and SISO control systems. The performance limitations of this design method are analogous to those of a SISO system.

Definition of Robustness

A system is considered to be robust and have **good robustness properties** if it has a large stability margin, good disturbance attenuation, and/or low sensitivity to parameter variations. The term **stability margin** refers to the gain margin and the phase margin, which are quantitative measures of stability. A proven method of obtaining good robustness properties is the use of a feedback control system, which can be designed to allow for variations in the system dynamics. Some causes of **variation** in system dynamics are

Modeling and data errors in the nominal plant and system.
Changes in environmental conditions, manufacturing tolerances, wear due to aging, and noncritical material failures.
Errors due to calibration, installation, and adjustments.

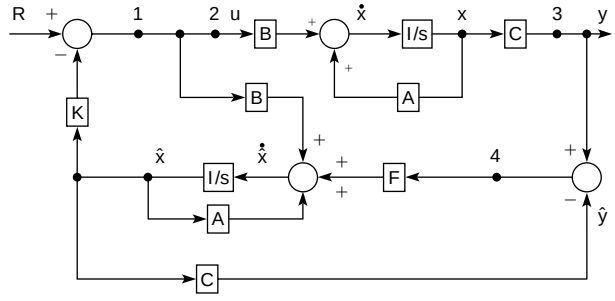
Feedback control systems with good feedback properties have been synthesized for SISO systems. Classical frequency domain techniques such as Nyquist, root-locus, Bode, and Nichols plots have been used to obtain the feedback control system for the SISO system. These design techniques have allowed the synthesis of feedback control systems yielding **insensitivity** to bounded parameter variations and a large stability margin. The success of the feedback control system for the SISO system has led to the direct extension of the classical frequency domain technique to the design of a multivariable feedback control system. This extension to the multivariable design problem examines an individual feedback loop as the phase and gain margins are varied, while the nominal phase and gain values in the remaining feedback paths are held constant. This technique, however, fails to consider the results of simultaneous variation of gain and phase in all paths, which is a real-world possibility and needs to be considered. A method of obtaining a feedback control system with good robustness properties that takes into consideration simultaneous gain and phase variation is the LQG/LTR.

The LQG/LTR technique can be applied to MIMO systems or SISO systems. The LQG/LTR technique not only has the good robustness properties of the classical frequency domain techniques but also is capable of minimizing the effects of unmodeled high-frequency dynamics, neglected nonlinearities, and a reduced-order model. **Computer control system design tools,** for instance, developed by Integrated Systems Incorporated (ISI) and The MathWorks can be used in synthesizing the LQG/LTR technique and other controller design techniques. The use of computer-aided design tools eliminates the numerical programming burden of programming the complex LQG/LTR algorithm.

In what follows, the needed concepts are defined and the LQG/LTR procedure for development of a **model-based compensator** (MBC) is described. A unity-feedback MBC is selected for the controller structure, because of its similarity to the SISO unity-feedback control system well known to classical control designers. The MBC closed-loop control system has proven to offer great practical considerations in the design of automatic control systems.

Robustness Concepts of the LQG and LQG/LTR Control Systems

The LQG/LTR design procedure is based on the system configuration of the LQG controller shown in Fig. 16.2.44. The LQG controller consists of a Kalman filter state estimator and a linear quadratic regulator. The Kalman filter state estimator has good robustness properties for plant perturbations at the plant output. The linear quadratic regulator (LQR) has good robustness properties for perturbations at the plant input. Even though its components separately have good robustness properties, the LQG controller is found to have no guaranteed robustness properties at either the input (point 2) or the output (point 3) of the plant.



- LQG guaranteed no robustness properties at the input or output of the plant.
- Point 1 has the good robustness properties of the full-state feedback system.
- Point 4 has the good robustness properties of the Kalman filter.
- Point 2 has no guaranteed robustness properties.
- Point 3 has no guaranteed robustness properties.
- The LQG/LTR design method permits recovery of the robustness properties of point 1 at point 2 or the robustness properties of point 4 at point 3.

Fig. 16.2.44 Summary of the robustness properties of the LQG block diagram.

The LQG/LTR design procedure allows us to recover robustness properties at either the input or the output of the plant. If robustness is desired at the input to the plant, first a nominal robust LQR design is made to satisfy the design constraints. Next, an LTR step is made to design a Kalman filter gain that recovers the robustness at the input to the plant of the LQG controller that is approximately that of the nominal LQR design. This implies from Fig. 16.2.44 that the robustness properties at points 1 and 2 are approximately the same.

If robustness is desired at the **output of the plant,** first a nominal robust Kalman filter design is made to satisfy the performance constraints. Next, an LTR step is made to design an LQR gain that recovers the robustness at the output of the plant that is approximately that of the nominal Kalman filter design. This implies from Fig. 16.2.44 that the robustness properties at points 3 and 4 are approximately the same.

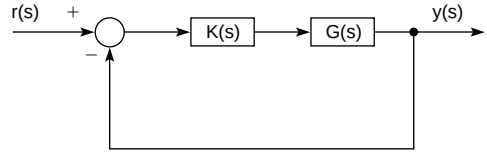


Fig. 16.2.45 Unity feedback model-based compensator (UFMBC).

The block diagram of the unity-feedback MBC is shown in Fig. 16.2.45. This control system structure allows tracking and regulation of a reference input at the output of the plant. The filter and regulator gains used in the controller $R(s)$ are obtained appropriately, depending on whether robustness is desired at the input or the output of the plant. Note that the robustness properties of the **unity-feedback MBC** (UFMBC) are the same as those of the LQG system, since the UFMBC is just an alternative structure of the LQG system.

MATHEMATICS AND CONTROL BACKGROUND

The use of **matrix algebra** is very important in the study of MIMO control system design methods. Some MIMO control system design methods use matrix algebra to decouple the system so that the plant matrix transfer function consists only of individual **decoupled transfer functions,** each represented conventionally as the ratio of two polynomials. This simplification then allows the use of conventional methods such as the Bode plot, Nyquist plot, and Nichols chart as a means of designing a controller for each loop of the MIMO plant. The LQG/LTR control system design method, however, does not require **decoupling** of the plant but rather preserves the coupling between the individual loops of the MIMO plant. Each step of the LQG/LTR design technique therefore requires the use of some matrix algebra, vector and matrix norms, and singular values during the control system design procedure. For instance, matrix algebra is used to derive a suitable representation of the closed-loop system for stability and robustness analysis. Matrix norms and singular values are used to examine the matrix magnitude of the MIMO transfer function. The **matrix magnitude** or **singular values** of a system allows preservation of the coupling between the various loops of the MIMO control system.

The concepts of matrix norm, singular values, controllability, observability, stabilizability, and detectability are very important in applying MIMO design procedures. It is therefore important that these concepts be well understood.

The purpose of this section is to provide **background material** in mathematics and control systems that has proved useful in deriving the procedures and analysis required to implement the control system design procedure.

Matrix Norm and Singular Values

The matrix norm of $\|\mathbf{A}\|_2$ is defined for the $m \times n$ matrix $\mathbf{A}$. The matrix $\mathbf{A}$ defines a linear transformation from the vector space V, called the domain to a vector space W, which is called the range. Thus the linear transformation

$$\mathbf{A}(x) = \mathbf{A}x \tag{16.2.68}$$

transforms a vector x in $V \in C^n$ into a vector $\mathbf{A}(x)$ in $W \in C^m$. The matrix norm of $\|\mathbf{A}\|_2$ is defined as

$$\|\mathbf{A}\|_2 = \max_{x \neq 0} \frac{\|\mathbf{A}\|_2}{\|x\|_2} = l_2 \text{ norm} \tag{16.2.69}$$

Since the **l_2 norm** is a useful tool in control system analysis, further insight into its definition is given. The l_2 norm of Eq. (16.2.69) can be written as

$$\|\mathbf{A}\|_2 = \max_i [\lambda_i(\mathbf{A}^H\mathbf{A})]^{1/2} \qquad i = 1, 2, \cdots, n \tag{16.2.70}$$

$$\|\mathbf{A}\|_2 = \max_i [\lambda_i(\mathbf{A}\mathbf{A}^H)]^{1/2} \qquad i = 1, 2, \cdots, n \tag{16.2.71}$$

The matrices $\mathbf{A}^H\mathbf{A}$ and $\mathbf{A}\mathbf{A}^H$ are Hermitian and positive semidefinite. The eigenvalues of these matrices are real and nonnegative. The definition of the **singular value** of a complex matrix $\mathbf{A} \in \mathbf{C}^{n \times n}$ is

$$\begin{aligned}\sigma_i(\mathbf{A}) &= [\lambda_i(\mathbf{A}^H\mathbf{A})]^{1/2} \\ &= [\lambda_i(\mathbf{A}\mathbf{A}^H)]^{1/2} \geq 0 \qquad i = 1, 2, \cdots, n\end{aligned} \tag{16.2.72}$$

If $\mathbf{A} \in \mathbf{C}^{m \times n}$, which indicates that $\mathbf{A}$ is a nonsquare matrix, then

$$\sigma_i(\mathbf{A}) = [\lambda_i(\mathbf{A}^H\mathbf{A})]^{1/2} = [\lambda_i(\mathbf{A}\mathbf{A}^H)]^{1/2} \tag{16.2.73}$$

for $1 \leq i \leq k$, where k is the number of singular values = min (m, n). The maximum and minimum singular values are defined respectively as

$$\overline{\sigma}(\mathbf{A}) = \|\mathbf{A}\|_2 \tag{16.2.74}$$

and

$$\underline{\sigma}(\mathbf{A}) = \max_{x \neq 0} \frac{\|\mathbf{A}x\|}{\|x\|_2} = \frac{\mathbf{1}}{\|\mathbf{A}^{-1}\|_2} \tag{16.2.75}$$

provided $\mathbf{A}$ has an inverse.

Controllability, Observability, Stabilizability, and Detectability

The LQG/LTR control system design method has a certain **requirement** of the plant for which the controller is being designed: the plant must be at least stabilizable and detectable. Therefore it is important that the conditions of controllability, observability, and the less restricted conditions of stabilizability and detectability be understood. A description of these concepts follows.

Given the state equation for the linear time invariant (LTI) system described as $\dot{x}(t) = \mathbf{A}x(t) + \mathbf{B}u(t)$ and $y(t) = \mathbf{C}x(t)$. The matrices **A, B,** and **C** are constant. Note that $x \in R^n$, $n \in R^p$, and $y \in R^p$ are respectively the state, input, and output vectors. Variables n and p are the dimensions of the states and the input and output vectors. The input and output vectors are of the same dimension.

A system is said to be controllable if it is possible to transfer the system from any initial state x_0 in state space to any other state x_F using the input in a finite period of time. The system is said to be uncontrollable otherwise. The condition of controllability places **no constraint** on the input in the system.

The algebraic test for controllability of the LTI system state equation uses the controllability matrix

$$\mathbf{Q} = [\mathbf{B}\ \mathbf{A}\mathbf{B} \cdots \mathbf{A}^{n-1}\mathbf{B}] \tag{16.2.76}$$

where n is the **number of states.** The LTI system is controllable if the rank of $\mathbf{Q}$ is equal to n. If the rank of $\mathbf{Q}$ is less than n, the system is not controllable. An uncontrollable system indicates that some modes or poles are not affected by the control input. Uncontrollable systems with **unstable modes** imply that a controller design is not possible to ensure system stability. If the uncontrollable modes of the uncontrollable system are stable, then the system is stabilizable.

An LTI system is considered **observable** at initial time t_0 if there exists a finite time t_1 greater than t_0 such that for any initial state x_0 in state space using knowledge of the input u and output y over the time interval $t_0 < t < t_1$, the state x_0 can be determined. The system is said to be unobservable otherwise.

The algebraic test for observability of the LTI system uses the observability matrix

$$\mathbf{N} = [\mathbf{C}^T\ \mathbf{A}^T\mathbf{C}^T \cdots (\mathbf{A}^T)^{n-1}\mathbf{C}^T] \tag{16.2.77}$$

The LTI system is observable if the rank of $\mathbf{N}$ is equal to n, where n is the number of states. If the rank of $\mathbf{N}$ is less than n, the system is unobservable. An unobservable system indicates that not all system modes contribute to the output of the system. If the **unobservable modes** (states) are stable, then the system is considered to be detectable.

EVALUATING MULTIVARIABLE PERFORMANCE AND STABILITY ROBUSTNESS OF A CONTROL SYSTEM USING SINGULAR VALUES

Performance Robustness

The requirements for the control system design are formulated by placing certain restrictions on the singular values of the **return ratio, return difference,** and **inverse return difference** of the control system. The block diagram in Fig. 16.2.46 shows the controller $K(s)$ and plant $G(s)$ with input $d_I(s)$ and output $d_o(s)$ disturbances and sensor noise $n(s)$.

Using Fig. 16.2.46, a **frequency-domain transfer function** representation for the plant output is first obtained. These transfer function representations, used in conjunction with concepts of singular values, allow the formulation of requirements in the frequency domain to obtain a control system with good performance and stability robustness.

In this design, performance robustness is of concern at the output of the plant. Therefore it is required that the plant output transfer function be derived in order to analyze and assist in obtaining a control system design to meet certain performance robustness requirements. Using Fig. 16.2.46, the plant output transfer function is derived and performance

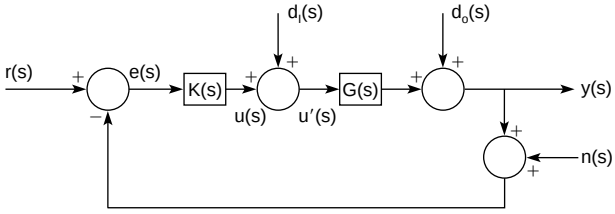


Fig. 16.2.46 Model-based unity feedback MIMO control system.

robustness requirements established. From Fig. 16.2.46 it can be shown that

$$y(s) = d_o(s) + G(s)u'(s) \tag{16.2.78}$$

$$e(s) = r(s) - y(s) - n(s) \tag{16.2.79}$$

and

$$u'(s) = d_o(s) + u(s) = d_I(s) + K(s)e(s). \tag{16.2.80}$$

where $G(s) = \mathbf{C}[s\mathbf{I} - \mathbf{A}]^{-1}\mathbf{B}$ is the plant transfer function and **A, B,** and **C** are system matrices. The transfer functions $y(s)$, $d_o(s)$, $e(s)$, $r(s)$, $n(s)$, $u(s)$, $d_I(s)$, and $K(s)$ are respectively the output, output disturbance, error, reference, noise, input, input disturbance, and controller transfer functions.

Using Eqs. (16.2.79) and (16.2.80) it can be shown that Eq. (16.2.78) becomes

$$\begin{aligned} y(s) &= [G(s)K(s)][r(s) - y(s) - n(s)] \\ &\quad + d_o(s) + G(s)d_I(s) \\ &= G(s)K(s)r(s) - G(s)K(s)y(s) - G(s)K(s)n(s) \\ &\quad + d_o(s) + G(s)d_I(s) \end{aligned} \tag{16.2.81}$$

Therefore,

$$\begin{aligned} y(s) &= [\mathbf{I} + G(s)K(s)]^{-1}[G(s)K(s)r(s) - G(s)K(s)n(s) \\ &\quad + d_o(s) + G(s)d_I(s)] \\ &= [\mathbf{I} + G(s)K(s)]^{-1}\{G(s)K(s)[r(s) - n(s)]\} \\ &\quad + [\mathbf{I} + G(s)K(s)]^{-1}[d_o(s) + G(s)d_I(s)] \\ &= y_r(s) + y_{d_I}(s) + y_{d_o}(s) + y_n(s), \end{aligned} \tag{16.2.82}$$

where $y_r(s)$, $y_{d_I}(s)$, $y_{d_o}(s)$, and $y_n(s)$ are, respectively, the contributions to the plant output due to the reference input, input disturbance, output disturbance, and noise input to the system. The requirements for good command following, disturbance rejection, and insensitivity to sensor noise are defined by using singular values to characterize the magnitude of MIMO transfer functions. These requirements are also **applicable to SISO systems** (Kazerooni and Narayan; Murphy and Bailey, April 1989).

