Section 8

Process Control

Thomas F. Edgar, Ph.D., *Professor of Chemical Engineering, University of Texas, Austin, TX. (Advanced Control Systems, Process Measurements, Section Editor)*

Cecil L. Smith, Ph.D., *Principal, Cecil L. Smith Inc., Baton Rouge, LA. (Batch Process Control, Telemetering and Transmission, Digital Technology for Process Control, Process Control and Plant Safety)*

F. Greg Shinskey, B.S.Ch.E., *Consultant (retired from Foxboro Co.), North Sandwich, NH. (Fundamentals of Process Dynamics and Control, Unit Operations Control)*

George W. Gassman, B.S.M.E., *Senior Research Specialist, Final Control Systems, Fisher Controls International, Inc., Marshalltown, IA. (Controllers, Final Control Elements, and Regulators)*

Paul J. Schafbuch, Ph.D., *Senior Research Specialist, Final Control Systems, Fisher Controls International, Inc., Marshalltown, IA. (Controllers, Final Control Elements, and Regulators)*

Thomas J. McAvoy, Ph.D., *Professor of Chemical Engineering, University of Maryland, College Park, MD. (Fundamentals of Process Dynamics and Control)*

Dale E. Seborg, Ph.D., *Professor of Chemical Engineering, University of California, Santa Barbara, CA. (Advanced Control Systems)*

FUNDAMENTALS OF PROCESS DYNAMICS AND CONTROL

ADVANCED CONTROL SYSTEMS

UNIT OPERATIONS CONTROL

BATCH PROCESS CONTROL

PROCESS MEASUREMENTS

TELEMETERING AND TRANSMISSION

DIGITAL TECHNOLOGY FOR PROCESS CONTROL

CONTROLLERS, FINAL CONTROL ELEMENTS, AND REGULATORS

PROCESS CONTROL AND PLANT SAFETY

Nomenclature

Symbol	Definition
A	Area
A_a	Actuator area
A_c	Amplitude of controlled variable
A_m	Output amplitude limits
A_v	Cross sectional area of valve
A_1	Cross sectional area of tank
b	Controller output bias
B	Bottoms flow rate
B_i^*	Limit on control
c,C	Controlled variable
c_A	Concentration of A
C_d	Discharge coefficient
C_i	Inlet concentration
C_i^*	Limit on control move
C_L	Specific heat of liquid
C_o	Integration constant
C_r	Heat capacity of reactants
C_v	Valve flow coefficient
D	Distillate flow rate
D_i^*	Limit on output
$D(s)$	Decoupler transfer function
e	Error
E	Economy of evaporator
f	Function of time
F,f	Feed flow rate
F_L	Pressure recovery factor
g_c	Unit conversion constant
g_i	Algebraic inequality constraint
G	Transfer function
G_c	Controller transfer function
G_f	Feedforward controller transfer function
G_L	Load transfer function
G_m	Sensor transfer function
G_p	Process transfer function
G_t	Transmitter transfer function
G_v	Valve transfer function
h_i	Algebraic equality constraints
h_1	Liquid head in tank
H	Latent heat of vaporization
i	Summation index
I_i	Impulse response coefficient
j	Time index
J	Objective function or performance index
k	Time index
k_f	Flow coefficient
k_r	Kinetic rate constant
K	Gain
K_c	Controller gain
K_L	Load transfer function gain
K_m	Measurement gain
K_p	Process gain
K_u	Ultimate controller gain (stability)
L	Disturbance or load variable
L_p	Sound pressure level
m,M	Manipulated variable
m_c	Number of constraints
M_o	Mass flow
M_r	Mass of reactants
M_w	Molecular weight
n	Number of data points, number of stages or effects
N	Number of inputs/outputs, model horizon
p_1	Pressure
p_a	Actuator pressure
p_v	Vapor pressure
P	Proportional band (%)
P_u	Proportional band (ultimate)
q	Radiated energy flux
q_b	Energy flux to a black body
Q_a	Flow rate
r_c	Number of constraints
R	Equal percentage valve characteristic

Symbol	Definition
R,r	Set point
R_T	Resistance in temperature sensor
R_1	Valve resistance
s	Laplace transform variable
$\mathbf{s}$	Search direction
S_i	Step response coefficient
t	Time
T	Temperature
T_b	Base temperature
T_f	Exhaust temperature
T_R	Reset time
u	Controller output
U	Heat transfer coefficient
V	Volume
V_s	Product value
w	Mass flow rate
w_i	Weighting factor
W	Steam flow rate
x	Mass fraction
x_i	Optimization variable
x_T	Pressure drop ratio factor
X	Transform of deviation variable
y	Process output, controlled variable, valve travel
Y	Controller tuning law, expansion factor
z	z-transform variable
z_i	Feed mole fraction (distillation)
Z	Compressibility factor

Greek symbols

Symbol	Definition
α	Digital filter coefficient
α_T	Temperature coefficient of resistance
β	Resistance thermometer parameter
γ	Ratio of specific heats
δ	Move suppression factor
Δq	Load step change
Δt	Time step
ΔT	Temperature change
Δu	Control move
ε	Spectral emissivity, step size
ζ	Damping factor (second order system)
θ	Time delay
λ	Relative gain array parameter, wavelength
Λ	Relative gain array
ξ	Deviation variable
ρ	Density
σ	Stefan-Boltzmann constant
Σ_τ	Total response time
τ	Time constant
τ_D	Derivative time (PID controller)
τ_F	Filter time constant
τ_I	Integral time (PID controller)
τ_L	Load time constant
τ_n	Natural period of closed loop
τ_p	Process time constant
τ_o	Period of oscillation
ϕ_{PI}	Phase lag

Subscripts

Symbol	Definition
A	Species A
b	Best
c	Controller
eff	Effective
F	Feedforward
i	Initial, inlet
L	Load, disturbance
m	Measurement or sensor
p	Process
s	Steady state
set	Set point value
t	Transmitter
u	Ultimate
v	Valve

FUNDAMENTALS OF PROCESS DYNAMICS AND CONTROL

THE GENERAL CONTROL SYSTEM

A process is shown in Fig. 8-1 with a manipulated input M, a load input L, and a controlled output C, which could be flow, pressure, liquid level, temperature, composition, or any other inventory, environmental, or quality variable that is to be held at a desired value identified as the set point R. The load may be a single variable or aggregate of variables acting either independently or manipulated for other purposes, affecting the controlled variable much as the manipulated variable does. Changes in load may occur randomly as caused by changes in weather, diurnally with ambient temperature, manually when operators change production rate, stepwise when equipment is switched in or out of service, or cyclically as the result of oscillations in other control loops. Variations in load will drive the controlled variable away from set point, requiring a corresponding change in the manipulated variable to bring it back. The manipulated variable must also change to move the controlled variable from one set point to another.

An open-loop system positions the manipulated variable either manually or on a programmed basis, without using any process measurements. This operation is acceptable for well-defined processes without disturbances. An automanual transfer switch is provided to allow manual adjustment of the manipulated variable in case the process or the control system is not performing satisfactorily.

A closed-loop system uses the measurement of one or more process variables to move the manipulated variable to achieve control. Closed-loop systems may include feedforward, feedback, or both.

Feedback Control In a feedback control loop, the controlled variable is compared to the set point R, with the difference, deviation, or error e acted upon by the controller to move m in such a way as to minimize the error. This action is specifically negative feedback, in that an increase in deviation moves m so as to decrease the deviation. (Positive feedback would cause the deviation to expand rather than diminish and therefore does not regulate.) The action of the controller is selectable to allow use on process gains of both signs.

The controller has tuning parameters related to proportional, integral, derivative, lag, deadtime, and sampling functions. A negative feedback loop will oscillate if the controller gain is too high, but if it is too low, control will be ineffective. The controller parameters must be properly related to the process parameters to ensure closed-loop stability while still providing effective control. This is accomplished first by the proper selection of control modes to satisfy the requirements of the process, and second by the appropriate tuning of those modes.

Feedforward Control A feedforward system uses measurements of disturbance variables to position the manipulated variable in such a way as to minimize any resulting deviation. The disturbance variables could be either measured loads or the set point, the former being more common. The feedforward gain must be set precisely to offset the deviation of the controlled variable from the set point.

Feedforward control is usually combined with feedback control to eliminate any offset resulting from inaccurate measurements and calculations and unmeasured load components. The feedback controller can either bias or multiply the feedforward calculation.

Computer Control Computers have been used to replace analog PID controllers, either by setting set points of lower level controllers in supervisory control, or by driving valves directly in direct digital control. Single-station digital controllers perform PID control in one or two loops, including computing functions such as mathematical operations, characterization, lags, and deadtime, with digital logic and alarms. Distributed control systems provide all these functions, with the digital processor shared among many control loops; separate processors may be used for displays, communications, file servers, and the like. A host computer may be added to perform high-level operations such as scheduling, optimization, and multivariable control. More details on computer control are provided later in this section.