Command Following The conditions required to obtain a MIMO control system design with good command following is described. Good command following of the reference input $r(s)$ by the plant output $y(s)$ implies that

$$y(s) \approx r(s) \qquad \forall\ s \in s_R, \tag{16.2.83}$$

where $s = j\omega$ and $s_R = j\omega_R$ and ω_R is the frequency range of the input $r(s)$. Letting $d_o(s) = d_I(s) = n(s) = 0$ in Eq. (16.2.82) results in

$$\begin{aligned} y_r(s) &= [\mathbf{I} + G(j\omega)K(j\omega)]^{-1}G(j\omega)K(j\omega)r(j\omega) \qquad \forall\ \omega \in \omega_R \\ &= \{\mathbf{I} + [G(j\omega)K(j\omega)]^{-1}\}^{-1}r(j\omega) \qquad \forall\ \omega \in \omega_R \end{aligned} \tag{16.2.84}$$

where $G(j\omega)K(j\omega)$ is defined as the return ratio. Therefore, to obtain good command following, the system **inverse return difference** $\mathbf{I} + [G(j\omega)K(j\omega)]^{-1}$ must have the property such that

$$\mathbf{I} + [G(j\omega)K(j\omega)]^{-1} \approx \mathbf{I} \qquad \forall\ \omega \in \omega_R \tag{16.2.85}$$

which implies that

$$\underline{\sigma}\,[G(j\omega)K(j\omega)] \gg 1 \qquad \forall\ \omega \in \omega_R \tag{16.2.86}$$

Thus the magnitude of the minimum singular value of the **return ratio** $\underline{\sigma}[G(j\omega)K(j\omega)]$ must be large over the frequency range of the input to obtain good command following of the control system.

Disturbance Rejection To minimize the effect of the output disturbance, $y_{d_o}(s)$, on the output signal $y(s)$, Eq. (16.2.82) is used with $n(s) = r(s) = d_I(s) = 0$, which results in

$$y_{d_o}(s) = [\mathbf{I} + G(s)K(s)]^{-1}\, d_o(s). \tag{16.2.87}$$

Rejection of this output disturbance at the plant output requires that the magnitude of the **return difference** $\mathbf{I} + G(s)K(s)$ to meet the following condition

$$\underline{\sigma}\,[\mathbf{I} + G(j\omega)K(j\omega)] \gg 1 \qquad \forall\ \omega \in \omega_{d_o} \tag{16.2.88}$$

which implies

$$\underline{\sigma}\,[G(j\omega)K(j\omega)] \gg 1 \qquad \forall\ \omega \in \omega_{d_o} \tag{16.2.89}$$

where y_{d_o} is the frequency range of the output disturbance $d_o(s)$.

With $n(s) = r(s) = d_o(s) = 0$ in Eq. (16.2.82), the effect of input disturbance on the plant output is examined. This effect is seen to be

$$y_{d_I}(s) = [\mathbf{I} + G(j\omega)K(j\omega)]^{-1}[G(j\omega)d_I(j\omega)] \qquad \forall\ \omega \in \omega_{d_I} \tag{16.2.90}$$

where ω_{d_I} is the frequency range of the input disturbance $d_I(s)$. To minimize the effect of $y_{d_I}(s)$ on the plant output requires that the magnitude of the return difference be large (in general, the larger the better). This requirement implies that

$$\underline{\sigma}\,[G(j\omega)K(j\omega)] \gg 1 \qquad \forall\ \omega \in \omega_{d_I} \tag{16.2.91}$$

to reject the input disturbance. It should be noted that this equation gives only general conditions for rejection. The control engineer must determine how much greater than 1 the return ratio magnitude must be.

Noise Rejection The requirements to minimize the effects of sensor noise on the plant output must be determined. First letting $n(s) = d_I(s) = d_o(s) = 0$ in Eq. (16.2.82) results in

$$\begin{aligned} y_n(s) &= [\mathbf{I} + G(s)K(s)]^{-1}[G(s)K(s)]\, n(s) \qquad \forall\ s \in s_n \\ &= \{\mathbf{I} + [G(s)K(s)]^{-1}\}^{-1}n(s) \qquad \forall\ s \in s_n \end{aligned} \tag{16.2.92}$$

where $s_n = j\omega_n$ and ω_n is the frequency range of the sensor noise. Thus to minimize the effects of sensor noise on the plant output it is desired to have the magnitude of $y_n(s)$ small over the frequency range of the noise. Thus

$$\overline{\sigma}\,[G(j\omega)K(j\omega)] \gg 1 \qquad \forall\ \omega \in \omega_n \tag{16.2.93}$$

is required to minimize the effect of noise on the plant output.

Any overlapping of any of the frequency ranges ω_R, ω_{d_I}, ω_{d_o}, ω_n will require system design specification tradeoffs to be made. This will definitely occur when the lower frequency ranges of ω_R and/or ω_{d_I} and ω_{d_o} overlap the higher frequency range ω_n.

Stability Robustness

All linear control system design methods are based on a linear model of the plant. Because this design model only approximates the actual plant, an error exists between the design model and the actual plant. This error can be evaluated as **absolute** or **relative.** The method of evaluation of this error (model uncertainty) depends on the nature of the differences between the actual design model and the actual plant. The evaluation of an absolute error defines an additive model uncertainty, whereas the evaluation of a relative error defines a multiplicative model uncertainty. The **additive** and multiplicative model uncertainties can be further classified as either structured or unstructured model uncertainties (Murphy and Bailey, ORNL 1989).

A **structured model uncertainty** is usually a low-frequency phenomenon characterized by variations in the parameters of the linear time-invariant plant design model. These parameter variations are caused by changes in the plant operating conditions, wear due to aging, and environmental conditions. A structured uncertainty also indicates that the gain and phase information concerning the uncertainty is known. Additive model uncertainties are generally considered structured uncertainties.

An **unstructured uncertainty** is typically a high-frequency phenome-

non in which the only information known about the uncertainty is its magnitude. Unstructured uncertainties usually have the characteristic of becoming significant at high frequencies. Multiplicative uncertainties are characterized as unstructured. Unstructured uncertainties are usually caused by the truncation of high-frequency plant dynamics or modes due to the linearization process, and the neglect of plant dynamics such as actuators and sensors. Some plant dynamics that characterize unstructured uncertainty are flexible-body dynamics, electrical and mechanical resonances, and transport delays. Most linear control-system design methods neglect these high-frequency plant dynamics. The result is that if a high-frequency bandwidth controller is implemented, the high-frequency modes could be excited, resulting in an unstable control system.

Model uncertainty clearly is seen to impose limitations on the achievable performance of a feedback control system design. Hence this section will focus on the limitations imposed by uncertainty, not on the difficult problem of exact representation of the system uncertainty.

Multiplicative Uncertainty Multiplicative uncertainty (perturbation) can be represented at the input or output of the plant, and is called respectively the input multiplicative uncertainty or the **output multiplicative uncertainty.** The model uncertainty due to actuator and sensor dynamics is modeled respectively as an input or output multiplicative perturbation. A reduced-order model of an already **linearized model** results in a multiplicative model uncertainty. **Time delay** at the input or output of a plant yields, respectively, input and output multiplicative uncertainties in the plant. Even though other multiplicative model uncertainties exist among various types of control system design problems, the model uncertainties mentioned above are a common consideration in most control problems.

Output Multiplicative Uncertainty

The output multiplicative uncertainty $\Delta G(s)$ is defined as

$$\Delta G(s) = [G'(s) - G(s)]\, G^{-1}(s) \tag{16.2.94}$$

where

$$G'(s) = L(s)G(s) \tag{16.2.95}$$

is the perturbed transfer function, $L(s)$ represents dynamics not included in the nominal plant $G(s)$. Thus resulting in

$$\Delta G(s) = [L(s) - \mathbf{I}] \tag{16.2.96}$$

A block diagram of a control system with output multiplicative perturbation is shown in Fig. 16.2.47. The conditions of stability for a control system with output multiplicative uncertainty require that

$$\underline{\sigma}\,\{\mathbf{I} + [G(j\omega)K(j\omega)]^{-1}\} > \overline{\sigma}\,[\Delta G(j\omega)] \qquad \forall\ \omega > 0 \tag{16.2.97}$$

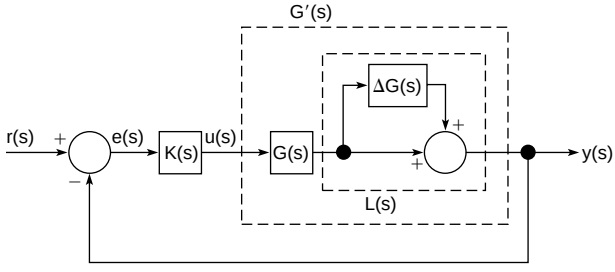


Fig. 16.2.47 Control system with output multiplicative perturbation.

Additive Uncertainty The block diagram of a control system with a plant additive model uncertainty is shown in Fig. 16.2.48. The representation of the additive model uncertainty using an absolute error criterion is defined as

$$\Delta G(s) = G'(s) - G(s) \tag{16.2.98}$$

where in this case the perturbed transfer function $G'(s)$ is

$$G'(s) = [\mathbf{C} + \mathbf{\Delta C}]\{s\mathbf{I} - [\mathbf{A} + \mathbf{\Delta A}]\}^{-1}[\mathbf{B} + \mathbf{\Delta B}] \tag{16.2.99}$$

and the nominal plant transfer function $G(s)$ is

$$G(s) = \mathbf{C}[s\mathbf{I} - \mathbf{A}]^{-1}\mathbf{B} \tag{16.2.100}$$

and where **A, B,** and **C** are the nominal system matrices and **ΔA, ΔB,** and **ΔC** are the perturbation matrices that indicate the respective deviations from the nominal system matrices. The condition for stability of the additively perturbed control system requires that

$$\underline{\sigma}[G(j\omega)K(j\omega)] > \overline{\sigma}[\Delta G(j\omega)K(j\omega)] \qquad \forall\ \omega > 0 \tag{16.2.101}$$

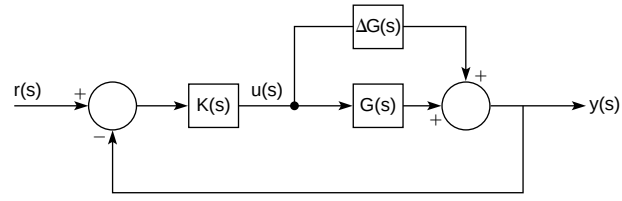


Fig. 16.2.48 Control system with additive perturbation.

Table 16.2.7 contains a summary of the general design requirements presented in this section to obtain a control system with good performance and stability robustness.

REVIEW OF OPTIMAL CONTROL THEORY

Linear Quadratic Regulator

Consider the linear time-invariant state-space system, first-order vector differential equation model

$$\dot{\mathbf{x}}(t) = \mathbf{A}x(t) + \mathbf{B}u(t) \tag{16.2.102}$$

$$y(t) = \mathbf{C}x(t) \tag{16.2.103}$$

where **A, B,** and **C** are constant system matrices, and $x(t)$, $u(t)$, and $y(t)$ are respectively the system state, input, and output vectors. The system is stabilizable and controllable. The goal of this procedure is to minimize the finite performance index

$$J = \int_0^\infty [x^{\mathrm{T}}(t)\mathbf{Q}x(t) + u^{\mathrm{T}}(t)\mathbf{R}u(t)]\, dt \tag{16.2.104}$$

$$J = \int_0^\infty [x^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\mathbf{C}x + u^{\mathrm{T}}\mathbf{R}u]\, dt$$

$$= \int_0^\infty [y^{\mathrm{T}}y + u^{\mathrm{T}}\mathbf{R}u]\, dt$$

where **Q** is a symmetric positive semidefinite matrix and **R** is positive definite.

Table 16.2.7 General System Design Specifications

System requirements	Range
Good command-following	$\underline{\sigma}[G(j\omega)K(j\omega)] \gg 0$ dB $\quad \forall\ \omega \in \omega_R$
Good disturbance rejection	$\underline{\sigma}[G(j\omega)K(j\omega)] \gg 0$ dB $\quad \forall\ \omega \in \omega_d$
Good immunity to noise	$\overline{\sigma}[G(j\omega)K(j\omega)] \ll 0$ dB $\quad \forall\ \omega \in \omega_n$
Good system response to high-frequency modeling error*	$\underline{\sigma}\{I + [G(j\omega)K(j\omega)]^{-1}\} > \|\Delta G(j\omega)\| \quad \forall\ \omega > 0$ rad/s
Good insensitivity to parameter variations at low frequencies*	$\underline{\sigma}[G(j\omega)K(j\omega)] > \overline{\sigma}[\Delta G(j\omega)K(j\omega)]$

* The $\Delta G(j\omega)$ indicated corresponds to the applicable multiplicative or additive model uncertainty.

The **performance index** J, referred to as the quadratic performance index, implies that a control $u(t)$ is sought to facilitate the minimization of J. The weighting matrices $\mathbf{Q}$ and $\mathbf{R}$ are selected to reflect the importance of particular states and control inputs. In general the weighting of the diagonal elements of $\mathbf{Q}$ can be determined by the importance of the state, as observed by the $\mathbf{C}$ matrix of the output equation. The effect of $\mathbf{Q}$ is the ability to control the **transient response** of the LQR. The effect of $\mathbf{R}$ is the ability to **control** the **energy** resulting from u.

Therefore a linear control law that minimizes J is defined as

$$u(t) = -\mathbf{K}x(t) \tag{16.2.105}$$

where

$$\mathbf{K} = \mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{S} \tag{16.2.106}$$

where, in turn, $\mathbf{S}$ is a constant symmetric positive semidefinite matrix that satisfies the algebraic Riccati equation

$$\mathbf{Q} - \mathbf{S}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{S} + \mathbf{A}^{\mathrm{T}}\mathbf{S} + \mathbf{S}\mathbf{A} = 0 \tag{16.2.107}$$

The closed loop regulator is given by

$$\dot{\mathbf{x}}(t) = [\mathbf{A} - \mathbf{B}\mathbf{K}]x(t) = [\mathbf{A} - \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{S}]x(t) \tag{16.2.108}$$

and is asymptotically stable provided the state equation (16.2.102) and the output equation (16.2.103) are detectable.

Kalman Filter

What follows is a method of reconstructing an estimate of the states using only the output measurements of the system. The method of reconstruction must be applicable when **process noise** and **measurement noise,** respectively, corrupt the plant state equations and the output measurement equations.

Let us consider the stochastic linear system state space vector differential equation model

$$\dot{\mathbf{x}}(t) = \mathbf{A}x(t) + \mathbf{B}u(t) + \Gamma d(t) \tag{16.2.109}$$

$$y(t) = \mathbf{C}x(t) + n(t) \tag{16.2.110}$$

where **A, B, C,** and Γ are constant system matrices. Vectors $s(t)$, $u(t)$, and $y(t)$ are respectively the state, input, and output vectors. Vector $d(t)$ is a process noise random vector and vector $n(t)$ is a measurement noise random vector. Assuming that $d(t)$ and $n(t)$ are zero-mean, uncorrelated, gaussian white noises,

$$E[d(t)]E[n(t)] = 0 \qquad \text{for all } t \tag{16.2.111}$$

$$E[d(t)d^{\mathrm{T}}(\tau)] = \mathbf{D}_o\delta(t - \tau) \tag{16.2.112}$$

$$E[n(t)n^{\mathrm{T}}(\tau)] = \mathbf{N}\delta(t - \tau) \qquad \text{for all } t < \tau \tag{16.2.113}$$

$$E[d(t)n^{\mathrm{T}}(\tau)] = 0. \tag{16.2.114}$$

where τ is the difference between two points in time, $E[\ ]$ is mean value, δ is an impulse function, and $\mathbf{N}$ and $\mathbf{D}_o$ are constant symmetric, positive definite, and positive semidefinite matrices, respectively. Note that $\mathbf{N}$ and $\mathbf{D}_o$ are constant because it is assumed that $d(t)$ and $n(t)$ are wide-sense stationary.