PROCESS DYNAMICS AND MATHEMATICAL MODELS

GENERAL REFERENCES: Seborg, Edgar, and Mellichamp, *Process Dynamics and Control*, Wiley, New York, 1989; Marlin, *Process Control*, McGraw-Hill, New York, 1995; Ogunnaike and Ray, *Process Dynamics Modeling and Control*, Oxford University Press, New York, 1994; Smith and Corripio, *Principles and Practices of Automatic Process Control*, Wiley, New York, 1985

Open-Loop versus Closed-Loop Dynamics It is common in industry to manipulate coolant in a jacketed reactor in order to control conditions in the reactor itself. A simplified schematic diagram of such a reactor control system is shown in Fig. 8-2. Assume that the reactor temperature is adjusted by a controller that increases the coolant flow in proportion to the difference between the desired reactor temperature and the temperature that is measured. The proportionality constant is K_c. If a small change in the temperature of the inlet stream occurs, then depending on the value of K_c, one might observe the reactor temperature responses shown in Fig. 8-3. The top plot shows the case for no control ($K_c = 0$), which is called the open loop, or the normal dynamic response of the process by itself. As K_c increases, several effects can be noted. First, the reactor temperature responds faster and faster. Second, for the initial increases in K_c, the maximum deviation in the reactor temperature becomes smaller. Both of these effects are desirable so that disturbances from normal operation have

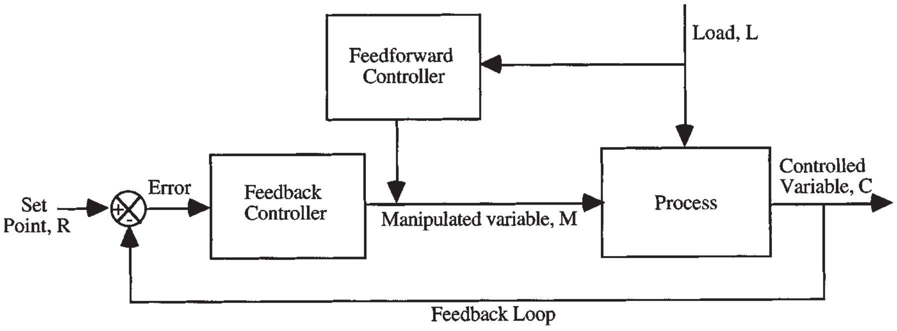


FIG. 8-1 Block diagram for feedforward and feedback control.

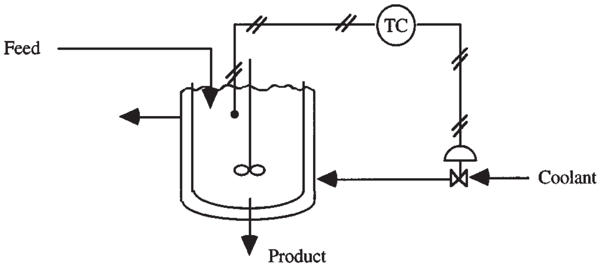


FIG. 8-2 Reactor control system.

as small an effect as possible on the process under study. As the gain is increased further, eventually a point is reached where the reactor temperature oscillates indefinitely, which is undesirable. This point is called the stability limit, where $K_c = K_u$, the ultimate controller gain. Increasing K_c further causes the magnitude of the oscillations to increase, with the result that the control valve will cycle between full open and closed.

The responses shown in Fig. 8-3 are typical of the vast majority of regulatory loops encountered in the process industries. Figure 8-3 shows that there is an optimal choice for K_c, somewhere between 0 (no control) and K_u (stability limit). If one has a dynamic model of a process, then this model can be used to calculate controller settings. In Fig. 8-3, no time scale is given, but rather the figure shows relative responses. A well-designed controller might be able to speed up the response of a process by a factor of roughly two to four. Exactly how fast the control system responds is determined by the dynamics of the process itself.

Physical Models versus Empirical Models In developing a dynamic process model, there are two distinct approaches that can be taken. The first involves models based on first principles, called physical models, and the second involves empirical models. The conservation laws of mass, energy, and momentum form the basis for developing physical models. The resulting models typically involve sets of differential and algebraic equations that must be solved simultaneously. Empirical models, by contrast, involve postulating the form of a dynamic model, usually as a transfer function, which is discussed below. This transfer function contains a number of parameters that need to be estimated. For the development of both physical and empirical models, the most expensive step normally involves verification of their accuracy in predicting plant behavior.

To illustrate the development of a physical model, a simplified treatment of the reactor, shown in Fig. 8-2 is used. It is assumed that the reactor is operating isothermally and that the inlet and exit volumetric flows and densities are the same. There are two components, *A* and *B*, in the reactor, and a single first order reaction of $A \rightarrow B$ takes place. The inlet concentration of *A*, which we shall call c_i, varies with time. A dynamic mass balance for the concentration of *A* (c_A) can be written as follows:

$$V\frac{dc_A}{dt} = Fc_i - Fc_A - k_rV_c \qquad (8\text{-}1)$$

In Eq. (8-1), the flow in of *A* is Fc_i, the flow out is Fc_A, and the loss via reaction is k_rVc_A, where V = reactor volume and k_r = kinetic rate constant. In this example, c_i is the input, or forcing variable, and c_A is the output variable. If *V, F,* and k_r are constant, Eq. (8-1) can be rearranged by dividing by $(F + k_rV)$ so that it only contains two groups of parameters. The result is:

$$\tau\frac{dc}{dt} = Kc_i - c_A \qquad (8\text{-}2)$$

where $\tau = V/(F + k_rV)$ and $K = F/(F + k_rV)$. For this example, the resulting model is a first-order differential equation in which τ is called the **time constant** and *K* the **process gain.**

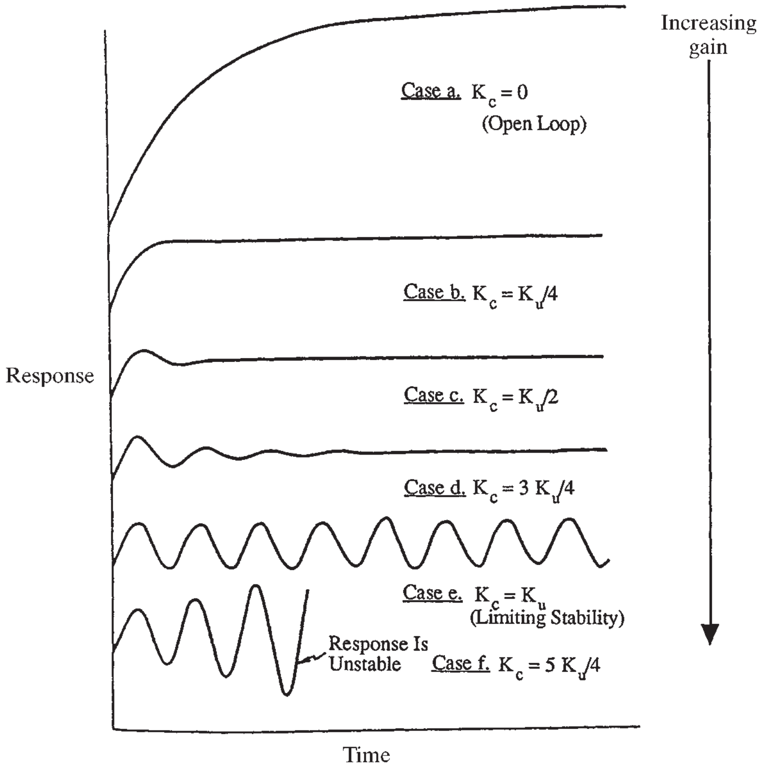


FIG. 8-3 Typical control system responses.

As an alternative to deriving Eq. (8-2) from a dynamic mass balance, one could simply postulate a first-order differential equation to be valid (empirical modeling). Then it would be necessary to estimate values for τ and *K* so that the postulated model described the reactor's dynamic response. The advantage of the physical model over the empirical model is that the physical model gives insight into how reactor parameters affect the values of τ, and *K,* which in turn affects the dynamic response of the reactor.

Nonlinear versus Linear Models If *V, F,* and *k* are constant, then Eq. (8-1) is an example of a linear differential equation model. In a linear equation, the output and input variables and their derivatives only appear to the first power. If the rate of reaction were second order, then the resulting dynamic mass balance would be:

$$V\frac{dc}{dt} = Fc_i - Fc_A - k_r\,Vc_A^2 \tag{8-3}$$

Since c_A appears in this equation to the second power, the equation is nonlinear.