The **state estimate $\hat{\mathbf{x}}$ of the actual measured state** x is obtained from the noisy measurement y. Therefore we can define a state error vector

$$e(t) = x(t) - \hat{\mathbf{x}}(t) \tag{16.2.115}$$

where we desire to minimize the mean square error

$$\bar{\mathbf{e}} = E[e^{\mathrm{T}}(t)e(t)] \tag{16.2.116}$$

The estimator can thus be shown to take the form

$$\dot{\hat{\mathbf{x}}} = \mathbf{A}x(t) + \mathbf{B}u(t) + \mathbf{F}[y(t) - \mathbf{C}x(t)] \tag{16.2.117}$$

where $\mathbf{F}$ is the filter gain which minimizes Eq. (16.2.116); $\mathbf{F}$ is defined as

$$\mathbf{F} = \Sigma\mathbf{C}^{\mathrm{T}}\mathbf{N}^{-1} \tag{16.2.118}$$

where Σ is a constant-variance matrix of the error $e(t)$. It is assumed here that $e(t)$ is wide-sense stationary. The matrix Σ is obtained by solving the algebraic variance Riccati equation

$$\mathbf{D} - \Sigma\mathbf{C}^{\mathrm{T}}\mathbf{N}^{-1}\mathbf{C}\Sigma + \mathbf{A}\Sigma + \Sigma\mathbf{A}^{\mathrm{T}} = 0 \tag{16.2.119}$$

where

$$\mathbf{D} = \Gamma\mathbf{D}_o\Gamma \tag{16.2.120}$$

A sufficient condition to obtain a unique and positive definite Σ from Eq. (16.2.117) requires that $[\mathbf{A}, \mathbf{C}]$ be completely observable. However, if the pair $[\mathbf{A}, \mathbf{C}]$ is required to be detectable, then Σ can be positive semidefinite.

The **reconstruction error** $e(t)$ satisfies the differential equation

$$\dot{e}(t) = [\mathbf{A} - \mathbf{F}\mathbf{C}]e(t) + [\Gamma - \mathbf{F}]\begin{bmatrix} d(t) \\ n(t) \end{bmatrix} \tag{16.2.121}$$

such that

$$e(t) \to 0 \text{ as } t \to \infty \qquad \forall\ t \geq t_0$$

if and only if the **observer** is asymptotically stable. The poles of the observer (filter) are found by using the closed-loop dynamics matrix $[\mathbf{A} - \mathbf{F}\mathbf{C}]$. To obtain asymptotic stability of the filter, it is required that the pair $[\mathbf{A}, \Gamma]$ be completely controllable. The necessary and sufficient condition of **stabilizability** of the pair $[\mathbf{A}, \Gamma]$ will ensure **stability** also.

PROCEDURE FOR LQG/LTR COMPENSATOR DESIGN

An LQG/LTR compensator will be designed for a **minimum phase plant (no right half plane zeroes).** The control system so designed will be guaranteed to be stable and will be robust at the plant output.

First, it is assumed that if **zero steady-state error** is desired in the control system design, the plant inputs will be augmented to obtain the desired **integral action.** In some cases the dynamics of the plant may have poles that are nearly zero, which establishes a **near integral action** in the plant. In such a case the addition of integrals at the plant input causes a much greater than **20-dB/decade roll-off,** thus yielding unreasonable plant dynamics characterized by the return ratio singular-value plots. Therefore caution is required in deciding how or whether to augment the plant dynamics. In the following description of the compensator design it is assumed that the nominal plant transfer function $G(s)$ has been augmented as required by the control engineers' desires.

Step 1 The first step requires selecting the appropriate transfer function $[G_{\mathrm{FOL}}(s)]$ that satisfies the desired **performance and stability robustness** of the final control system. Therefore, the appropriate scalar $\mu > 0$ and a constant matrix $\mathbf{L}$ such that the singular values of

$$G_{\mathrm{FOL}}(s) = (1/\sqrt{\mu})[\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{L}] \tag{16.2.122}$$

approximately meet the desired control system requirements. The system matrices $\mathbf{C}$ and $\mathbf{A}$ of the minimum phase component are assumed to have been augmented already to obtain the necessary integral action. The selection of $\mathbf{L}$ is made so that at low frequencies ($s = j\omega \to j0$), high frequencies ($s = j\omega \to j\omega$), or intermediate frequencies ($s = j\omega \to j\omega_i$) the singular values of $G_{\mathrm{FOL}}(s)$ are approximately the same. After $\mathbf{L}$ is selected, then μ is adjusted to obtain the desired crossover frequency $\sigma[G_{\mathrm{FOL}}(j\omega)]$; therefore μ functions as a gain parameter of the transfer function. It should be noted that it is desired to balance the singular values of $G_{\mathrm{FOL}}(s)$ in the region of the crossover frequency in order to obtain similar responses for each loop of the system. Also, typically it is desired to have the return ratio singular values roll off at a rate no greater than -20 dB/decade in the region of gain crossover.

Step 2 In this step the Kalman filter transfer function is defined as

$$G_{\mathrm{KF}}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{F} \tag{16.2.123}$$

where $\mathbf{F}$ is the Kalman filter gain and $\mathbf{C}$ and $\mathbf{A}$ are the system matrices of the linear time-invariant state-space system.

To compute the **filter gain F,** first the following Kalman filter **algebraic Riccati equation (ARE)** is solved

$$\mathbf{L}\mathbf{L}^{\mathrm{T}} - (1/\sqrt{\mu})\Sigma\mathbf{C}^{\mathrm{T}}\mathbf{C} + \mathbf{A}\Sigma + \Sigma\mathbf{A}^{\mathrm{T}} = 0 \tag{16.2.124}$$

where $\mathbf{L}$ and μ are as obtained in step 1 and Σ (the covariance matrix) is the solution to the equation. The filter gain is thus computed to be

$$\mathbf{F} = (1/\sqrt{\mu})\Sigma\mathbf{C}^{\mathrm{T}} \tag{16.2.125}$$

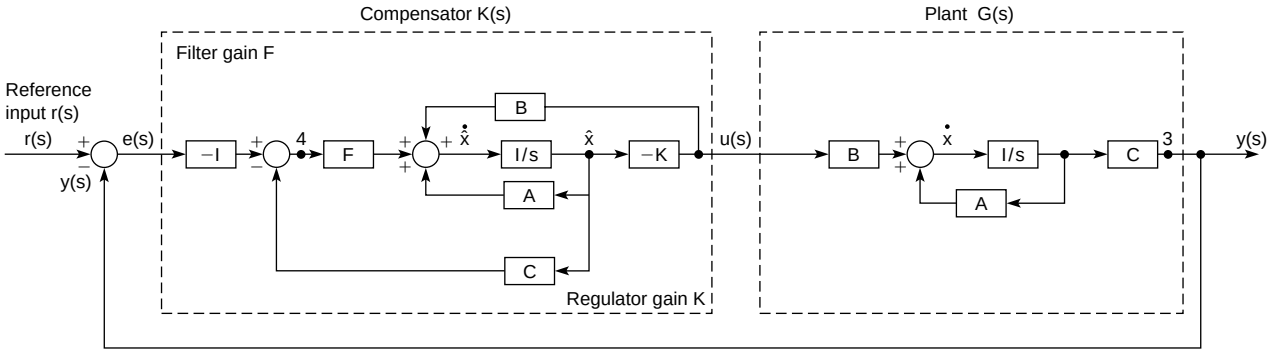


Fig. 16.2.49 Block diagram of unity-feedback LQG/LTR control system.

Thus as seen above, **L** and μ are tunable parameters used to produce the necessary filter gain **F** to meet the desired system performance and stability robustness constraints. This step completes the LQG/LTR design requirement of designing the target Kalman filter transfer function matrix with the desired properties which will be recovered at the plant output during the loop transfer recovery step. The general block diagram of the LQG/LTR control system is shown in Fig. 16.2.49.

Step 3 The transfer function of the LQG/LTR controller is defined as

$$K(s) = \mathbf{K}[s\mathbf{I} - (\mathbf{A} - \mathbf{BK} - \mathbf{FC})]\mathbf{F} \qquad (16.2.126)$$

where **A, B,** and **C** are the system matrices of Eqs. (16.2.102) and (16.2.103). The filter gain **F** is as stated in Eq. (16.2.125). The **regulator gain K** will be computed in this step to achieve loop transfer recovery. The selection of **K** will yield a return ratio at the plant output such that

$$G(s)K(s) \rightarrow G_{\mathrm{KF}}(s) \qquad (16.2.127)$$

where $G(s)$ is the plant transfer function, and $K(s)$ is the controller transfer function.

The first step in selecting **K** requires solving the LQR problem, which has the following ARE:

$$\mathbf{Q}(q) - \mathbf{SBR}_0^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{S} + \mathbf{SA} + \mathbf{A}^{\mathrm{T}}\mathbf{S} = 0 \qquad (16.2.128)$$

where $\mathbf{R}_0 = \mathbf{I}$ is the control weighting matrix and $\mathbf{Q}(q) = \mathbf{Q}_0 + q^2\mathbf{CC}^{\mathrm{T}}$ is the modified state weighting matrix. The scalar q is a free design parameter. The resulting regulator gain **K** is computed as

$$\mathbf{K} = \mathbf{B}^{\mathrm{T}}\mathbf{S} \qquad (16.2.129)$$

As $q \rightarrow 0$ Eq. (16.2.128) becomes the ARE to the optimal regulator problem. As $q \rightarrow \infty$ the LQG/LTR technique guarantees that, since the design model $G(s) = \mathbf{C}(s)\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$ has no nonminimum phase zeroes, then pointwise in s

$$\lim_{q\rightarrow\infty} G(s)K(s) \rightarrow G_{\mathrm{KF}}(s) \qquad (16.2.130)$$

$$\lim_{q\rightarrow\infty} \{\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{BK}[s\mathbf{I} - (\mathbf{A} - \mathbf{BK} - \mathbf{FC})]^{-1}\mathbf{F}\} \rightarrow \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{F} \qquad (16.2.131)$$

This completes the LQG/LTR design procedure. Now that the loop transfer recovery step is complete, the singular-value plots of the return ratio, return difference, and inverse return difference must be examined to verify that the desirable loop shape has been obtained.

EXAMPLE CONTROLLER DESIGN FOR A DEAERATOR

In this section, three linear control methods are used to obtain a level-control system design for a deaerator. The **three linear control system design methods** that will be used are proportional plus integral, linear quadratic gaussian, and linear quadratic gaussian with loop transfer recovery. In this investigation, the dual of the procedure developed for the LQG/LTR design at the plant input will be used to obtain a **robust control system design** at the **plant output** (Murphy and Bailey, ORNL 1989).

The **Bode gain** and **phase plots** of the resulting control system design will be presented for each controller. Also, the singular-value plots of the **return ratio, return difference,** and **inverse return difference** for opening the loop at the plant output (point 3 in Fig. 16.2.50) will be presented.

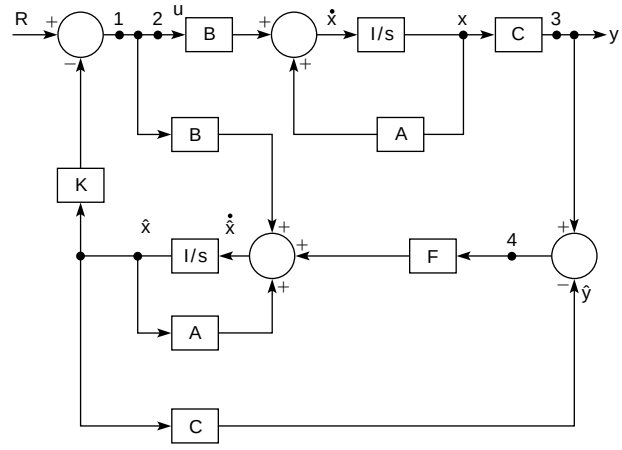


Fig. 16.2.50 Block diagram of LQG system.

The deaerator was chosen because of its **simple mathematical structure;** one input, one output, and three state variables. Thus the model is easy to follow but complex enough to illustrate the design technique.

The Linearized Model

The mathematical model of the deaerator is **nonlinear,** with a process flow diagram as shown in Fig. 16.2.51. This nonlinear plant is **linearized**

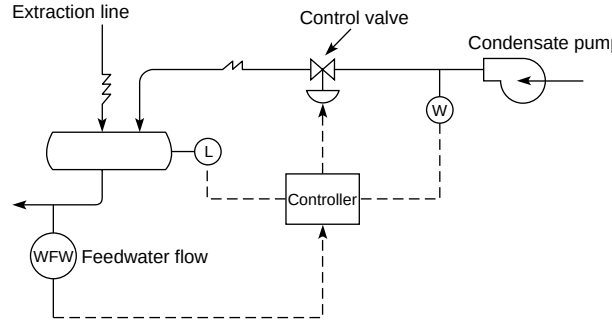


Fig. 16.2.51 General process flow diagram of the deaerator.

about a nominal operating condition, and the resulting linearized plant model will be used to obtain the controller design. The linear **deaerator model** is described by the state space linear time-invariant (LTI) differential equation

$$\dot{x}(t) = \mathbf{A}x(t) + \mathbf{B}u(t) \qquad (16.2.132)$$
$$y(t) = \mathbf{C}x(t) \qquad (16.2.133)$$

where **A, B,** and **C** are the system matrices given respectively as

$$\mathbf{A} = \begin{bmatrix} -53.802 & 1.7093 & 9.92677 \\ 0.0001761 & -0.0009245 & -0.0053004 \\ -0.001124 & -0.0243636 & -0.139635 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} -9526.25 \\ 0.265148 \\ -1.72463 \end{bmatrix}$$

$$\mathbf{C} = [0.0 \quad 3.2833 \quad 0.06373].$$

The system states are defined as

$$x = \begin{bmatrix} \delta P \\ \delta \rho \\ \delta H \end{bmatrix}$$

where δP is operating pressure between the pump and the extraction line, $\delta\rho$ is fluid density, and δH is internal energy of the tank level. Plant output y = change in deaerator tank level, and the plant input u = change in control valve.

LQG Controller Design

The design considerations for the tracking LQG control system will first be discussed. The block diagram of the LQG control system is shown in Fig. 16.2.50. It is apparent that output tracking of a reference input will not occur with the present LQG controller configuration. An alternative compensator structure, shown in Fig. 16.2.52, is therefore used to obtain a tracking LQG controller. The controller $K(s)$ uses the same filter gain **F** and regulator gain **K** computed by the LQG design procedure. The detailed block diagram is the same structure that will be used for the LQG/LTR controller shown in Fig. 16.2.49.

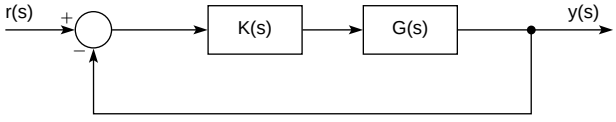


Fig. 16.2.52 Alternative compensator structure.

Using a **control system design package** MATRIX$_X$, the regulator gain **K** and filter gain **F** are computed as

$$\mathbf{K} = 10^5 \times [0.0223 \quad 3.145 \quad -0.1024]$$

and

$$\mathbf{F} = \begin{bmatrix} 0.1962 \\ 9.9998 \\ -0.0083 \end{bmatrix}$$

given **user-defined weighting matrices.** The frequency response of the open-loop transfer function of the closed-loop compensated system

$$\frac{y(s)}{r(s)} = \frac{\text{tank level}}{\text{reference input}}$$

where $y(s)$ is the **tank level** and $r(s)$ is the **reference input,** as shown in Fig. 16.2.53.

LQG/LTR Controller Design

This LQG/LTR control system has the structure shown in Fig. 16.2.52. This LQG/LTR control system design is a two-step process. First, a Kalman filter (KBF) is designed to obtain **good command following** and **disturbance rejection** over a specified low-frequency range. Also, the KBF design is made to meet the required robustness criteria with system

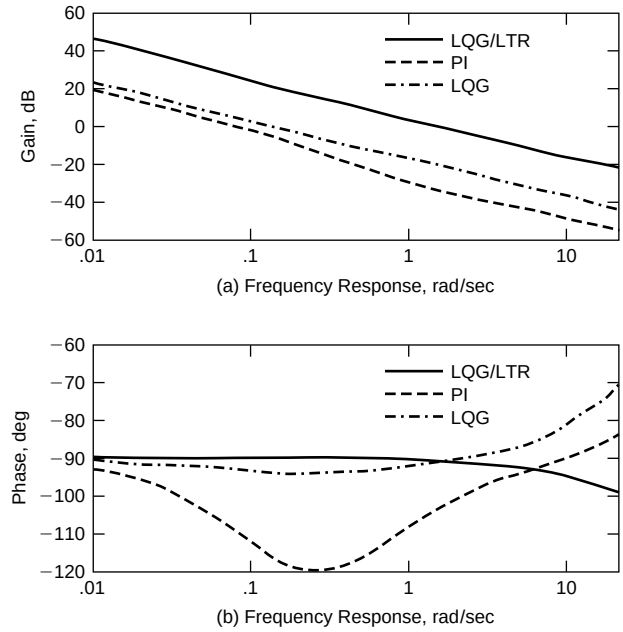


Fig. 16.2.53 Open-loop frequency response of the PI, LQG, LQG/LTR control system. (*a*) Frequency response of system gain; (*b*) frequency response of system phase.

uncertainty of $\Delta G(s)$ at the system's output. To obtain good command following and disturbance rejection requires that

$$\underline{\sigma}(G_4(s)) \geq 20 \text{ dB} \qquad \forall\ \omega < 0.1 \text{ rad/s} \qquad (16.2.134)$$

where $G_4(s)$ is the loop transfer function at the output of the KBF (point 4 of Fig. 16.2.50). Assuming the **high-frequency uncertainty** $\Delta G(s)$ for this system becomes significant at 5 rad/s, then for robustness

$$\underline{\sigma}\{\mathbf{I} + [G_4(s)]^{-1}\} \geq \|\Delta G(j\omega)\| = 5 \text{ dB} \qquad \forall\ \omega \geq 5 \text{ rad/s}. \qquad (16.2.135)$$

In this application, the frequency at which system uncertainty becomes significant, 5 rad/s, and the magnitude of the uncertainty, 5 dB at 5 rad/s, are assumed values. This assumption is required because of the lack of information on the high-frequency modeling errors of the process being studied.

The **second step** is to **recover the good robustness properties** of the loop transfer function $G_4(s)$ of point 4 at the output node y (point 3). This will be accomplished by applying the loop transfer recovery (LTR) step at the output node (point 3). The LTR step requires the computation of a regulator gain (**K**) to obtain the robustness properties of the loop transfer function $G_4(s)$ at the output node y.

The tool used to obtain the LQG/LTR design for this system, considering the low- and high-frequency bound requirements is CASCADE, a computer-aided system and control analysis and design environment that synthesizes the LQG/LTR design procedure (Birdwell et al.). Note that more recent computer-aided control system design tools synthesize this design procedure.

The resulting filter gain **F** and regulator **K** gain for this system are respectively

$$\mathbf{F} = \begin{bmatrix} -0.0107 \\ 0.472 \\ -0.02299 \end{bmatrix}$$

and

$$\mathbf{K} = [0.660 \quad 27703.397 \quad 589.4521]$$

The resulting open-loop frequency response of the compensated system is shown in Fig. 16.2.53.

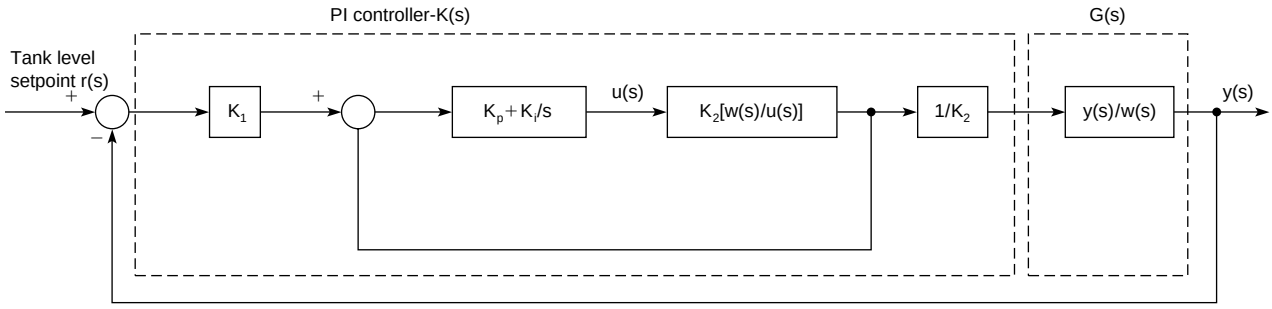


Fig. 16.2.54 Block diagram of a PI control system.

PI Controller Design

The PI controller for this system design takes on a structure similar to the three-element controller prevalent in process control. Therefore the PI control system structure takes the form shown in Fig. 16.2.54. The classical design details of the deaerator are shown in Murphy and Bailey, ORNL 1989.

The transfer functions used in this control structure are defined as

$$\frac{w(s)}{u(s)} = \frac{\text{condensate flow}}{\text{controller output}} \tag{16.2.136}$$

$$= 4.5227 \times 10^8 \times \frac{s + 0.142}{(s + 0.1405)(s + 53.8015)}$$

and

$$\frac{y(s)}{w(s)} = \frac{\text{tank level}}{\text{condensate}} \tag{16.2.137}$$

$$= 1.6804 \times 10^{-9} \times \frac{(s + 0.1986)(s + 47.458)}{s(s + 0.142)}$$

The constants K_1 and K_2 are used for unit conversion and defined respectively as 1.0 and 10^{-6}. The terms K_p and K_i are defined respectively as the proportional gain and the integral gain.

Using the practical experience of a prior three-element controller design, the proportional and integral gains are selected to be $K_p = 0.1$ and $K_i = 0.05$. The resulting open-loop frequency response plot of the compensated system $y(s)/r(s)$ is shown in Fig. 16.2.54.

Precompensator Design

The transient responses of the PI, LQG, and LQG/LTR control systems are shown in Fig. 16.2.55. The transient response of the LQG/LTR controller is seen to have a faster time to peak than the LQG and PI transient responses. Considering the practical physical limitations of the plant, it appears unlikely that the required rise time dictated by the transient response of the LQG/LTR control system is possible. The obvious solution would be to redesign the LQG/LTR control system to obtain a slower, more practical response. In the case of the LQG/LTR control system, the possibility of a redesign is eliminated, because the control system was designed to meet certain command-tracking, disturbance-rejection, and stability-robustness requirements.

Therefore, to eliminate this difficulty, a second-order precompensator will be placed in the forward path of the reference input signal. This precompensator will shape the output transient response of the system, since the control system design requires the output to follow the reference input. The precompensator in this investigation will be required to have a time to peak of 30 s with a maximum overshoot of about 3 percent. The requirements are selected to emulate somewhat the transient response of the PI controller.

The transfer function of the second-order precompensator used is as follows:

$$\frac{r(s)}{r_i(s)} = \frac{\omega_n^2}{s^2 + 2\xi\omega_n s + \omega_n^2} \tag{16.2.138}$$

where ξ is the damping ratio, ω_n is the natural frequency, $r(s)$ is the output of the precompensator, and $r_i(s)$ is the actual reference input. The

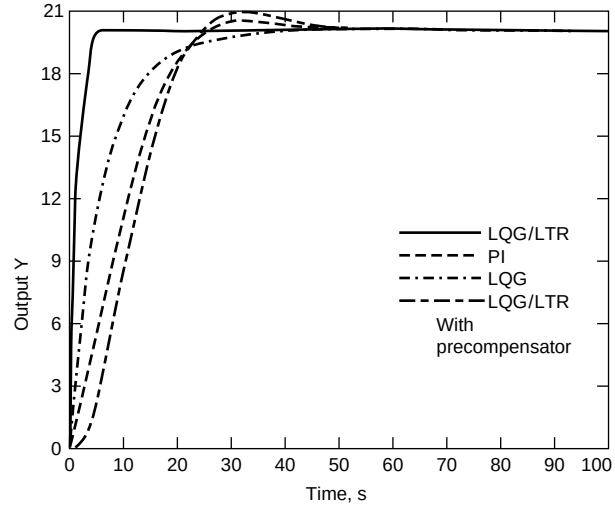


Fig. 16.2.55 PI, LQG, LQG/LTR, LQG/LTR (with precompensator) control system closed-loop transient responses.

precompensator Eq. (16.2.138) can be represented in state equation form as follows:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_n^2 & -\zeta\omega_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} r_i(t) \tag{16.2.139}$$

and

$$r(t) = [\omega_n^2 \quad 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{16.2.140}$$

where $r_i(t)$ is the time-domain representation of the reference input signal and $r(t)$ is the time-domain signal at the output of the precompensator.

Considering the 3 percent overshoot requirement, then ξ must be 0.7. Using the expression

$$t_{\max} = \frac{\pi}{\omega_n\sqrt{1 - (0.7)^2}} \tag{16.2.141}$$

where $t_{\max}$ must be 30 s, gives

$$\omega_n = \frac{\pi}{30\sqrt{1 - (0.7)^2}} \tag{16.2.142}$$

The complete state equation of the precompensator is

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -0.0215 & -0.2052 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} r_i(t) \tag{16.2.143}$$

and

$$r(t) = [0.0215 \quad 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{16.2.144}$$

The resulting transient response of the LQG/LTR control system using the precompensator is shown in Fig. 16.2.55.

ANALYSIS OF SINGULAR-VALUE PLOTS

Systems with large stability margins, good disturbance rejection/command following, low sensitivity to plant parameter variations, and stability in the presence of model uncertainties are described as being robust and having good robustness properties. The singular-value plots for the PI, LQG, and LQG/LTR control systems are examined to evaluate the robustness properties. The singular-value plots of the return ratio, return difference, and the inverse return difference are shown in Figs. 16.2.56, 16.2.57, and 16.2.58.

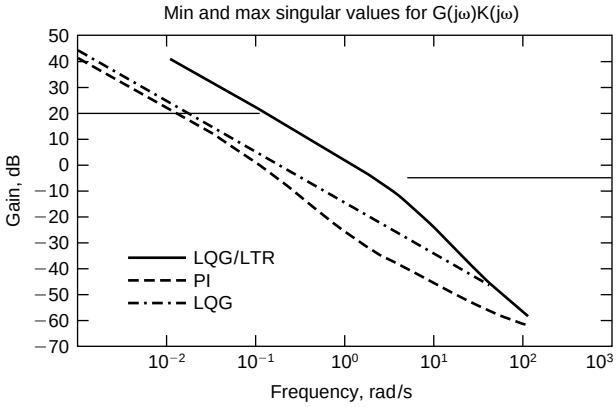


Fig. 16.2.56 Singular-value plots of return ratio.

Deaerator Study Summary and Conclusions

Summary of Control System Analysis The design criteria used for analysis in this system is defined as shown in Table 16.2.8. The performance and robustness results summarized from the analysis using the singular-value plots are shown in Table 16.2.9, from which it appears that the LQG/LTR control system has the widest low-frequency range of disturbance rejection and insensitivity to parameter variations. The PI control system has the worst disturbance rejection property and the most sensitivity to parameter variations.

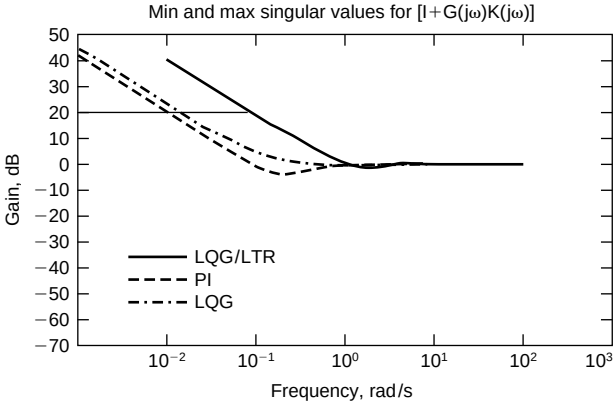


Fig. 16.2.57 Singular-value plots of return difference.

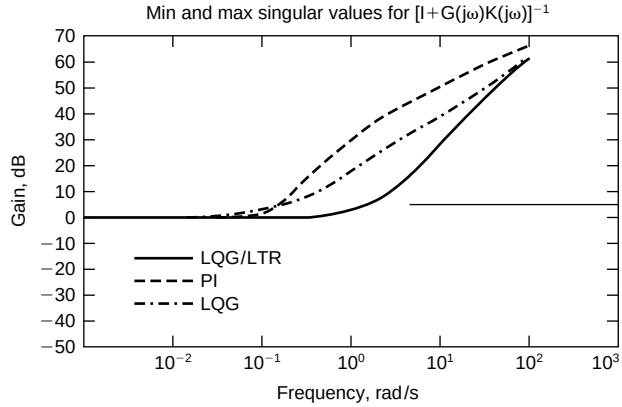


Fig. 16.2.58 Singular-value plots of inverse return difference.

All of the control systems are capable of maintaining a stable robust system in the presence of high-frequency modeling errors, but the PI control system is best. Though all the control systems also have good immunity to noise at frequencies significantly greater than 5 rad/s, the PI control system has the best immunity.

Unlike the other systems, the LQG/LTR control system meets all the design criteria. Its design was obtained in a systematic manner, in contrast to the trial-and-error method used for the LQG control system. The LQG controller design was obtained by shaping the output transient response using the system state weighting matrix, then examining the system's singular-value plots. The PI control system used is simply a three-element control system strategy that has previously been used in the simulation study of the deaerator flow control system.

Conclusions It should be evident that performance characteristics such as a suitable transient response do not imply that the system will have good robustness properties. Obtaining suitable performance characteristics and a stable robust system are two separate goals. Classical unity-feedback design methods have been used for transient response

Table 16.2.8 System Design Specifications

System requirements	Range
Good command-following/ disturbance rejection	$\underline{\sigma}[L(j\omega)] > 20$ dB for $\omega \le 0.1$ rad/s
Good system response to high-frequency modeling error	$\underline{\sigma}(I + [L(j\omega)]^{-1}) \ge \|\Delta L\| = 5$ dB for $\omega \le 5$ rad/s
Good insensitivity to parameter variations at low frequencies	$\underline{\sigma}[I + L(j\omega)] \ge 26$ dB for low frequencies
Good immunity to noise, $\omega \ge 5$ rad/s	$\overline{\sigma}[L(j\omega)] \ll 0$ dB for high frequencies

Table 16.2.9 Performance and Robustness Results

System properties	PI control system	LQG control system	LQG/LTR control system
Command following/disturbance	$\underline{\sigma}[L(j\omega)] \ge 20$ dB for $\omega \le 0.01$ rad/s	$\underline{\sigma}[L(j\omega)] \ge 20$ dB for $\omega \le .017$ rad/s	$\underline{\sigma}[L(j\omega)] \ge 20$ dB for $\omega \le 0.1$ rad/s
System response to high-frequency modeling error ($\omega \ge 5$ rad/s)	$\underline{\sigma}(I + [L(j\omega)]^{-1}) \ge 43.0$ dB	$\underline{\sigma}(I + [L(j\omega)]^{-1}) \ge 30$ dB	$\underline{\sigma}(I + [L(j\omega)]^{-1}) \ge 16.7$ dB
Insensitivity to parameter variations at low frequencies	$\underline{\sigma}[I + L(j\omega)] \ge 26$ dB for $\omega \le 0.0055$ rad/s	$\underline{\sigma}[I + L(j\omega)] \ge 26$ dB for $\omega \le 0.008$ rad/s	$\underline{\sigma}[I + L(j\omega)] \ge 26$ dB for $\omega \le 0.055$ rad/s
Immunity to noise ($\omega \ge 5$ rad/s)	$\overline{\sigma}[L(j\omega)] \le -42.5$ dB	$\overline{\sigma}[L(j\omega)] \le -30$ dB	$\overline{\sigma}[L(j\omega)] \le -16$ dB

shaping, with little consideration for robustness properties. The main advantage of a closed-loop feedback system is that good performance and stability robustness properties are obtainable. As has been shown, transient response shaping can be obtained easily using a precompensator. Therefore it appears that the goal of the control system designer is first to design a stable robust system and then use prefiltering to obtain the desired transient response.