The difference between linear systems and nonlinear systems can be seen by considering the steady state behavior of Eq. (8-1) compared to Eq. (8-3) (the left-hand side is zero; i.e., $dc_A/dt = 0$). For a given change in c_i, Δc_i, the change in c_A calculated from Eq. (8-1), or Δc, is always proportional to Δc_i, and the proportionality constant is K [see Eq. (8-2)]. The change in the output of a system divided by a change in the input to the system is called the *process gain.* Linear systems have constant process gains for all changes in the input. By contrast, Eq. (8-3) gives a Δc that varies in proportion to Δc_i but with the proportionality factor being a function of the concentration levels in the reactor. Thus, depending on where the reactor operates, a change in c_i produces different changes in c_A. In this case, the process has a nonlinear gain. Systems with nonlinear gains are more difficult to control than linear systems that have constant gains.

Simulation of Dynamic Models Linear dynamic models are particularly useful for analyzing control-system behavior. The insight gained through linear analysis is invaluable. However, accurate dynamic process models can involve large sets of nonlinear equations. Analytical solution of these models is not possible. Thus, in these cases, one must turn to simulation approaches to study process dynamics and the effect of process control. Equation (8-3) will be used to illustrate the simulation of nonlinear processes. If dc_A/dt on the left-hand side of Eq. (8-3) is replaced with its finite difference approximation, one gets:

$$c_A(t+\Delta t) = \frac{c_A(t) + \Delta t\cdot[Fc_i(t) - Fc_A(t) - k_rVc_A(t)^2]}{V} \tag{8-4}$$

Starting with an initial value of c_A and knowing $c_i(t)$, Eq. (8-4) can be solved for $c_A(t+\Delta t)$. Once $c_A(t+\Delta t)$ is known, the solution process can be repeated to calculate $c_A(t+2\Delta t)$, and so on. This approach is called the Euler integration method; while it is simple, it is not necessarily the best approach to numerically integrating nonlinear differential equations. To achieve accurate solutions with an Euler approach, one often needs to take small steps in time, Δ*t.* A number of more sophisticated approaches are available that allow much larger step sizes to be taken but require additional calculations. One widely used approach is the fourth-order Runge Kutta method, which involves the following calculations:

define
$$f(c_A,t) = \frac{Fc_i(t) - Fc_A - k_rVc_A^2}{V} \tag{8-5}$$

then
$$c_A(t+\Delta t) = c_A(t) + \Delta t(m_1 + 2m_2 + 2m_3 + m_4) \tag{8-6}$$

with
$$m_1 = f\,[c_A(t),\, t] \tag{8-7}$$

$$m_2 = f\left[c_A(t) + \frac{m_1\Delta t}{2},\, t + \frac{\Delta t}{2}\right] \tag{8-8}$$

$$m_3 = f\left[c_A(t) + \frac{m_2\Delta t}{2},\, t + \frac{\Delta t}{2}\right] \tag{8-9}$$

$$m_4 = f\,[c_A(t) + m_3\Delta t,\, t + \Delta t] \tag{8-10}$$

In this method, the m_i's are calculated sequentially in order to take a step in time. Even though this method requires calculation of the four additional m_i values, for equivalent accuracy the fourth-order Runge Kutta method can result in a faster numerical solution, since a larger step, Δ*t,* can be taken with it. Increasingly sophisticated simulation packages are being used to calculate the dynamic behavior of processes and test control system behavior. These packages have good user interfaces, and they can handle stiff systems where some variables respond on a time scale that is much much faster or slower than other variables. A simple Euler approach cannot effectively handle stiff systems, which frequently occur in chemical-process models.

Laplace Transforms When mathematical models are used to describe process dynamics in conjunction with control-system analysis, the models generally involve linear differential equations. Laplace transforms are very effective for solving linear differential equations. The key advantage of using Laplace transforms is that they convert differential equations into algebraic equations. The resulting algebraic equations are easier to solve than the original differential equations. When the Laplace transform is applied to a linear differential equation in time, the result is an algebraic equation in a new variable, *s,* called the Laplace variable. To get the solution to the original differential equation, one needs to invert the Laplace transform. Table 8-1 gives a number of useful Laplace transform pairs, and more extensive tables are available (Seborg, Edgar, and Mellichamp, *Process Dynamics and Control,* Wiley, New York, 1989).

To illustrate how Laplace transforms work, consider the problem of solving Eq. (8-2), subject to the initial condition that $c_A = 0$ at $t = 0$, and c_i is constant. If c_A were not initially zero, one would define a deviation variable between c_A and its initial value $(c_A - c_0)$. Then the transfer function would be developed using this deviation variable. Taking the Laplace transform of both sides of Eq. (8-2) gives:

$$\pounds\left(\frac{\tau dc_A}{dt}\right) = \pounds(Kc_i) - \pounds(c_A) \tag{8-11}$$

Denoting the £(*c*) as $C_A(s)$ and using the relationships in Table 8-1 gives:

$$\tau sC_A(s) = \frac{Kc_i}{s} - C_A(s) \tag{8-12}$$

Equation (8-12) can be solved for C_A to give:

$$C_A(s) = \frac{Kc_i/s}{\tau s + 1} \tag{8-13}$$

Using the entries in Table 8-1, Eq. (8-13) can be inverted to give the transient response of c_A as:

$$c_A(t) = (Kc_i)(1 - e^{-t/\tau}) \tag{8-14}$$

Equation (8-14) shows that c_A starts from 0 and builds up exponentially to a final concentration of Kc_i. Note that to get Eq. (8-14), it was only necessary to solve the algebraic Eq. (8-12) and then find the inverse of $C_A(s)$ in Table 8-1. The original differential equation was not solved directly. In general, techniques such as partial fraction expansion must be used to solve higher order differential equations with Laplace transforms.

Transfer Functions and Block Diagrams A very convenient and compact method of representing the process dynamics of linear systems involves the use of transfer functions and block diagrams. A transfer function can be obtained by starting with a physical model as

TABLE 8-1 Frequently Used Laplace Transforms

Time function, $f(t)$	Transform, $F(s)$
A	A/s^2
At	A/s
Ae^{-at}	$A/(s+a)$
$A(1-e^{-t/\tau})$	$A/[s(\tau s+1)]$
$A\sin(\omega t)$	$A\omega/(s^2+\omega^2)$
$f(t-\theta)$	$e^{-\theta s}F(s)$
df/dt	$sF(s) - f(0)$
$\int f(t)\,dt$	$F(s)/s$

discussed previously. If the physical model is nonlinear, then it first needs to be linearized around an operating point. The resulting linearized model is then approximately valid in a region around this operating point. To illustrate how transfer functions are developed, Eq. (8-2) will again be used. First, one defines deviation variables, which are the process variables minus their steady state values at the operating point. For Eq. (8-2), there would be deviation variables for both c_A and c_i, and these are defined as:

$$\xi = c_A - c_s \tag{8-15}$$

$$\xi_i = c_i - c_{is} \tag{8-16}$$

where the subscript s stands for steady state. Substitution of Eq. (8-15) and (8-16) into Eq. (8-2) gives:

$$\tau \frac{d\xi}{dt} = K\xi_i - \xi + (Kc_{is} - c_s) \tag{8-17}$$

The term in parentheses in Eq. (8-17) is zero at steady state and thus it can be dropped. Next the Laplace transform is taken, and the resulting algebraic equation solved. Denoting $X(s)$ as the Laplace transform of ξ and $X_i(s)$ as the transform of ξ_i, the final transfer function can be written as:

$$\frac{X}{X_i} = \frac{K}{\tau s + 1} \tag{8-18}$$

Equation (8-18) is an example of a first-order transfer function. As mentioned above, an alternative to formally deriving Eq. (8-18) involves simply postulating its form and then identifying its two parameters, the **process gain** K and **time constant** τ, to fit the process under study. In fitting the parameters, data can be generated by forcing the process. If step forcing is used, then the resulting response is called the **process reaction curve.** Often transfer functions are placed in block diagrams, as shown in Fig. 8-4. Block diagrams show how changes in an input variable affect an output variable. Block diagrams are a means of concisely representing the dynamics of a process under study. Since linearity is assumed in developing a block diagram, if more than one variable affects an output, the contributions from each can be added together.

Continuous versus Discrete Models The preceding discussion has focused on systems where variables change continuously with time. Most real processes have variables that are continuous in nature, such as temperature, pressure, and flow. However, some processes involve discrete events, such as the starting or stopping of a pump. In addition, modern plants are controlled by digital computers, which are discrete by nature. In controlling a process, a digital system samples variables at a fixed rate, and the resulting system is a sampled data system. From one sampling instant until the next, variables are assumed to remain fixed at their sampled values. Similarly, in controlling a process, a digital computer sends out signals to control elements, usually valves, at discrete instants of time. These signals remain fixed until the next sampling instant.

Figure 8-5 illustrates the concept of sampling a continuous function. At integer values of the sampling rate, Δt, the value of the variable to be sampled is measured and held until the next sampling instant. To deal with sampled data systems, the z transform has been developed. The z transform of the function given in Fig. 8-5 is defined as

$$Z(f) = \sum_{n=0}^{\infty} f(n\,\Delta t) z^{-n} \tag{8-19}$$

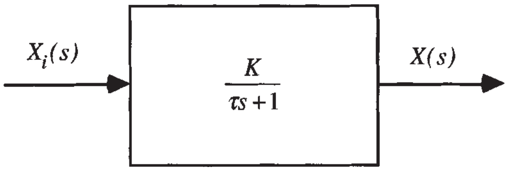


FIG. 8-4 First-order transfer function.

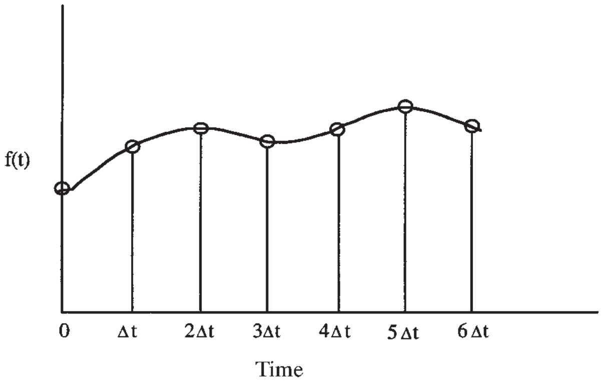


FIG. 8-5 Sampled data example.

In an analogous manner to Laplace transforms, one can develop transfer functions in the z domain as well as block diagrams. Tables of z transform pairs have been published (Seborg, Edgar, and Mellichamp, *Process Dynamics and Control,* Wiley, New York, 1989) so that the discrete transfer functions can be inverted back to the time domain. The inverse gives the value of the function at the discrete sampling instants. Sampling a continuous variable results in a loss of information. However, in practical applications, sampling is fast enough that the loss is typically insignificant and the difference between continuous and discrete modeling is small in terms of its effect on control. Increasingly, model predictive controllers that make use of discrete dynamic models are being used in the process industries. The purpose of these controllers is to guide a process to optimum operating points. These model predictive control algorithms are typically run at much slower sampling rates than are used for basic control loops such as flow control or pressure control. The discrete dynamic models used are normally developed from data generated from plant testing as discussed hereafter. For a detailed discussion of modeling sampled data systems, the interested reader is referred to textbooks on digital control (Astrom and Wittenmark, *Computer Controlled Systems,* Prentice Hall, Englewood Cliffs, NJ, 1984).

Process Characteristics in Transfer Functions In many cases, process characteristics are expressed in the form of transfer functions. In the previous discussion, a reactor example was used to illustrate how a transfer function could be derived. Here, another system involving flow out of a tank, shown in Fig. 8-6, is considered.

Proportional Element First, consider the outflow through the exit valve on the tank. If the flow through the line is turbulent, then Bernoulli's equation can be used to relate the flow rate through the valve to the pressure drop across the valve as:

$$f_1 = k_f A_v \sqrt{2g_c(h_1 - h_0)} \tag{8-20}$$

where f_1 = flow rate, k_f = flow coefficient, A_v = cross sectional area of the restriction, g_c = constant, h_1 = liquid head in tank, and h_0 = atmo-

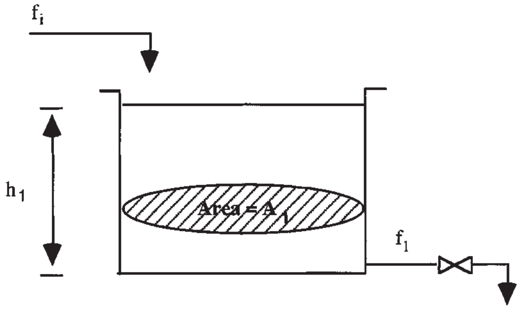


FIG. 8-6 Single tank with exit valve.

spheric pressure. This relationship between flow and head is nonlinear, and it can be linearized around a particular operating point to give:

$$f_1 - f_{1s} = \left(\frac{1}{R_1}\right)(h_1 - h_{1s}) \qquad (8\text{-}21)$$

where $R_1 = f_{1s}/(g_c k_f^2 A^2)$ is called the resistance of the valve in analogy with an electrical resistance. The transfer function relating changes in flow to changes in head is shown in Fig. 8-7, and it is an example of a pure gain system with no dynamics. In this case, the process gain is $K = 1/R_1$. Such a system has an instantaneous dynamic response, and for a step change in head, there is an immediate step change in flow, as shown in Fig. 8-8. The exact magnitude of the step in flow depends on the operating flow, f_{1s}, as the definition of R_1 shows.

First-Order Lag (Time Constant Element) Next consider the system to be the tank itself. A dynamic mass balance on the tank gives:

$$A_1 \frac{dh_1}{dt} = f_i - f_1 \qquad (8\text{-}22)$$

where A_1 is the cross sectional area of the tank and f_i is the inlet flow. By substituting Eq. (8-21) into Eq. (8-22), and following the approach discussed above for deriving transfer functions, one can develop the transfer function relating changes in h_1 to changes in f_i. The resulting transfer function is another example of a first-order system, shown in Fig. 8-4, and it has a gain, $K = R_1$, and a time constant, $\tau_1 = R_1A_1$. For a step change in f_i, h_1 follows a decaying exponential response from its initial value, h_{1s}, to a final value of $h_{1s} + R_1\Delta f_i$ (Fig. 8-9). At a time equal to τ_1, the transient in h_1 is 63 percent finished; and at $3\tau_1$, the response is 95 percent finished. These percentages are the same for all first-order processes. Thus, knowledge of the time constant of a first-order process gives insight into how fast the process responds to sudden input changes.

Capacity Element Now consider the case where the valve in Fig. 8-7 is replaced with a pump. In this case, it is reasonable to assume that the exit flow from the tank is independent of the level in the tank. For such a case, Eq. (8-22) still holds, except that f_1 no longer depends on h_1. For changes in f_i, the transfer function relating changes in h_1 to changes in f_i is shown in Fig. 8-10. This is an example of a pure capacity process, also called an integrating system. The cross sectional area of the tank is the chemical process equivalent of an electrical capacitor. If the inlet flow is step forced while the outlet is held constant, then the level builds up linearly as shown in Fig. 8-11. Eventually the liquid would overflow the tank.

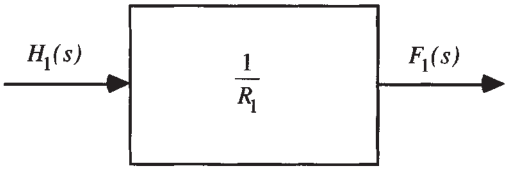


FIG. 8-7 Proportional element transfer function.

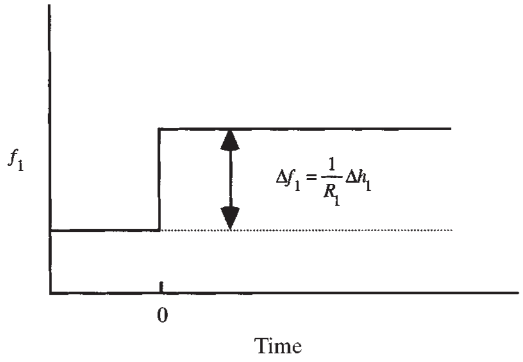


FIG. 8-8 Response of proportional element.

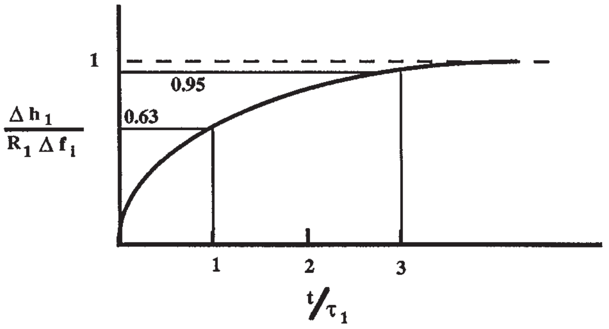


FIG. 8-9 Response of first-order system.