Optimal control methods as demonstrated by the LQG control system do not guarantee good robustness properties when applied systematically to meet minimization requirements of a performance index. Methods such as **pole placement** emphasize transient response characteristics without regard to robustness, which could be detrimental to the system integrity in the presence of model uncertainties.

TECHNOLOGY REVIEW

Fuzzy Control

Fuzzy controllers are a popular method for construction of simple control systems from **intuitive knowledge of a process.** They generally take the form of a set of **IF** . . . **THEN** . . . rules, where the conditional parts of the rules have the structure ''Variable *i* is <fuzzy value 1> AND Variable *j* is <fuzzy value 2> . . . .'' Here the <fuzzy value> refers to a **membership function,** which defines a range of values in which the variable may lie and a number between zero and one for each element of that range specifying the possibility (where certainty is one) that the variable takes that value. Most fuzzy control applications use an **error signal** and the rate of change of an error signal as the two variables to be tested in the conditionals.

The actions of the rules, specified after THEN, take a similar form. Because only a finite collection of membership functions is used, they can be named by mnemonics which convey meaning to the designer. For **example,** a fuzzy controller's rule base might contain the following rules:

IF temperature-error is high and temperature-error-rate is slow
THEN fuel-rate is negative.
IF temperature-error is zero and temperature-error-rate is slow
THEN fuel-rate is zero.

Fuzzy controllers are an attractive control design approach because of the ease with which fuzzy rules can be interpreted and can be made to follow a designer's **intuitive knowledge** regarding the control structure. They are inherently nonlinear, so it is easy to incorporate effects which mimic variable gains by adjusting either the rules or the definitions of the membership functions. Fuzzy controller technology is attractive, in part because of its intuitive appeal and its **nonlinear nature,** but also because of the frequent claim that fuzzy rules can be utilized to approximate arbitrary continuous functional relationships. While this is true, in fact it would require substantial complexity to construct a fuzzy rule base to mimic a desired function sufficiently well for most engineers to label it an approximation. Herein also lie the potential liabilities of fuzzy controller technology: there is very little theoretical guidance on how rules and membership functions should be specified. Fuzzy controller design is an **intuitive process;** if intuition fails the designer, one is left with very little beyond trial and error.

One remedy for this situation is the use of machine, or **automated learning technology** to **infer** fuzzy controller **rules** from either collected data or simulation of a process model. This process is inference, rather than deductive reasoning, in that proper operation of the controller is **learning by example,** so if the examples do not adequately describe the process, problems can occur. This inference process, however, is quite similar to the methods of system identification, which also infer model structure from examples, and which is an accepted method of model construction. Machine learning technology enables utilization of more complex controller structures as well, which introduce additional degrees of freedom in the design process but also enhance the capabilities of the controller. One **example** of this approach is the **Fuzzy PID controller** structure (Wang and Birdwell), which utilizes a fuzzy rule base to generate gains for a PID controller. Other similar approaches have been reported in the literature. An advantage of this approach is the ability to establish theoretical results on closed loop stability, which are otherwise lacking for fuzzy controllers.

Signal Validation Technology

Continuous monitoring of instrument channels in a process industry facility serves many purposes during plant operation. In order to achieve the desired operating condition, the system states must be measured accurately. This may be accomplished by implementing a **reliable signal validation procedure** during both normal and transient operations. Such a system would help reduce challenges on control systems, **minimize plant downtime,** and help plan maintenance tasks. Examples of measurements include pressure, temperature, flow, liquid level, electrical parameters, machinery vibration, and many others. The performance of control, safety, and plant monitoring systems depends on the accuracy of signals being used in these systems. Signal validation is defined as the detection, isolation, and characterization of faulty signals. This technique must be applicable to systems with redundant or single sensor configuration.

Various methods have been developed for signal validation in aerospace, power, chemical, metals, and other process industries. Most of the early development was in the aerospace industry. The signal validation techniques vary in complexity depending on the level of information to be extracted. Both **model-based** and **direct data-based techniques** are now available. The following is a list of **techniques** often implemented for **signal validation:**

Consistency checking of redundant sensors
Sequential probability ratio testing for incipient fault detection
Process empirical modeling (static and dynamic) for state estimation
Computational neural networks for state estimation
Kalman filtering technique for state estimation in both linear and nonlinear systems
Time-series modeling techniques for sensor response time estimation and frequency bandwidth monitoring

PC-based signal validation systems (Upadhyaya, 1989) are now available and are being implemented on-line in various industries. The computer software system consists of one or more of the above signal processing modules, with a decision maker that provides sensor status information to the operator.

An example of **signal validation** (Upadhyaya and Eryurek) in a power plant is the monitoring of water level in a steam generator. The objective is to estimate the water level using other related measurements and compare this with the measured level. An empirical model, a neural network model, or the Kalman filtering technique may be used. For example, the inputs to an empirical model consist of steam generator main feedwater flow rate, steam generator pressure, and inlet and outlet temperatures of the primary water through the steam generator. Such a model would be developed during normal plant operation, and is generally referred to as the *training phase*. The signal estimation models should be updated according to the plant operational status. The use of multiple signal validation modules provides a high degree of confidence in the results.

Chaos

Recent advances in dynamical process analysis have revealed that nonlinear interactions in simple deterministic processes can often result in highly complex, aperiodic, sometimes apparently ''random'' behavior. Processes which exhibit such behavior are said to be exhibit *deterministic chaos*. The term *chaos* is perhaps unfortunate for describing this behavior. It evokes images of purely erratic behavior, but this phenomenon actually involves highly structured patterns. It is now recognized that deterministic chaos dominates many engineering systems of practical interest, and that, in many cases, it may be possible to exploit previously unrecognized deterministic structure for improved understanding and control.

Deterministic chaos and nonlinear dynamics both apply to phenomena that arise in process equipment, motors, machinery, and even control systems as a result of nonlinear components. Since nonlinearities are almost always present to some extent, nonlinear dynamics and chaos are the rule rather than the exception, although they may sometimes occur to such a small degree that they can be safely ignored. In many cases, however, nonlinearities are sufficiently large to cause significant changes in process operation. The effects are typically seen as unstable operation and/or apparent noise that is difficult to diagnose and control with conventional linear methods. These instabilities and noise can cause reduced operating range, poor quality, excessive downtime and maintenance, and, in worst cases, catastrophic process failure.

The apparent erratic behavior in chaotic processes is caused by the extreme sensitivity to initial conditions in the system, introduced by nonlinear components. Though the system is deterministic, this sensitivity and the inherent uncertainty in the initial conditions make long-term predictions impossible. In addition, small external disturbances can cause the process to behave in unexpected ways if this sensitivity has not been considered. Conversely, if one can learn how the process reacts to small perturbations, this inherent sensitivity can be taken advantage of. Once the system dynamics are known, small changes in system parameters can be used to drive the system to a desired state and to keep it there. Thus, great effects can be achieved through minimal input. Various control algorithms have been developed that use this principle to achieve their goals. A recent example from engineering is the control of a slugging fluidized bed by Vasadevan et al.

The following example illustrates the use of chaos analysis and control which involves a laboratory fluidized-bed experiment. Several recent studies have recognized that some modes of fluidization in fluidized beds can be classified as deterministic chaos. These modes involve large scale cyclic motions where the mass of particles move in a piston-like manner up and down in the bed. This mode is usually referred to as *slugging* and is usually considered undesirable because of its inferior transfer properties. Slugging is notoriously difficult to eradicate by conventional control techniques. Vasadevan et al. have shown that the inherent sensitivity to small perturbations can be exploited to alleviate the slugging problem. They did this by adding an extra small nozzle in the bed wall at the bottom of the bed. The small nozzle injected small pulses of process gas into the bed, thus disrupting the ''normal'' gas flow. The timing of the pulses was shown to be crucial to achieve the desired effect. Different injection timing caused the slugging behavior to be either enhanced or destroyed.

16.3 SURVEYING

by W. David Teter

REFERENCES: Moffitt, Bouchard, ''Surveying,'' HarperCollins. Wolf and Brinker, ''Elementary Surveying,'' Harper & Rowe. Kavanaugh and Bird, ''Surveying Principles and Applications,'' Prentice-Hall. Kissam, ''Surveying for Civil Engineers,'' McGraw-Hill. Anderson and Mikhail, ''Introduction to Surveying,'' McGraw-Hill. McCormac, ''Surveying Fundamentals,'' Prentice-Hall.

INTRODUCTION

Surveying is often defined as the art and science of measurement for location or establishment of position above, on, or beneath the earth's surface. The principles of surveying practice have remained consistent from their earliest inception, but in recent years the equipment and technology have changed rapidly. The emergence of the total station device, electronic distance measurements (EDM), Global Positioning System (GPS), and geographic information systems (GIS) is of significance in comparing modern surveying to past practice.

The information required by a surveyor remains basically the same and consists of the measurement of direction (angle) and distance, both horizontally and vertically. This requirement holds regardless of the type of survey such as land boundary description, topographic, construction, route, or hydrographic.

HORIZONTAL DISTANCE

Most surveying and engineering measurements of distance are horizontal or vertical. Land measurement referenced to a map or plat is reduced to horizontal distance regardless of the manner in which the field measurements are made.

The methods and devices used by the surveyor to measure distance include pacing, odometers, tachometry (stadia), steel tape, and electronic distance measurement. These methods produce expected precision ranging from ± 2 percent for pacing to + 0.0003 percent for EDM. The surveyor must be aware of the necessary precision for the given application.

Tapes

Until the advent of EDM technology, the primary distance-measuring instrument was the steel tape, usually 100 ft or 30 m (or multiples) in length, with 1 ft or 1 dm on the end graduated to read with high precision (0.01 ft or 1 mm). Cloth tapes can be used when low precision is acceptable.

Corrections Measurements with a steel tape are subject to variations that normally must be corrected for when high precision is required. Tapes are manufactured to a standard length at usually 68°F (20°C), for a standard pull of between 10 and 20 lb (4.5 and 9 kg), and supported over the entire length. The specifications for the tape are provided by the manufacturer. If any of the conditions vary during field measurement, corrections for temperature, pull, and sag will have to be made according to the manufacturer's instructions.

Tape Use Most surveying measurements are made with reference to the horizontal, and if the terrain is level, involves little more than laying down the tape in a sequential manner to establish the total distance. Care should be taken to measure in a straight line and to apply constant tension. When measuring on a slope, either of two methods may be employed. In the first method, the tape is held horizontal by raising the low end and employing a plumb bob for location over the point. In cases of steep terrain a technique known as *breaking tape* is used, where only a portion of the total tape length is used. With the second method, the slope distance is measured and converted by trigonometry to the horizontal distance. In Fig. 16.3.1 the horizontal distance $D = S \cos(\alpha)$ where S is the slope distance and α is the slope angle which must be estimated or determined.

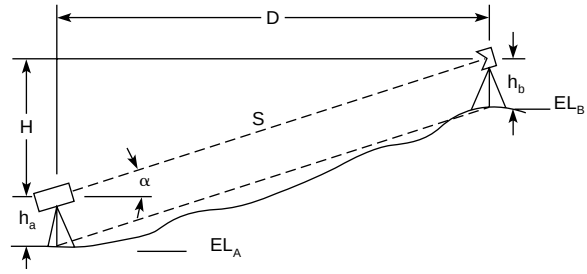


Fig. 16.3.1 Slope-reduction measurements used with EDM devices.

Electronic Distance Measurement

Modern surveying practice employs EDM technology for precise and rapid distance measurement. A representation of an EDM device is shown in Fig 16.3.2. The principle of EDM is based on the comparison of the modulated wavelength of an electromagnetic energy source (beam) to the time required for the beam to travel to and return from a point at an unknown distance. The EDM devices may be of two types: (1) electrooptical devices employing light transmission within or just beyond the visible region or (2) devices transmitting in the microwave spectrum. Electrooptical devices require an active transmitter and a passive reflector, while microwave devices require an active transmitter

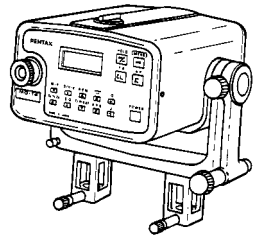


Fig. 16.3.2 Electronic distance meter (EDM). *(PENTAX Corp.)*

and an identical unit for reception and retransmission at the endpoint of the measured line. Ideally the EDM beam would propagate at the velocity of light; however, the actual velocity of propagation is affected by the index of refraction of the atmosphere. As a result, the atmospheric conditions of temperature, pressure, and humidity must be monitored in order to apply corrections to the measurements. Advances in electrooptical technology have made microwave devices, which are highly sensitive to atmospheric relative humidity, nearly obsolete.

Atmospheric Corrections EDM devices built prior to about 1982 require a manual computation for atmospheric correction; however, the more modern instruments allow for keystroke entry of values for temperature and pressure which are processed and applied automatically as a correction for error to the output reading of distance. The relationships between temperature, pressure, and error are shown in Fig. 16.3.3. Microwave devices require an additional correction for relative humidity. A psychrometer wet-dry bulb temperature difference of 3°F (2°C) causes an error of about 0.001 percent in distance measurement.

Instrument Corrections Electrooptical devices may also be subject to error attributed to *reflector constant.* This results from the physical

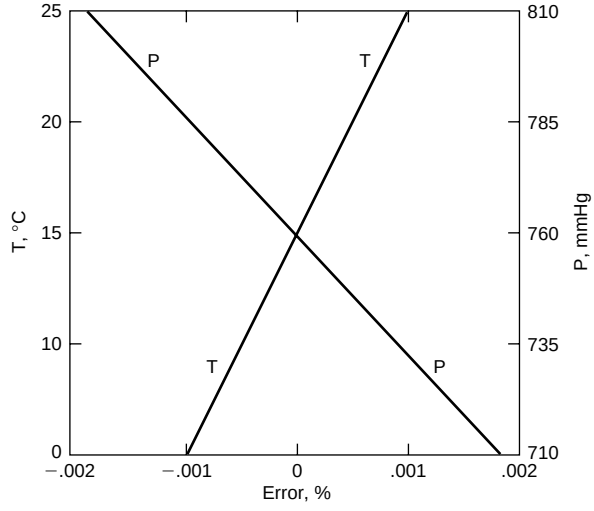


Fig. 16.3.3 Relationship of error to atmospheric pressure and temperature.

center of the reflector not being coincident with the effective or optical center. The error can range to 30 or 40 mm and is unique, but is usually known, for each reflector. The older EDM devices required this value to be subtracted from each reading, but in more modern devices the reflector constant can be preset into the instrument and the compensation occurs automatically. Finally, microwave EDM devices are subject to error from ground reflection which is known as *ground swing.* Care must be taken to minimize this effect by deploying the devices as high as possible and averaging multiple measurements.

EDM Field Practice Practitioners have developed many variations in employing EDM devices. Generally the field procedures are governed by the fact that the distance readout from the device is a slope measurement which must be reduced to the horizontal equivalent. Newer EDM devices may allow for automatic reduction by keystroke input of the vertical angle, if known, as would be the case when the EDM device is employed with a theodolite or if the device is a total station (see later in the section). Otherwise the slope reduction can be accomplished by using elevation differences. With reference to Fig. 16.3.1, it is seen that $H = (\mathrm{EL}_A + h_a) - (\mathrm{EL}_B + h_b)$ and that the horizontal distance $D = (S^2 - H^2)^{1/2}$. EDM devices are sometimes mounted on a conventional theodolite. In such cases the vertical angle as measured by the theodolite must be adjusted to be equivalent to the angle seen by the EDM device. It is common to sight the theodolite at a point below the reflector equal to the vertical distance between the optical centers of the theodolite and the EDM device.