Second-Order Element Because of their linear nature, transfer functions can be combined in a straightforward manner. Consider the two tank system shown in Fig. 8-12. For tank 1, the transfer function relating changes in f_1 to changes in f_i can be obtained by combining two first order transfer functions to give:

$$\frac{F_1(s)}{F_i(s)} = \frac{1}{R_1A_1s + 1} \qquad (8\text{-}23)$$

Since f_1 is the inlet flow to tank 2, the transfer function relating changes in h_2 to changes in f_1 has the same form as that given in Fig. 8-4:

$$\frac{H_2(s)}{F_1(s)} = \frac{R_2}{A_2R_2s + 1} \qquad (8\text{-}24)$$

Equations (8-23) and (8-24) can be multiplied together to give the final transfer function relating changes in h_2 to changes in f_i as shown in Fig. 8-13. This is an example of a second-order transfer function. This transfer function has a gain R_1R_2 and two time constants, R_1A_1 and R_2A_2. For two equal tanks, a step change in f_i produces the S-shaped response in level in the second tank shown in Fig. 8-14.

General Second-Order Element Figure 8-3 illustrates the fact that closed loop systems often exhibit oscillatory behavior. A general

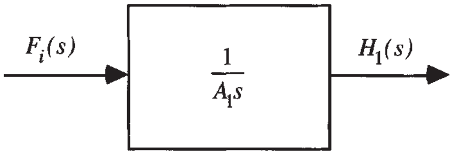


FIG. 8-10 Pure capacity transfer function.

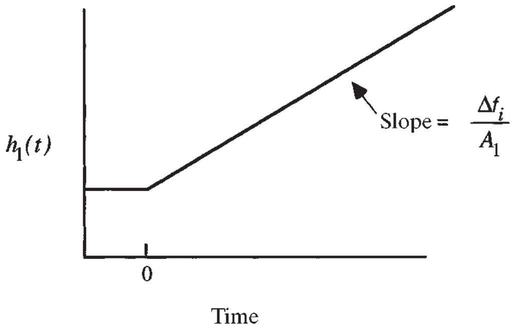


FIG. 8-11 Response of pure capacity system.

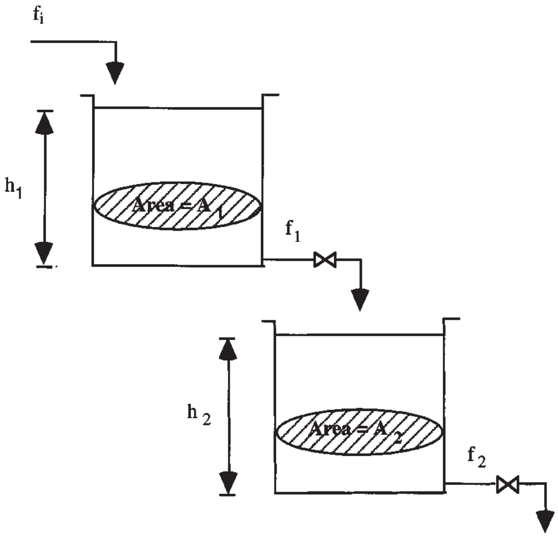


FIG. 8-12 Two tanks in series.

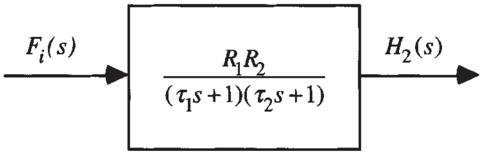


FIG. 8-13 Second-order transfer function.

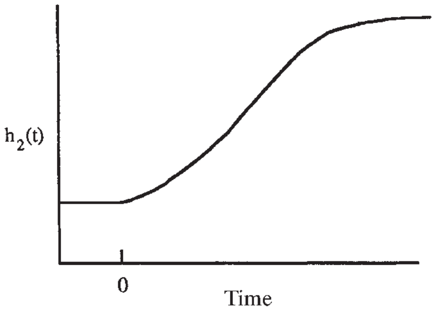


FIG. 8-14 Response of second-order system.

second-order transfer function that can exhibit oscillatory behavior is important for the study of automatic control systems. Such a transfer function is given in Fig. 8-15. For a step input, the transient responses shown in Fig. 8-16 result. As can be seen when $\zeta < 1$, the response oscillates and when $\zeta > 1$, the response is S-shaped. Few open-loop chemical processes exhibit an oscillating response; most exhibit an S-shaped step response.

Distance-Velocity Lag (Dead-Time Element) The dead-time element, commonly called a distance-velocity lag, is often encountered in process systems. For example, if a temperature-measuring element is located downstream from a heat exchanger, a time delay occurs before the heated fluid leaving the exchanger arrives at the temperature measurement point. If some element of a system produces a dead-time of θ time units, then an input to that unit, $f(t)$, will be reproduced at the output as $f(t - \theta)$. The transfer function for a pure dead-time element is shown in Fig. 8-17, and the transient response of the element is shown in Fig. 8-18.

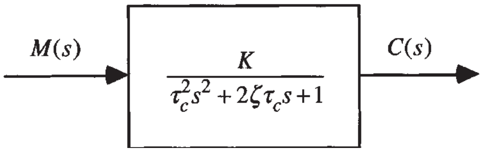


FIG. 8-15 General second-order transfer function.

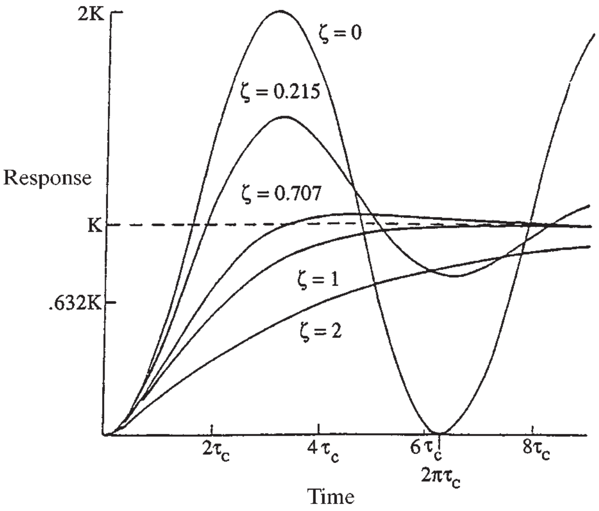


FIG. 8-16 Response of general second-order system.

Higher-Order Lags If a process is described by a series of n first-order lags, the overall system response becomes proportionally slower with each lag added. The special case of a series of n first-order lags with equal time constants has a transfer function given by:

$$G(s) = \frac{K}{(\tau s + 1)^n} \qquad (8\text{-}25)$$

The step response of this transfer function is shown in Fig. 8-19. Note that all curves reach about 60 percent of their final value at $t = n\tau$.

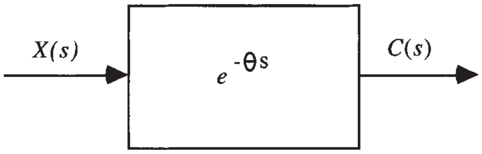


FIG. 8-17 Dead-time transfer function.

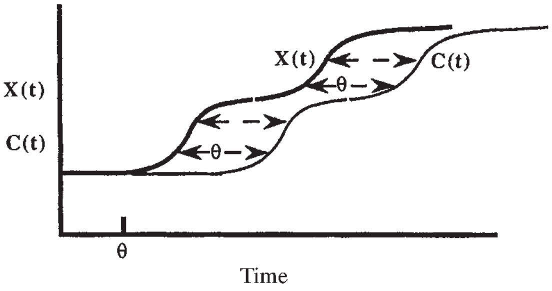


FIG. 8-18 Response of dead-time system.

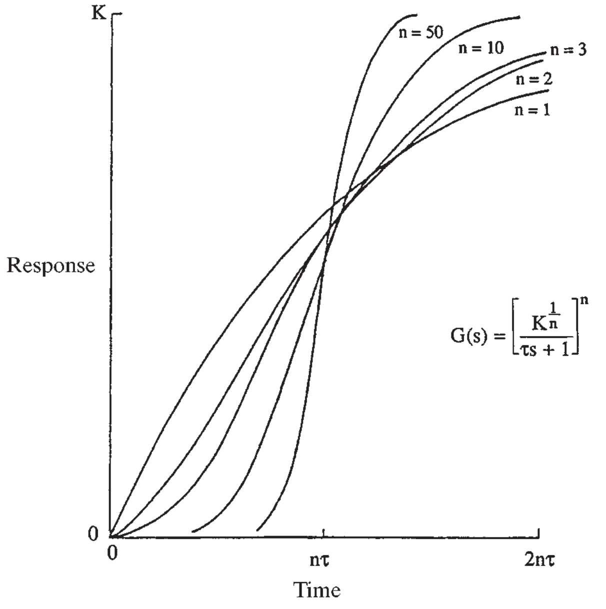


FIG. 8-19 Response of *n*th order lags.

Higher-order systems can be approximated by a first or second-order plus dead-time system for control system design.

Multiinput, Multioutput Systems The dynamic systems considered up to this point have been examples of single-input, single-output (SISO) systems. In chemical processes, one often encounters systems where one input can affect more than one output. For example, assume that one is studying a distillation tower in which both reflux and boilup are manipulated for control purposes. If the output variables are the top and bottom product compositions, then each input affects both outputs. For this distillation example, the process is referred to as a 2 × 2 system to indicate the number of inputs and outputs. In general, multiinput, multioutput (MIMO) systems can have n inputs and m outputs with $n \neq m$, and they can be nonlinear. Such a system would be called an $n \times m$ system. An example of a transfer function for a 2 × 2 linear system is given in Fig. 8-20. Note that since linear systems are involved, the effects of the two inputs on each output are additive. In many process-control systems, one input is selected to control one output in a MIMO system. For m output there would be m such selections. For this type of control strategy, one needs to consider which inputs and outputs to couple together, and this problem is referred to as loop pairing. Another important issue that arises involves interaction between control loops. When one loop makes a change in its manipulated variable, the change affects the other loops in the system. These changes are the direct result of the multivariable nature of the process. In some cases, the interaction can be so severe that overall control-system performance is drastically reduced. Finally, some of the modern approaches to process control tackle the MIMO problem directly, and they simultaneously use all manipulated variables to control all output variables rather than pairing one input to one output (see later section on multivariable control).

Fitting Dynamic Models to Experimental Data In developing empirical transfer functions, it is necessary to identify model parameters from experimental data. There are a number of approaches to process identification that have been published. The simplest approach involves introducing a step test into the process and recording the response of the process, as illustrated in Fig. 8-21. The *x*'s in the figure represent the recorded data. For purposes of illustration, the process under study will be assumed to be first order with dead-time and have the transfer function:

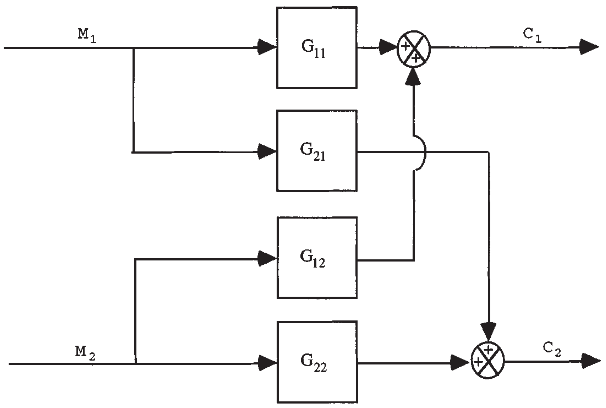


FIG. 8-20 Example of 2 × 2 transfer function.

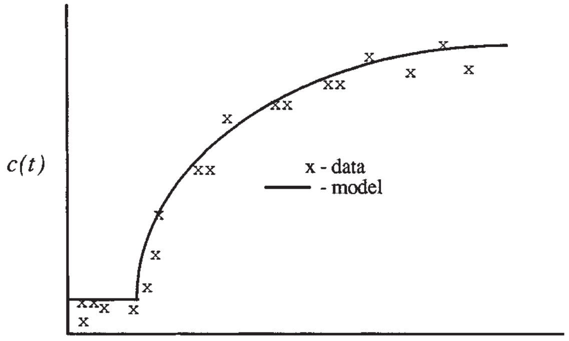


FIG. 8-21 Plot of experimental data.

$$G(s) = \frac{C(s)}{M(s)} = K \exp(-\theta s)/(\tau s + 1) \qquad (8\text{-}26)$$

The response produced by Eq. (8-26), $c(t)$, can be found by inverting the transfer function, and it is also shown in Fig. 8-21 for a set of model parameters, K, τ, and θ, fitted to the data. These parameters are calculated using optimization to minimize the squared difference between the model predictions and the data, i.e., a least squares approach. Let each measured data point be represented by c_j (measured response), t_j (time of measured response), $j = 1$ to n. Then the least squares problem can be formulated as:

$$\min_{\tau,\theta,K} \sum_{j=0}^{n} [c_j - \hat{c}(t_j)]^2 \qquad (8\text{-}27)$$

which can be solved to calculate the optimal values of K, τ, and θ. A number of software packages are available for minimizing Eq. (8-27).

One operational problem that step forcing causes is the fact that the process under study is moved away from its steady state operating point. Plant managers may be reluctant to allow large steady state changes, since normal production will be disturbed by the changes. As a result, alternative methods of forcing actual processes have been developed, and these included pulse testing and pseudo random binary signal (PRBS) forcing, both of which are illustrated in Fig. 8-22. With pulse forcing, one introduces a step, and then after a period of time the input is returned to its original value. The result is that the process dynamics are excited, but after the forcing, the process returns to its original steady state. PRBS forcing involves a series of pulses of fixed height and random duration, as shown in Fig. 8-22. The advantage of PRBS is that forcing can be concentrated on particular frequency ranges that are important for control-system design.

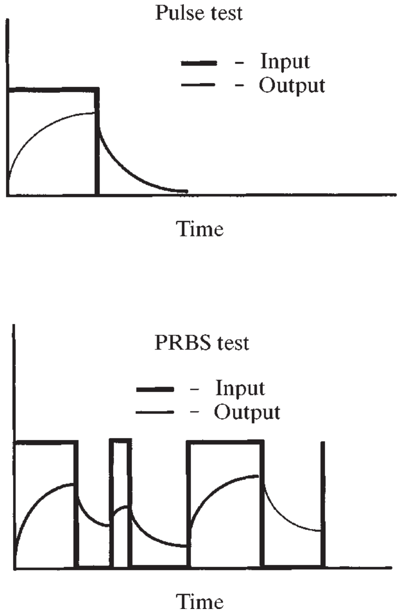


FIG. 8-22 Pulse and PRBS testing.

Transfer function models are linear in nature, but chemical processes are known to exhibit nonlinear behavior. One could use the same type of optimization objective as given in Eq. (8-26) to determine parameters in nonlinear first-principle models, such as Eq. (8-3) presented earlier. Also, nonlinear empirical models, such as neural network models, have recently been proposed for process applications. The key to the use of these nonlinear empirical models is having high-quality process data, which allows the important nonlinearities to be identified.

FEEDBACK CONTROL SYSTEM CHARACTERISTICS

General References: Shinskey, *Feedback Controllers for the Process Industries,* McGraw-Hill, New York, 1994; Seborg, Edgar, and Mellichamp, *Process Dynamics and Control,* Wiley, New York, 1989.

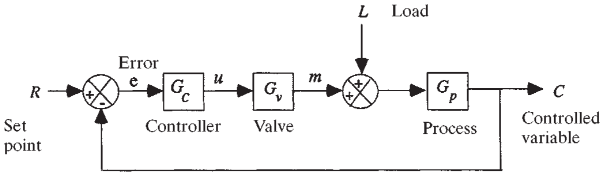


FIG. 8-23 Both load regulation and setpoint response require high gains for the feedback controller.

There are two objectives in applying feedback control: regulating the controlled variable at set point following changes in load, and responding to set-point changes; the latter called servo operation. In fluid processes, almost all control loops must contend with variations in load; therefore, regulation is of primary importance. While most loops will operate continuously at fixed set points, frequent changes in set points can occur in flow loops and in batch production. The most common mechanism for achieving both objectives is feedback control, because it is the simplest and most universally applicable approach to the problem.

Closing the Loop The simplest representation of the closed feedback loop is shown in Fig. 8-23. The load is shown entering the process at the same point as the manipulated variable because that is the most common point of entry, and also because, lacking better information, the transfer function gains in the path of the manipulated variable are the best estimates of those in the load path. In general, the load never impacts directly on the controlled variable without passing through the dominant lag in the process. Where the load is unmeasured, its current value can be observed to be the controller output required to keep the controlled variable *C* at set point *R.*

If the loop is opened, either by placing the controller in manual operation or by setting its gains to zero, the load will have complete influence over the controlled variable, and the set point will have none. Only by closing the loop with controller gains as high as possible will the influence of the load be minimized and that of the set point be maximized. There is a practical limit to the controller gains, however, at the point where the controlled variable develops a uniform oscillation (see Fig. 8-24). This is defined as the limit of stability, and it is reached when the product of gains in the loop $|G_c G_v G_p|$ for that frequency of oscillation is equal to 1.0. If a change in a parameter in the loop causes an increase from this condition, oscillations will expand, creating a dangerous situation where safe limits of operation could be exceeded. Consequently, control loops should be left in a condition

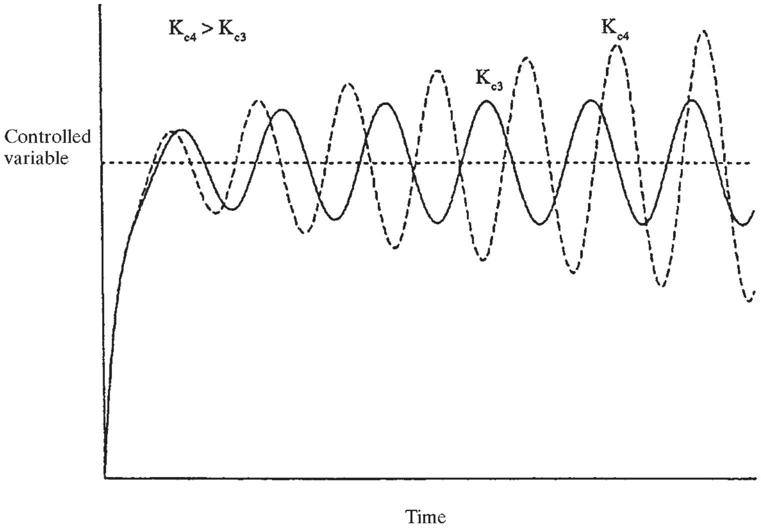


FIG. 8-24 Transition to instability as controller gain increases.

where the loop gain is less than 1.0 by a safe margin that allows for possible variations in process parameters.