VERTICAL DISTANCE

The acquisition of data for vertical distance is also called **leveling**. Methods for the determination of vertical distance include direct measurement, tachometry or stadia leveling, trigonometric leveling, and differential leveling. Direct measurement is obvious, and stadia leveling is discussed later.

Trigonometric Leveling This method requires the measurement of the vertical angle and slope distance between two points and is illustrated in Fig. 16.3.1. The vertical distance $H = S \sin(\alpha)$. A precise value for the vertical angle and slope distance will not be obtained if the EDM device is mounted on the theodolite or the instrument heights h_a and h_b are not equal. See ''EDM Field Practice,'' above. Before the emergence of EDM devices, the slope distance was simply taped and the vertical angle measured with a theodolite or transit.

If the elevation of an inaccessible point is required, the EDM reflector cannot be placed there and a procedure employing two setups of a transit or theodolite may be used, as shown in Fig. 16.3.4. Angle $Z_3 = Z_1 - Z_2$, and the application of the law of sines shows that distance $A/\sin Z_2 = C/\sin Z_3$. C is the measured distance between the two instrument setups. The vertical distance $H = A \sin Z_1$. The height of the stack above the instrument point is equal to $H + h_1$, where h_1 is the height of the instrument.

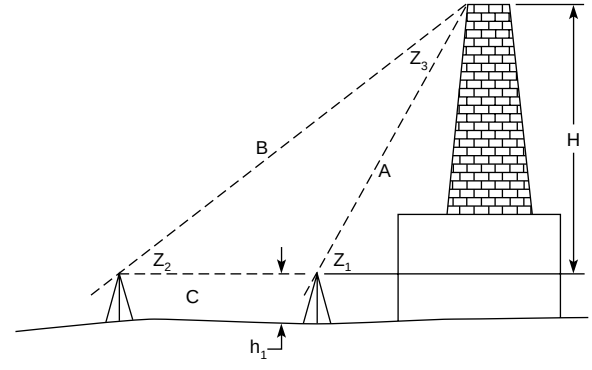


Fig. 16.3.4 Determination of elevation of an inaccessible point.

Differential Leveling This method requires the use of an instrument called a **level,** a device that provides a horizontal line of sight to a rod graduated in feet or metres. Figure 16.3.5 shows a vintage level, and Fig. 16.3.6 shows a modern ''automatic'' level. The difference is that with the older device, four leveling screws are used to center the spirit-level bubble in two directions so that the instrument is aligned with the

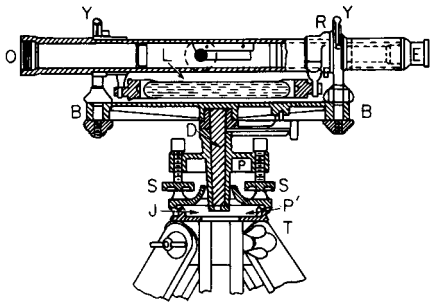


Fig. 16.3.5 Y level.

horizontal. The modern device adjusts its own optics for precise levelness once the three leveling screws have been used to center the rough-leveling/circular level bubble within the scribed circle. The instrument legs should be planted on a firm footing. The eyepiece is checked and adjusted to the user's eye for clear focus on the crosshairs, and the objective-lens focus knob is used to obtain a clear image of the graduated rod.

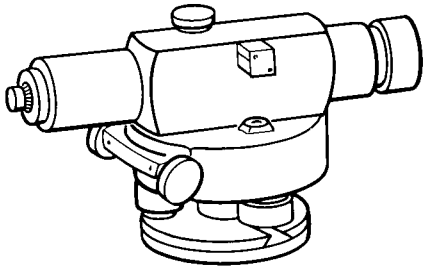


Fig. 16.3.6 Automatic (self-adjusting) level.

To determine the difference in elevation between two points, set the level nearly midway between the points, hold a rod on one, look through the level and see where the line of sight, as defined by the eye and horizontal cross wire, cuts the rod, called **rod reading**. Move the rod to the second point and read. The difference of the readings is the difference in level of the two points. If it is impossible to see both points from a single setting of the level, one or more intermediate points, called **turning points,** are used. The readings taken on points of known or assumed elevation are called **plus sights,** those taken on points whose elevations are to be determined are called **minus sights**. The elevation of a point plus the rod readings on it gives the elevation of the line of sight; the elevation of the line of sight minus the rod reading on a point of unknown elevation gives that elevation. In Fig. 16.3.7, I_1 and I_2 are intermediate points between *A* and *B*. The setups are numbered. Assuming *A* to be of known elevation, the reading on *A* is a + sight; the reading on I_1 from 1 is a − sight; the reading on I_1 from 2 is a + sight and on I_2 is a

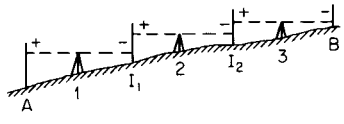


Fig. 16.3.7 Determining elevation difference with intermediate points.

− sight. The algebraic sum of the plus and minus sights is the difference of elevation between *A* and *B*. Target rods can usually be read by vernier to thousandths of a foot. In grading work the nearest tenth of a foot is good; in lining shafting the finest possible reading is none too good. It is desirable that the sum of the distances to the plus sights approximately equal the sum of the distances to the minus sights to ensure compensation of errors of adjustment. On a side hill this can be accomplished by zigzagging. When the direction of pointing is changed, the position of the level bubble should be checked. In the case of a vintage instrument, the bubble should be adjusted back to a centered position along the spirit level. An automatic level requires the bubble to be repositioned within the centering circle.

To Make a Profile of a Line A **bench mark** is a point of reasonably permanent character whose elevation above some surface—as sea level—is known or assumed and used as a reference point for elevation. The level is set up either on or a little off the line some distance—not more than about 300 ft (90 m)—from the starting point or a convenient benchmark (BM), as at *K* in Fig. 16.3.8. A reading is taken on the BM and added to the known or assumed elevation to get the height of the instrument, called HI. Readings are then taken at regular intervals (or **stations**) along the line and at such irregular points as may be necessary to show change of slope, as at *B* and *C* between the regular points. The regular points are marked by stakes previously set ''on line'' at distances of 100 ft (30 m), 50 ft (15 m), or other distance suitable to the

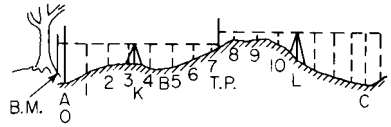


Fig. 16.3.8 Making a profile.

character of the ground and purpose of the work. When the work has proceeded as far as possible—not more than about 300 ft (90 m) from the instrument for good work—a **turning point** (TP) is taken at a regular point or other convenient place, the instrument moved ahead and the operation continued. The first reading on the BM and the first reading on a TP after a new setup are plus sights (+ S); readings to points along the line and the first reading on a TP to be established are minus sights (− S).

The **notes** are taken in the form shown in Fig. 16.3.9. The elevation of a given point, both sights taken on it, and the HI determined from it all appear on a line with its station (Sta) designation. In plotting the **profile,** the vertical scale is usually exaggerated from 10 to 20 times.

Inspection and Adjustment of Levels Leveling instruments are subject to maladjustment through use and should be checked from time to time. Vintage instruments require much more attention than modern devices. Generally, the devices should be checked for verticality of the crosshair, proper orientation of the leveling bubble, and the optical line of sight. The casual user would normally not attempt to physically adjust an instrument, but should be familiar with methods for inspection for maladjustment as follows.

1. *Verticality of crosshair.* Set up the level and check for coincidence of the vertical crosshair on a suspended plumb line or vertical corner of a building. The crosshair ring must be rotated if this condition is not satisfied. This test applies to all types of instruments.

2. *Alignment of the bubble tube.* This check applies to older instruments having a spirit level. In the case of a Y level, the instrument is carefully leveled and the telescope removed from the Ys and turned end for end. In the case of a dumpy level, the instrument is carefully leveled and then rotated 180°. In each case, if the bubble does not remain centered in the new position, the bubble tube requires adjustment.

3. *Alignment of the circular level.* This check applies to tilting levels and modern automatic levels equipped with a circular level consisting of a bubble to be centered within a scribed circle. The instrument is carefully adjusted to center the bubble within the leveling circle and

Left-hand page

This space for a heading, telling what the work is, who does it, and the date on which it is done.

Sta.	+ S	H.I.	− S	Elev.
B.M.	6.42	506.42	.....	500.0
$A = 0$	.....	.....	10.4	496.0
1	.....	.....	8.2	98.2
2	.....	.....	6.1	500.3
+30	.....	.....	5.5	0.9
3	.....	.....	6.1	0.3
4	.....	.....	7.9	498.5
$B = +40$	.....	.....	8.4	98.0
5	.....	.....	7.5	98.9
6	.....	.....	5.1	501.3
7	.....	.....	3.2	3.2
+ 10 T.P.	4.27	509.13	1.56	504.86
8	.....	.....	2.2	506.9

Right-hand page

This page is for remarks describing B.M.s and T.P.s or other important particulars.

Fig. 16.3.9 Form for surveyor's notes.

then rotated through 180°. If the bubble does not remain centered, adjustment is required.

4. *Line of sight coincident with the optical axis.* This check (commonly called the *two-peg* test) applies to all instruments regardless of construction. The check is performed by setting the instrument midway between two graduated rods and taking readings on both. The instrument is then moved to within the 6 ft of one of the rods, and again readings are taken on both. The difference in readings should agree with the first set. If they do not, an adjustment to raise or lower the crosshair is necessary.

ANGULAR MEASUREMENT

The instruments used for measurement of angle include the transit (Fig. 16.3.10), the optical theodolite, and the electronic total station (Fig. 16.3.11). Angular measurements in both horizontal and vertical planes are obtained with the quantitative value derived in one of three ways depending on the era in which the instrument was manufactured. In principle, the effective manner in which these devices operate is shown in Fig. 16.3.12. The graduated outer circle is aligned at zero with the inner-circle arrow, and an initial point *A* is sighted on. Then, with the outer circle held fixed by means of clamps on the instrument, the inner circle (with pointer and moving with the telescope) is rotated to a position for sighting on point *B*. The relative motion between the two circles describes the direction angle which is read to precision with the aid of the scale's vernier. In the actual case, only the transit operates in the mechanical manner described. The operation of an optical theodolite occurs internally with the user viewing internal scales that move relative to each other. Many optical theodolites have scale microscopes or scale micrometers for precise reading. The electronic total station (discussed later) is also dependent on internal optics, but in addition it gives its angle values in the form of a digital display.

Examination of these devices reveals the presence of an *upper clamp* (controlling the movement of the inner circle described before) and a *lower clamp* (controlling the movement of the outer circle). Both clamps have associated with them a *tangent* screw for very fine adjustment. Setting up the instrument involves leveling the device and, unlike

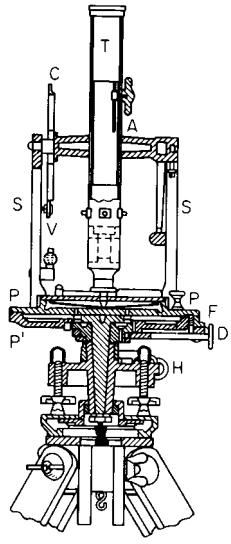


Fig. 16.3.10 Vintage transit.

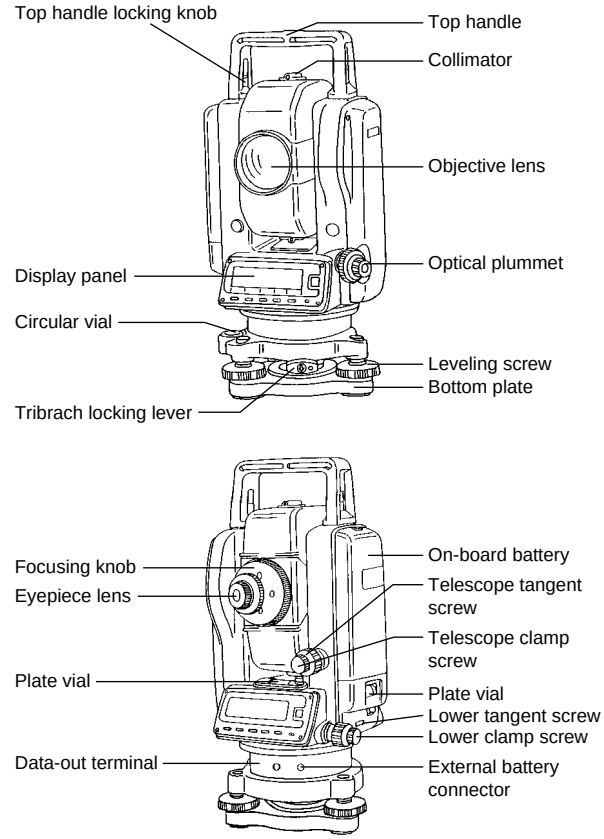


Fig. 16.3.11 Electronic total station. *(PENTAX Corp.)*

the procedure with a level, locating the instrument exactly over a particular point (the apex of the angle). This is accomplished with a suspended plumb (in the case of older transits) or an optical (line-of-sight) plummet in modern instruments. Prior to any measurement the clamps are manipulated so as to zero the initial angular reading prior to turning the angle to be read. In the case of the digital output total station, the initial angular reading is zeroed by pushing the zero button on the display. A total station, in effect, has only one motion clamp.

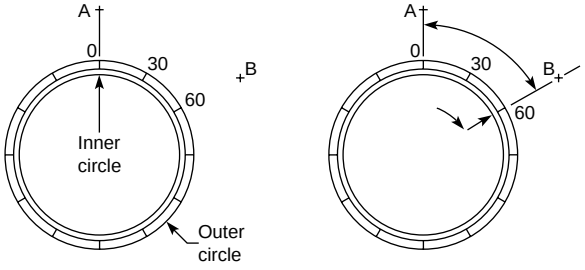


Fig. 16.3.12 Use of graduated circles to measure the angles between points.

Angle Specification Figure 16.3.13 shows several ways in which direction angles are expressed. The angle describing the direction of a line may be expressed as a **bearing,** which is the angle measured east or west from the north or south line and not exceeding 90°, e.g.: N45°E, S78°E, N89°W. Another system for specifying direction is by **azimuth** angle. Azimuths are measured clockwise from north (usually) up to 360°. Still other systems exist where direction is expressed by angle to the right (or left), deflection angles, or interior angles.

To Produce a Straight Line Set up the instrument over one end of the line; with the lower motion clamp and tangent screw bring the telescopic line of sight to the other end of the line marked by a flag, a pencil, a pin, or other object; transit the telescope, i.e., plunge it by revolving on its horizontal axis, and set a point (drive a stake and ''center'' it with a tack or otherwise) a desired distance ahead in line with the telescopic line of sight; loosen the lower motion clamp and turn the instrument in azimuth until the line of sight can be again pointed to the other end of the line; again transit and set a point beside the first point set. If the instrument is in adjustment, the two points will coincide; if not, the point marking the projection of the line lies midway between the two established points.

To Measure a Horizontal Angle Set up the instrument over the apex of the angle; with the lower motion bring the line of sight to a distant point in one side of the angle; unclamp the upper motion and bring the line of sight to a distant point in the second side of the angle, clamp and set exactly with the tangent screw; read the angle as displayed using scale micrometer or vernier if required.

To Measure a Vertical Angle Set up the instrument over a point marking the apex of the angle *A* (see Fig. 16.3.14) by the lower motion and the motion of the telescope on its horizontal axis, bring the intersection of the vertical and horizontal wires of the telescope in line with a point as much above the point defining the lower side of the angle as the telescope is above the apex; read the vertical angle, turn the telescope to a point which is the height of the instrument above the point marking the upper side of the angle and read the vertical angle. How to combine the readings to find the angle will be obvious.

To Run a Traverse A traverse is a broken line marking the line of a road, bank of a stream, fence, ridge, or valley, or it may be the boundary of a piece of land. The bearing or azimuth and length of each portion of the line are determined, and this constitutes ''running the traverse.''