In controller design, a choice must be made between performance and robustness. Performance is a measure of how well a given controller with certain parameter settings regulates a variable relative to the best loop performance with optimal controller settings. Robustness is a measure of how small a change in a process parameter is required to bring the loop from its current state to the limit of stability. Increasing controller performance by raising its gains can be expected to decrease robustness. Both performance and robustness are functions of the process being controlled, the selection of the controller, and the tuning of the controller parameters.

On/Off Control An on/off controller is used for manipulated variables having only two states. They commonly control temperatures in homes, electric water-heaters and refrigerators, and pressure and liquid level in pumped storage systems. On/off control is satisfactory where slow cycling is acceptable because it always leads to cycling when the load lies between the two states of the manipulated variable. The cycle will be positioned symmetrically about the set point only if the normal value of the load is equidistant between the two states of the manipulated variable. The period of the symmetrical cycle will be approximately 4θ, where θ is the deadtime in the loop. If the load is not centered between the states of the manipulated variable, the period will tend to increase, and the cycle follows a sawtooth pattern.

Every on/off controller has some degree of deadband, also known as lockup, or differential gap. Its function is to prevent erratic switching between states, thereby extending the life of contacts and motors. Instead of changing states precisely when the controlled variable crosses set point, the controller will change states at two different points for increasing and decreasing signals. The difference between these two switching points is the deadband (see Fig. 8-25); it increases the amplitude and period of the cycle, similar to the effect of dead time.

A three-state controller is used to drive either a pair of independent on/off actuators such as heating and cooling valves, or a bidirectional motorized actuator. The controller is actually two on/off controllers, each with deadband, separated by a dead zone. When the controlled variable lies within the dead zone, neither output is energized. This controller can drive a motorized valve to the point where the manipulated variable matches the load, thereby avoiding cycling.

Proportional Control A proportional controller moves its output proportional to the deviation in the controlled variable from set point:

$$u = K_c e + b = \frac{100}{P} e + b \tag{8-28}$$

where $e = \pm(r - c)$, the sign selected to produce negative feedback. In some controllers, proportional gain K_c is expressed as a pure number; in others, it is set as 100/P, where P is the proportional band in percent. The output bias b of the controller is also known as manual reset. The proportional controller is not a good regulator, because any change in output to a change in load results in a corresponding change in the controlled variable. To minimize the resulting offset, the bias should be set at the best estimate of the load and the proportional band set as low as possible. Processes requiring a proportional band of more than a few percent will control with unacceptable values of offset.

Proportional control is most often used for liquid level where variations in the controlled variable carry no economic penalty, and where other control modes can easily destabilize the loop. It is actually recommended for controlling the level in a surge tank when manipulating the flow of feed to a critical downstream process. By setting the proportional band just under 100 percent, the level is allowed to vary over the full range of the tank capacity as inflow fluctuates, thereby minimizing the resulting rate of change of manipulated outflow. This technique is called averaging level control.

Proportional-plus-Integral (PI) Control Integral action eliminates the offset described above by moving the controller output at a rate proportional to the deviation from set point. Although available alone in an integral controller, it is most often combined with proportional action in a PI controller:

$$u = \frac{100}{P}\left(e + \frac{1}{\tau_I}\int e\,dt\right) + C_0 \tag{8-29}$$

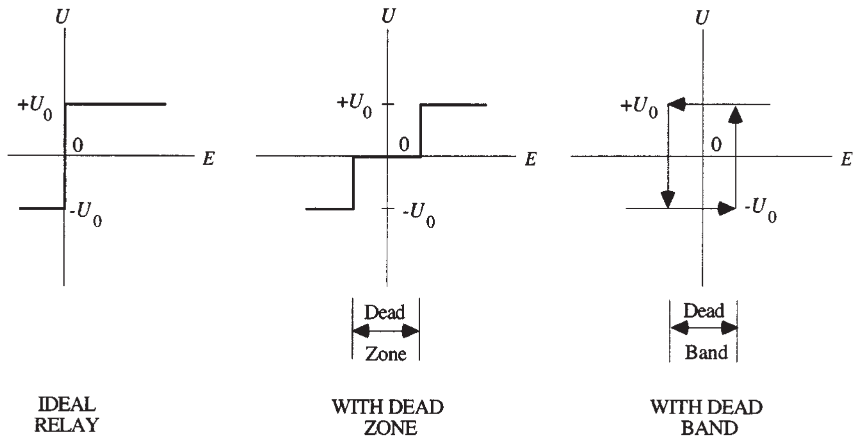
where τ_I is the integral time constant in minutes; in some controllers, it is introduced as integral gain or reset rate 1/τ_I in repeats per minute. The last term in the equation is the constant of integration, the value the controller output has when integration begins.

The PI controller is by far the most commonly used controller in the process industries. The summation of the deviation with its integral in the above equation can be interpreted in terms of frequency response of the controller (Seborg, Edgar, and Mellichamp, *Process Dynamics and Control,* Wiley, New York, 1989). The PI controller produces a phase lag between zero and 90 degrees:

$$\phi_{PI} = -\tan^{-1}\frac{\tau_0}{2\pi\tau_I} \tag{8-30}$$

where τ_0 is the period of oscillation of the loop. The phase angle should be kept between 15 degrees for lag-dominant processes and 45 degrees for dead-time-dominant processes for optimum results.

Proportional-plus-Integral-plus-Derivative (PID) Control The derivative mode moves the controller output as a function of the rate-of-change of the controlled variable, which adds phase lead to the controller, increasing its speed of response. It is normally combined



IDEAL RELAY

WITH DEAD ZONE

WITH DEAD BAND

FIG. 8-25 On/off controller characteristics.

with proportional and integral modes. The noninteracting form of the PID controller appears functionally as:

$$u = \frac{100}{P}\left(e + \frac{1}{\tau_I}\int e\,dt + \tau_D\frac{dc}{dt}\right) + C_0 \qquad (8\text{-}31)$$

where τ_D is the derivative time constant. Note that derivative action is applied to the controlled variable rather than to the deviation, as it should not be applied to the set point; the selection of the sign for the derivative term must be consistent with the action of the controller. Figure 8-26 compares typical loop responses for P, PI, and PID controllers, along with the uncontrolled case.

In some analog PID controllers, the integral and derivative terms are combined serially rather than in parallel as done in the last equation. This results in interaction between these modes, such that the effective values of the controller parameters differ from their set values as follows:

$$\tau_{I_{\text{eff}}} = \tau_I + \tau_D$$

$$\tau_{D_{\text{eff}}} = \frac{1}{1/\tau_D + 1/\tau_I}$$

$$K_c = \frac{100}{P}\left(1 + \frac{\tau_D}{\tau_I}\right) \qquad (8\text{-}32)$$

The performance of the interacting controller is almost as good as the noninteracting controller on most processes, but the tuning rules differ because of the above relationships. With digital PID controllers, the noninteracting version is commonly used.

There is always a gain limit placed upon the derivative term—a value of 10 is typical. However, interaction decreases the derivative gain below this value by the factor $1 + \tau_D/\tau_I$, which is the reason for the decreased performance of the interacting PID controller. Sampling in a digital controller has a similar effect, limiting derivative gain to the ratio of derivative time to the sample interval of the controller. Noise on the controlled variable is amplified by derivative action, preventing its use in controlling flow and liquid level. Derivative action is recommended for control of temperature and composition, reducing the integrated error (IE) by a factor of two over PI control with no loss in robustness (Shinskey, *Feedback Controllers for the Process Industries,* McGraw-Hill, New York, 1994).