To Establish Bearing The bearing of a line may be specified relative to a north-south line (meridian) that is true, magnetic, or assumed, and whose angle is less than 90°. Whatever the reference meridian, the instrument is set over one end of the line, and the horizontal angle readout is set to zero with the instrument pointing in the direction of the north-south meridian. This is shown in Fig. 16.3.12 when point *A* is in the direction of the meridian (north). Using appropriate clamps, the angle is turned by pointing the telescope to the other end of the desired line. In Fig. 16.3.12 the bearing angle would be N60°E. Note that modern electron total stations generally do not have a magnetic needle and, as a result, the bearings or azimuths are typically measured relative to an arbitrary (assumed) meridian.

To establish azimuth the same procedure as described for the determination of bearing may be used. An azimuth may have values up to 360°. When the preceding line of a traverse is used to orient the instrument, the azimuth of this line is known as the **back azimuth** and its value should be set into the instrument as it is pointed along the preceding line

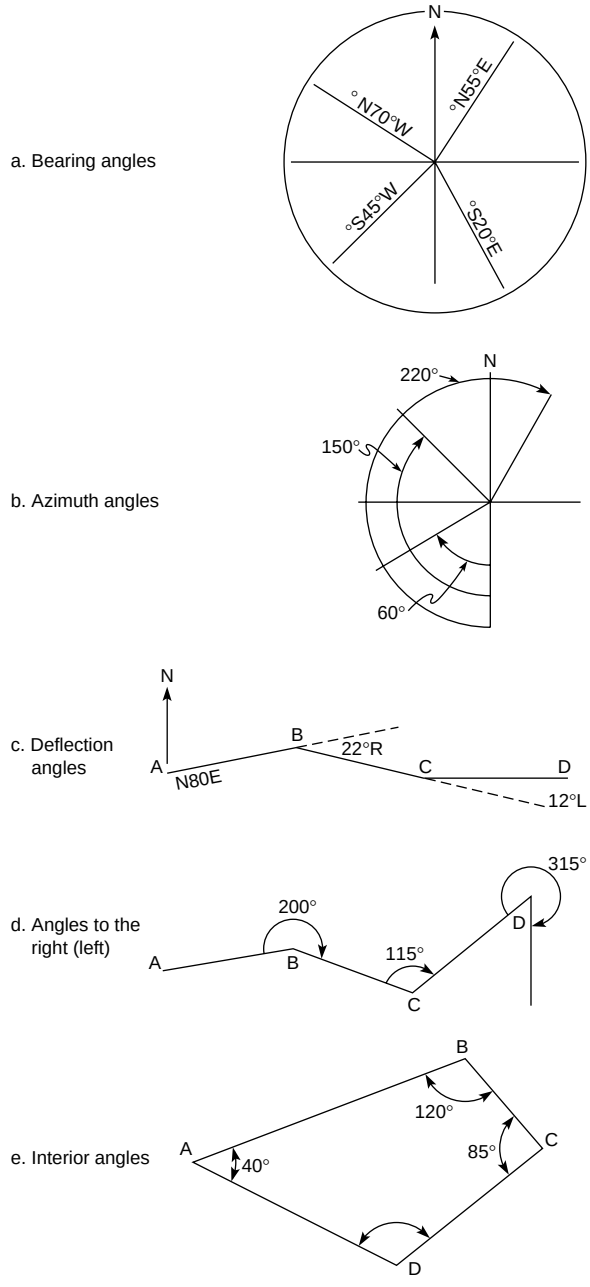


Fig. 16.3.13 Diagram showing various methods of specifying direction angles.

of the traverse. The telescope is turned clockwise to the next line of the traverse to establish the **forward azimuth**.

In traverse work, azimuths and bearings are not usually measured except for occasional checks. Instead, deflection angles or angles to the right from one course to the next are measured. The initial line (course)

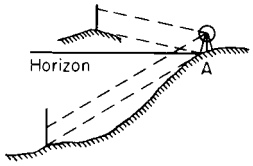


Fig. 16.3.14 Measuring a vertical angle.

is taken as an assumed meridian, and the bearings or azimuths of the other courses with respect to the initial course are calculated. The calculations are derived from the measured deflection angles or angles to the right. If the magnetic or true meridian for the initial course is determined, then the derived bearings or azimuths can likewise be adjusted.

In Fig. 16.3.15, the bearing of a is N40°E, of b is N88° 30′E, of c is S49°20′E = 180° − (40° + 48°30′ + 42°10′), of d is S36°40′W = 86° − 49°20′, or 40° + 48°30′ + 42°10′ + 86° − 180°, of e is N81°20′W = 180° − (36°40′ + 62°). No azimuths are shown in Fig. 16.3.15, but note that the azimuth of a is 40° and the back azimuth of a is 220° (40° + 180°). The azimuths of b and c are 88°30′ and 130°40′ respectively.

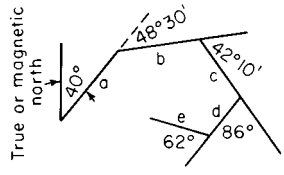


Fig. 16.3.15 Bearings derived from measured angles.

Inspection and Adjustment of the Instrument The transit, optical theodolite, and total station devices are subject to maladjustment through use and should be checked from time to time. Regardless of the type of instrument, the basic relationships that should exist are (1) the plate level-bubble tubes must be horizontal, (2) the line of sight must be perpendicular to the horizontal axis, (3) the line of sight must move in the vertical plane, (4) the bubble (if there is one) of the telescope tube must be centered when the scope is horizontal, and (5) the reading of vertical angles must be zero when the instrument is level and the telescope is level.

All modern devices are subject to the following checks: (1) adjustment of the plate levels, (2) verticality of the crosshair, (3) line of sight perpendicular to the horizontal axis of the telescope, (4) the line of sight must move in a vertical plane, (5) the telescope bubble must be centered (two-peg test), (6) the vertical circle must be indexed, (7) the circular level must be centered, (8) the optical plummet must be vertical, (9) the reflector constant should be verified, (10) the EDM beam axis and the line of sight must be closely coincident, and (11) the vertical and horizontal zero points should be checked.

To Measure Distances with the Stadia In the transit and optical theodolite telescope are two extra horizontal wires so spaced (when fixed by the maker) that they are 1/100 of the focal length of the objective apart. When looking through the telescope at a rod held in a vertical position, 100 times the rod length S intercepted between the two extra horizontal wires plus an instrumental constant C is the distance D from the center of the instrument to the rod if the line of sight is horizontal, or $D = 100\,S + C$ (see Fig. 16.3.16). If the line of sight is inclined by a vertical angle A, as in Fig. 16.3.17, then if S is the space intercepted on the rod and C is the instrumental constant, the distance is given by the formula $D = 100\,S \cos^2 \mathrm{A} + \mathrm{C} \cos A$. For angles less than 5 or 6°, the distance is given with sufficient exactness by $D = 100S$. Although theory would indicate that distances can thus be determined to within 0.2 ft, in practice it is not well to rely on a precision greater than the nearest foot for distances of 500 ft or less.

Only the oldest transits, known as *external-focusing devices*, have an instrument constant. They can be recognized by noting that the objective lens will physically move as the focus knob is turned. Such devices

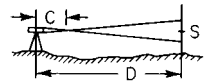


Fig. 16.3.16 Horizontal stadia measurement.

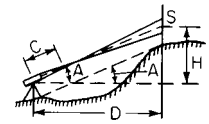


Fig. 16.3.17 Sloping stadia measurement.

have not been manufactured for at least 50 years, and therefore it is most probable that, for stadia applications, the value of $C = 0$ should be used in the equations. Likewise, it is certain that the stadia interval factor K in the following equation will be exactly 100. Stadia surveying falls into the category of low-precision work, but in certain cases where low precision is acceptable, the speed and efficiency of stadia methods is advantageous.

Stadia Leveling This method is related to the previously described trigonometric leveling. The elevation of the instrument is determined by sighting on a point of known elevation with the center crosshair positioned at the height of the instrument (HI) at the setup position. Record the elevation angle A from the setup point to the distant point of known elevation, read the rod intercept S, and apply the following equation to determine the difference in elevation H between the known point and the instrument:

$$H = KS \cos A \sin A + C \sin A$$

Since K is sure to be equal to 100 and that the stadia constant C for most existing instruments will be equal to zero, the value of H is easily determined and the elevation of the instrument HI is known. With known HI, sights can now be performed on other points and the above equation applied to determine the elevation differences between the instrument and the selected points. Note also that the horizontal distance D can be determined by a similar equation:

$$D = \tfrac{1}{2}\, KS \sin 2A + C \cos A$$

Stadia Traverse A stadia traverse of low precision can be quickly performed by recording the stadia data for rod intercept S and vertical angle A plus the horizontal direction angle (or deflection angle, angle to the right, or azimuth) for each occupied point defining a traverse. Application of the equation for distance D eliminates the need for laborious taping of a distance. Note that modern surveying with EDM devices makes stadia surveying obsolete.

Topography Low-precision location of topographic details is also quickly performed by stadia methods. From a known instrument setup position, the direction angle to the landmark features, along with the necessary values of rod intercept S and vertical angle A, allows computation of the vertical and horizontal location of the landmark position from the foregoing equations for H and D. If sights are made in a regular pattern as described in the next paragraph, a contour map can be developed.

Contour Maps A contour map is one on which the configuration of the surface is shown by lines of equal elevation called **contour lines**. In Fig. 16.3.18, contour lines varying by 10 ft in elevation are shown. H, H are hill peaks, R, R ravines, S, S saddles or low places in the ridge $HSHSH$. The horizontal distance between adjacent contours shows the distance for a fall or rise of the contour interval—10 ft in the figure. A profile of any line as AB can be made from the contour map as shown in the lower part of the figure. Conversely, a contour map may be made from a series of profiles, properly chosen. Thus, a profile line run along the ridge $HSHSH$ and radiating profile lines from the peaks down the hills and from the saddles down the ravines would give data for project-

ing points of equal elevation which could be connected for contour lines. This is the best method for making contour maps of very limited areas, such as city squares, or very small parks. If the ground is not too much broken, the small tract is divided into squares and elevations are taken at each square, corner, and between two corners on some lines if necessary to get correct profiles.

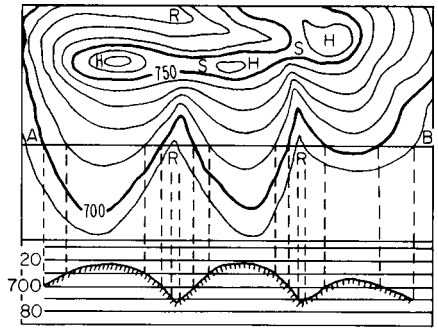


Fig. 16.3.18 Contour map.

SPECIAL PROBLEMS IN SURVEYING AND MEASURATION

Volume of Earth in Foundation and Area Grading The volume of earth removed from a foundation pit or in grading an area can be computed in several ways, of which two follow.

1. The area (Fig. 16.3.19) is divided into squares or rectangles, elevations are taken at each corner before and after grading, and the volumes are computed as a series of prisms. If A is the area (ft^2) of one of the squares or rectangles—all being equal—and b_1, b_2, b_3, b_4 are corner heights (ft) equal to the differences of elevation before and after grading, the subscripts referring to the number of prisms of which b is a corner, then the volume in cubic yards is

$$Q = A(\Sigma h_1 + 2\Sigma h_2 + 3\Sigma h_3 + 4\Sigma h_4) / (4 \times 27)$$

In Fig. 16.3.19 the h's at A_0, D_0, D_3, C_5, and A_5, would be h_1's; those at B_0, C_0, D_1, D_2, C_4, B_5, A_4, A_3, A_2, and A_1 would be h_2's; that at C_3 an h_3; and the rest h_4's. The rectangles or squares should be of such size that their tops and bottoms are practically planes.

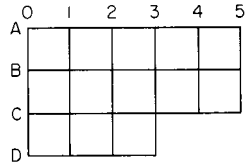


Fig. 16.3.19 Estimating volume of earth by squares.

2. A large-scale profile of each line one way across the area is carefully made, as the *A, B, C* and *D* lines of Fig. 16.3.19, the final grade line is drawn on it, and the areas in excavation and embankment are separately measured with a planimeter or by estimation from the drawing. The excavation area of profile *A* is averaged with that of profile *B*, and the result multiplied by the distance *AB* and divided by 27 to reduce to cubic yards. Similarly, the material between *B* and *C* is found.

To Pass an Obstacle Four cases are shown in Fig. 16.3.20. If the obstacle is large, as a building, (1) turn right angles at *B, C, D* and *E*, making $BC = DE$ when $CD = BE$. All distances should be long enough to ensure sufficiently accurate sighting. (2) At *B* turn the angle *K* and measure *BC* to a convenient point. At *C* turn left $= 360° - 2K$; measure $CD = BC$. At *D* turn *K* for line *DE*. $BD = 2BC \cos (180° - K)$. (3) At *B* lay off a right angle and measure *BC*. At *C* measure any angle to clear object and measure $CD = BC/\cos C$. At *D* lay off $K = 90° + C$ for the line *DE*. $BC = BC \times \tan C$. If the obstacle is small, as a tree. (4) at *A*, some distance back, turn the small angle *a* necessary to pass the obstacle and measure *AB*. At *B* turn the angle $2a$ and measure $BC = AB$. At *C* turn the small angle *a* for the line *AC*, and transit, or turn the large angle $K = 180° - a$. If *a* is but a few minutes of arc, $AC = AB + BC$ with

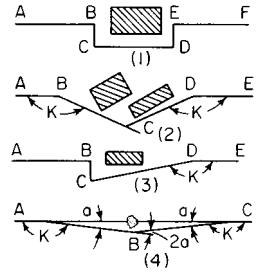


Fig. 16.3.20 Surveying past an obstacle.

sufficient exactness. If only a tape is available, the right-angle method (1) above given may be used, or an equilateral triangle, *ABC* (Fig. 16.3.21) may be laid out, *AC* produced a convenient distance to *F*, the similar triangle *DEF* laid out, *FE* produced to *H* making $FH = AF$, and the similar triangle *GHI* then laid out for the line *GH*. $AH = AF$.

To Measure the Distance across a Stream To measure *AB*, Fig. 16.3.22, *B* being any established point, tree, stake, or building corner: (1) Set the instrument over *A*; turn a right angle from *AB* and measure any distance *AC*; set over *C* and measure the angle *ACB*. $AB = AC \tan ACB$. (2) Set over *A*, turn any convenient angle BAC' and measure AC';

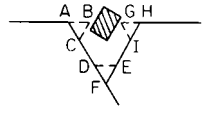


Fig. 16.3.21 Surveying past an obstacle by using an equilateral triangle.

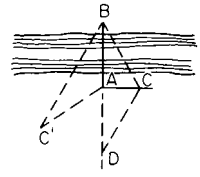


Fig. 16.3.22 Measuring across a stream.

set over C' and measure $AC'B$. Angle $ABC' = 180° - AC'B - BAC'$. $BA = AC' \times \sin AC'B / \sin ABC'$. (3) Set up on *A* and produce *BA* any measured distance to *D*; establish a convenient point *C* about opposite *A* and measure *BAC* and *CAD;* set over *D* and measure *ADC;* set over *C*, and measure *DCA* and *ACB;* solve *ACD* for *AC*, and *ABC* for *AB*. For best results the acute angles of either method should lie between 30 and 60°.

To Measure a Visible but Inaccessible Distance (as *AB* in Fig. 16.3.23) Measure *CD*. Set the instrument at *C* and measure angles *ACB* and *BCD;* set at *D* and measure angles *CDA* and *ADB*. $CAD = 180° - (ACB + BCD + CDA)$. $AD = CD \times \sin ACD / \sin CAD$. $CBD = 180° - (BCD + CDA + ADB)$. $BD = CD \times \sin BCD / \sin CBD$. In the triangle *ABD*, $\frac{1}{2}(B + A) = 90° - \frac{1}{2}D$, where *A, B* and *D* are the angles of the triangle; $\tan \frac{1}{2}(B - A) = \cot \frac{1}{2}D(AD - BD) / (AD + BD)$; $AB = BD \sin D / \sin A = AD \sin D / \sin B$.

Random Line On many surveys it is necessary to run a random line from point *A* to a nonvisible point *B* which is a known distance away. On the basis of compass bearings, a line such as *AB* is run. The distances *AB* and *BC* are measured, and the angle *BAC* is found from its calculated tangent (see Fig. 16.3.24).

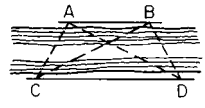


Fig. 16.3.23 Measuring an inaccessible distance.

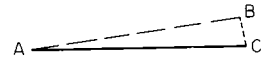


Fig. 16.3.24 Running a random line.

To Stake Out a Simple Horizontal Curve A simple horizontal curve is composed of a single arc. Usually the curve must be laid out so that it joins two straight lines called tangents, which are marked on the ground by PT (points on tangent). These tangents are run to intersection, thus locating the PI (point of intersection). The plus of the PI and the angle I are measured (see Fig. 16.3.25). With these values and any given value of R (the radius desired for the curve), the data required for staking out the curve can be computed.

$$R = \frac{5{,}729.58}{D} \qquad L = 100\,\frac{\Delta}{D} \qquad T = R \tan \frac{\Delta}{2}$$

where R = radius, T = tan distance, L = curve length, C = long chord. In a sample computation, assume that $\Delta = 8°24'$ and a 2° curve is required. $R = 5{,}729.58/2 = 2{,}864.79$ ft (873 m); $L = (100)\,(8.40/2) = 420.0$ ft (128 m); $T = 2{,}864.79 \times 0.07344 = 210.39$ ft (64 m). The degree of the curve is always twice as great as the deflection angle for a chord of 100 ft (30 m).

Setting Stakes for Trenching A common way to give line and grade for trenching (see Fig. 16.3.26) is to set stakes K ft from the center line, driving them so that the near face is the measuring point and the top is some whole inch or tenth of a foot above the bottom grade or grade of the center or top of the pipe to be laid. The top of the pipe barrel is perhaps the better line of reference. If preferred, two stakes can be driven on opposite sides and a board nailed across, on which the centerline is marked and the depth to pipeline given. When only one stake is used, a graduated pole sliding on one end of a level board at right angles is convenient for workmen and inspectors. On long grades, the grade stakes are set by ''shooting in.'' Two grade stakes are set, one at each end of the grade, the instrument is set over one, its height above grade determined, and a rod reading calculated for the distance stake such as to make the line of sight parallel to the grade line; the line of sight is then set at this rod reading; when the rod is taken to any intermediate stake, the height of instrument above grade less the rod reading will be the height of the top of the stake above grade. If the ground is uniform, the stakes may all be set at the same height above grade by driving them so as to give the same rod readings throughout.

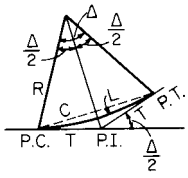


Fig. 16.3.25 Staking out a horizontal curve.

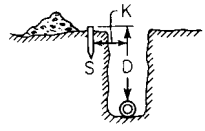


Fig. 16.3.26 Setting stakes for trenching.

To Reference a Point The Point P (Fig. 16.3.27), which must be disturbed during construction operations and will be again required as a line point in a railway, pipeline, or other survey, is referenced as follows: (1) Set the instrument over it and set four points, A, B, and C, D on two intersecting lines. When P is again required, the transit is set over B and, with foresight on A, two temporary points close together near P but on opposite sides of the line DC are set; the instrument is then set on D and, with foresight on C, a point is set in the lines DC and BA by setting it in DC under a string stretched between the two temporary points on BA. (2) Points A and E and C and F may be established instead of A, B, C, D. (3) If the ground is fairly level and is not to be much disturbed, only points A and C need be located, and these by simple tape measurement from P. They should be less than a tape length from P. When P is wanted, arcs struck from A and C with the measured distance for radii will give P at their intersection.

Foundations The corners and lines of a foundation are preserved by setting stakes outside the area to be disturbed, as in Fig. 16.3.28. Cords stretched around nails in the stakes marking the reference points will give the referenced corners at their intersections and the main lines of the building. These corners can be plumbed down to the level desired if the height of the stakes above grade is given. It is well to nail boards across the stakes at AB, putting nails in the top edge of the board to mark the points A and B, and if the ground permits, to put all the boards at the same level.

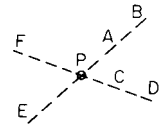


Fig. 16.3.27 Referencing a point which will be disturbed.

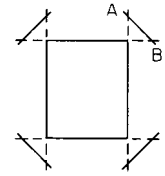


Fig. 16.3.28 Reference points for foundation corners.

To Test the Alignment and Level of a Shaft Having placed the shaft hangers as closely in line as possible by the use of a chalk line, the shaft is finally adjusted for line by hanging plumb lines over one side of the shaft at each hanger and bringing these lines into a line found by stretching a cord or wire or by setting a theodolite at one end and adjusting at each hanger till its plumb line is in the line of sight. The position of the line will be known either on the floor or on the ceiling rafters or beams to which the hangers are attached. If the latter, the instrument may be centered over a point found by plumbing down, and sighted to a plumb line at the farther end.

To level the shaft, an ordinary carpenter's level may be used near each hanger, or, better, a pole with an improvised sliding target may be hung over the shaft at each hanger by a hook in one end. The target is brought to the line of sight of a leveling instrument set preferably about under the middle of the shaft, by adjusting the hanger.

When the hangers are attached to inclined roof rafters, the two extreme hangers can be put in a line at right angles to the vertical planes of the rafters by the use of a square and cord. The other hangers will then be put as nearly as possible without instrumental test in the same line. The shaft being hung, the two extreme hangers, which have been attached to the rafters about midway between their limits of adjustment, are brought to line and level by trial, using an instrument with a well-adjusted telescope bubble, a plumb line, and inverted level rod or target pole. Each intermediate hanger is then tested and may be adjusted by trial.

To Determine the Verticality of a Stack If the stack is not in use and its top is accessible, a board can be fitted across the top, the center of the opening found, and a plumb line suspended to the bottom, where its deviation from the center will show any leaning. If the stack is in use or its top not accessible and its sides are battered, the following procedure may be followed. Referring to Fig. 16.3.29, set up an instrument at any point T and measure the horizontal angles between vertical planes tangent respectively to both sides of the top and the base and also the angle a to a second point T_1. On a line through T approximately at right angles to the chimney diameter, set the transit at T_1 and perform the same operations as at T_1, measuring also K and the angle b. On the drawing board, lay off K to as large a scale as convenient, and from the plotted T and T_1 lay off the several angles shown in the figure. By trial, draw circumferences tangent to the two quadrilaterals formed by the intersecting tangents of the base and top, respectively. The line joining the centers of these circumferences will be the deviation from the vertical in direction and amount. If the base is square, T and T_1 should be established opposite the middle points of two adjacent sides, as in Fig. 16.3.30.

To Determine Land Area and Boundaries Two procedures may be followed to gather data to establish and define land area and boundaries. Generally the data is gathered by the traverse method or the radial/radiation-survey method.

The **traverse** method entails the occupation of each end point on the boundary lines for a land area. The length of each boundary is measured (in more recent times with EDM) and the direction is measured by deflection angle or angle to the right. Reference sources explain how

direction angles and boundary lengths are converted to **latitudes** and **departures** (north-south and east-west change along a boundary). The latitudes and departures of each course/boundary allow for a simple computation of double meridian distance (DMD), which in turn is directly related by computation to land area. In Fig. 16.3.31 each of the points 1 through 4 would be occupied by the instrument for direction and distance measurement.

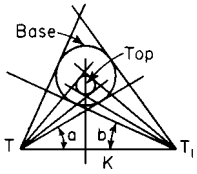


Fig. 16.3.29 Determining the verticality of a stack.

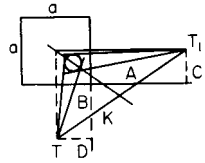


Fig. 16.3.30 Determining the verticality of a stack.

The **radial** method requires, ideally, only a single instrument setup. In Fig. 16.3.31 the setup point might be at *A* or any boundary corner with known or assumed coordinates and with line of sight available to all other corners. This method relies on gathering data to permit computation of boundary corners relative to the instrument coordinates. The data required is the usual direction angle from the instrument to the point (e.g., angle α in Fig. 16.3.31) and the distance to the point. The references detail the computation of the coordinates of each corner and the computation of land area by the **coordinate method**. The coordinate method used with the radial survey technique lends itself to the use of modern EDM and total station equipment. In very low precision work (approximation) the older methods might be used.

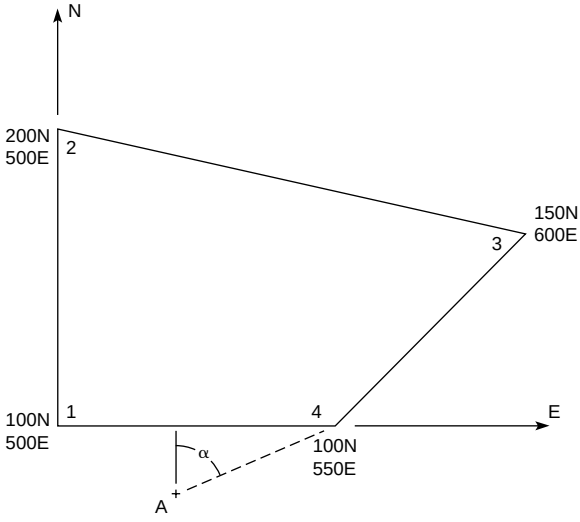


Fig. 16.3.31 Land area survey points.

Stream flow estimates may be obtained by using leveling methods to determine depth along a stream cross section and then measuring current velocity at each depth location with a current meter. As shown in Fig. 16.3.32, the total stream flow *Q* will be the sum of the products of area *A* and velocity *V* for each (say) 10-ft-wide division of the cross section.

To define limits of cut and fill once the route alignment for the roadway has been established (see Fig. 16.3.15), the surveyor uses leveling methods to define transverse ground profiles. From this information and the required elevation of the roadbed, the distance outward from the roadway center is determined, thus defining the limits of cut and fill. The surveyor then implants slope stakes to guide the earthmoving equipment operators. See Fig. 16.3.33.

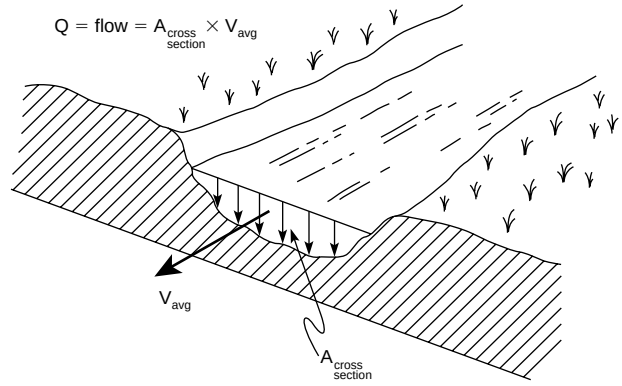


Fig. 16.3.32 Determination of stream flow.

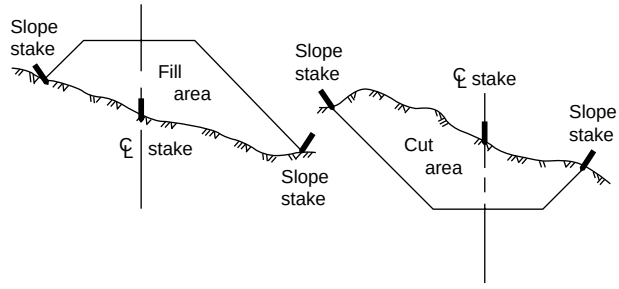


Fig. 16.3.33 Stakes showing limits of roadway cut and fill.

GLOBAL POSITIONING SYSTEM

The GPS concept is based on a system of 24 satellites in six orbital planes with four satellites per orbit. Two elements of data support the underlying principle of GPS surveying. First, radio-frequency signals are transmitted from the satellites and received by GPS receiver units at or near the earth's surface. The signal transmission time is combined with the velocity of propagation (approximately the speed of light) to result in a pseudodistance $d = Vt$, where V is the velocity of propagation and t is the time interval from source to receiver. A second requirement is the precise location of the satellites as published in satellite ephemerides.

The satellite data (ephemeride and time) is transmitted on two frequencies referred to as L1 (1575.42 MHz) and L2 (1227.60 MHz). The transmissions are band-modulated with codes that can be interpreted by the GPS receiver–signal processor. The L1 frequency modulation results in two signals: the precise positioning service (PPS) P code and the C/A or standard positioning service (SPS) code. The L2 frequency contains the P code only. The most recent GPS receivers use dual-frequency receiving and process the C/A code from L1 and the P code from L2.

Single-Receiver Positioning A single point position can be established using one GPS receiver where C/A or SPS code is processed from several satellites (at least four for both horizontal and vertical position). The reliability depends on the uncertainty of the satellite position, which is + 15 m in the ephemeride data transmitted. This handicap can be circumvented through the application of Loran C technology. The surveyor may also obtain high precision through the use of two receivers and differential positioning. The U.S. government reserves the right to degrade ephemeride data for reasons of national security.

Differential Positioning Multiple GPS receivers may be used to attain submeter accuracy. This is done by placing one receiver at a known location and multiple receivers at locations to be determined. Virtually all error inherent in the transmissions is eliminated or cancels out, since the difference in coordinates from the known position is being deter-

mined. By averaging data taken over time and postprocessing, the precision of this technique can approach millimeter precision.

Field Practice A procedure, sometimes called *leapfrogging*, can be used with three GPS receivers (R1, R2, and R3) operating simultaneously (see Fig. 16.3.34). Receiver R1 is placed at a known control point *P*, while R2 and R3 are placed at *L* and *M*. The three units are operated for a minimum of 15 min to obtain time and position data from at least four satellites. The R1 at *P* and R2 at *L* are leapfrogged to *N* and *O* respectively. This procedure is repeated until all points have been occupied. It is also desirable to include a partial leapfrog to obtain data for the *OP* line if a closed traverse is wanted.

Fig. 16.3.34 Determining positions by leapfrogging.

A second procedure requires two receivers to be initially located at two known control points. The units are operated for at least 15 min in order to establish data defining the relative location of the two points. Then one of the GPS receivers is moved to subsequent positions until all points have been occupied. As in the previous procedure, signals from a minimum of four satellites must be acquired.

GPS Computing The data for time and position of a satellite acquired by a GPS station is subsequently postprocessed on a computer. The object is to convert the time and satellite-position data to the earth-surface position of the GPS receiver. At least four satellites are needed for a complete (three-dimensional) definition of coordinate location and elevation of a control point. The computer program typically requires input of the ephemerides-coordinate location of the (at least) four satellites, the propagation velocity of the radio transmission from the satellite, and a satellite-clock offset determined by calibration or by receipt of correction data from the satellite.

General Survey Computing The microcomputer is now an integral part of the processing of survey data. Software vendors have developed many programs that perform virtually every computation required by the professional surveyor. Some programs will also produce the graphic output required for almost every survey: site drawings, plat drawings, profile and transverse sections, and topographic maps.

The Total Station The construction of the total station device is such that it performs electronically all the functions of a transit, optical theodolite, level, and tape. It has built-in EDM capability. Thus the total data-gathering requirements are available in a single instrument: horizontal and vertical angles plus the distance from the instrument's optical center to the EDM reflector. A built-in microprocessor converts the slope distance to the desired horizontal distance, and also to vertical distance for leveling requirements. If the height of the instrument's optical center and the height of the reflector are keyed in, the on-board computer can display that actual elevation difference between ground points with correction for earth curvature and atmospheric refraction. A typical total station is shown in Fig. 16.3.11.

Another feature found in some total stations is data-collection capability. This permits data as read by the instrument (angle, distance) to be electronically output to a data collector for transfer to a computer for later processing. A total station operating in conjunction with a remote-processing unit allows a survey to be conducted by a single person. The variety of features found on total station devices is enormous.

The total station changes some of the traditional procedures in surveying practice. An example is the laying out of a route curve as shown in Fig. 16.3.24. The references describe the newer methods.