CONTROLLER TUNING

The performance of a controller depends as much on its tuning as its design. Tuning must be applied by the end user to fit the controller to the controlled process. There are many different approaches to controller tuning based on the particular performance criteria selected, whether load or set-point changes are most important, whether the process is lag- or deadtime-dominant, and the availability of information about the process dynamics. The earliest definitive work in this field was done at the Taylor Instrument Company by Ziegler and Nichols (Trans. ASME, 759, 1942), tuning PI and interacting PID controllers for optimum response to step load changes applied to lag-dominant processes. While these tuning rules are still in use, they are approximate and do not apply to set-point changes, dead-time-dominant processes, or noninteracting PID controllers (Seborg, Edgar, and Mellichamp, *Process Dynamics and Control,* Wiley, New York, 1989).

Controller Performance Criteria The most useful measures of controller performance in an industrial setting are the maximum deviation in the controlled variable resulting from a disturbance and its integral. The disturbance could be to the set point or to the load, depending on the variable being controlled and its context in the process. The size of the deviation and its integral are proportional to the size of the disturbance (if the loop is linear at the operating point). While actual disturbances in a plant setting may appear to be random, the controller needs a reliable test to determine how well it is tuned. The disturbance of choice for test purposes is the step, because it can be applied manually, and by containing all frequencies including zero, it exercises all modes of the controller. When tuned optimally for step disturbances, the controller should be well-tuned for most other disturbances as well.

Figure 8-27 shows the optimum response of a controlled variable to a step change in load. A step change in load may be simulated by stepping the controller output while it is in the manual mode followed immediately by transfer to automatic. The maximum deviation is the most important criterion for variables that could exceed safe operating levels such as steam pressure, drum level, and steam temperature in a boiler. The same rule applies to product quality, which could violate specifications and therefore be rejected. If the product can be accumulated in a downstream storage tank, however, its average quality is more important, and this is a function of the deviation integrated over the residence time in the tank. Deviation in the other direction, where the product is better than specification, is safe, but it increases production costs in proportion to the integrated deviation because quality is given away.

For a PI or PID controller, the integrated deviation—better known as the integrated error IE—is related to the controller settings:

$$\text{IE} = \frac{\Delta u P \tau_I}{100} \qquad (8\text{-}33)$$

where Δu is the difference in controller output between two steady states, as required by a change in load or set point. The proportional band P and integral time τ_I are the indicated settings of the controller

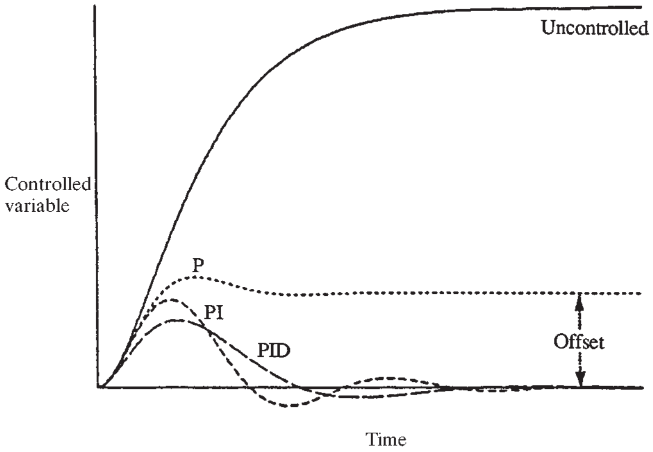


FIG. 8-26 Response for a step change in disturbance with tuned P, PI, and PID controllers and with no control.

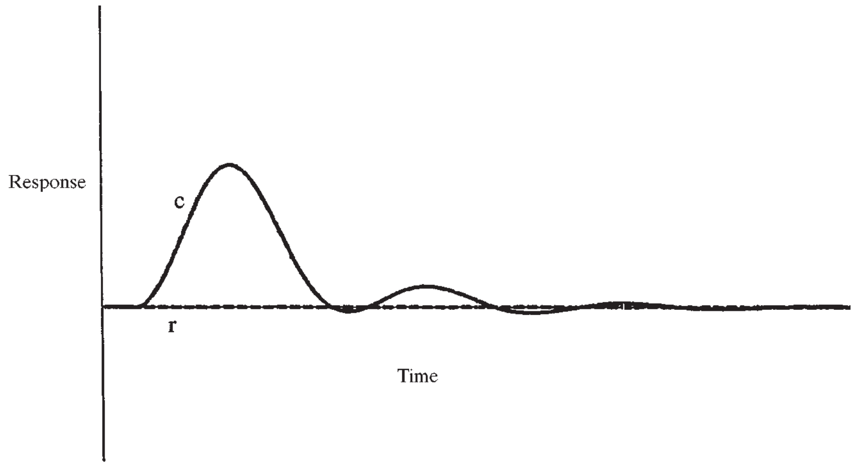


FIG. 8-27 The minimum-IAE response to a step load change has little overshoot and is well-damped.

for both interacting and noninteracting PID controllers. Although the derivative term does not appear in the relationship, its use typically allows a 50 percent reduction in integral time and therefore in IE. The integral time in the IE expression should be augmented by the sample interval if the controller is digital, the time constant of any filter used, and the value of any deadtime compensator.

It would appear from the above that minimizing IE is simply a matter of minimizing the P and τ_I settings of the controller. However, settings will be reached that produce uniform oscillations—an unacceptable situation. It is preferable, instead, to find a combination of controller settings that minimize integrated absolute error IAE, which for both load and set-point changes is a well-damped response with minimal overshoot. Figure 8-27 is an example of a minimum-IAE response to a step change in load for a lag-dominant process. Because of the very small overshoot, the IAE will be only slightly larger than the IE. Loops that are tuned to minimize IAE tend to be close to minimum IE and also minimum peak deviation.

The performance of a controller (and its tuning) must be based on what is achievable for a given process. The concept of best practical IE (IE_b) for a step change in load Δq can be estimated (Shinskey, *Feedback Controllers for the Process Industries,* McGraw-Hill, New York, 1994):

$$\text{IE}_b = \Delta q K_L \tau_L (1 - e^{-\theta/\tau_L}) \tag{8-34}$$

where K_L is the gain and τ_L the primary time constant in the load path, and θ the dead time in the manipulated path to the controlled variable. If the load or its gain is unknown, Δu and $K(= K_v K_p)$ may be substituted. If the process is non-self-regulating (i.e., it is an integrator), the relationship is

$$\text{IE}_b = \frac{\Delta q \theta^2}{\tau_1} \tag{8-35}$$

where τ_1 is the time constant of the process integrator. The peak deviation with the best practical response curve is:

$$e_b = \frac{\text{IE}_b}{\theta + \tau_2} \tag{8-36}$$

where τ_2 is the time constant of a common secondary lag (e.g., in the measuring device).

The performance for any controller can be measured against this standard by comparing the IE it achieves in responding to a step load change with the best practical IE. Potential performance improvements by tuning PI controllers on lag-dominant processes lie in the 20–30 percent range, while for PID controllers they fall between 40–60 percent, varying with secondary lags.

Tuning Methods Based on Known Process Models The most accurate tuning rules for controllers have been based on simulation, where the process parameters can be specified and IAE and IE can be integrated during the simulation as an indication of performance. Controller settings are then iterated until a minimum IAE is reached for a given disturbance. These optimum settings are then related to the parameters of the simulated process in tables, graphs, or equations, as a guide to tuning controllers for processes whose parameters are known (Seborg, Edgar, and Mellichamp, *Process Dynamics and Control,* Wiley, New York, 1989). This is a multidimensional problem, however, in that the relationships change as a function of process type, controller type, and source of disturbance.

Table 8-2 summarizes these rules for minimum-IAE load response for the most common controllers. The process gain K and time constant τ_m are obtained from the product of G_v and G_p in Fig. 8-23. Derivative action is not effective for dead-time-dominant processes. For non-self-regulating processes, τ is the time constant of the integrator. The last category of distributed lag includes all heat-transfer processes, backmixed vessels, and processes having multiple interacting lags such as distillation columns; $\sum \tau$ represents the total response time of these processes (i.e., the time required for 63 percent complete response to a step input). Any secondary lag, sampling interval, or filter time constant should be added to deadtime θ.

The principal limitation to using these rules is that the true process parameters are often unknown. Steady-state gain K can be calculated from a process model or determined from the steady-state results of a step test as $\Delta c/\Delta u$, as shown in Fig. 8-28. The test will not be viable, however, if the time constant of the process τ_m is longer than a few

TABLE 8-2 Tuning Rules Using Known Process Parameters

Process	Controller	P	τ_I	τ_D
Dead-time-dominant	PI	250K	0.5 θ	
Lag-dominant	PI	106K θ/τ_m	4.0 θ	
	PID_n	77K θ/τ_m	1.8 θ	0.45 θ
	PID_i	106K θ/τ_m	1.5 θ	0.55 θ
Non-self-regulating	PI	106 θ/τ_1	4.0 θ	
	PID_n	78 θ/τ_1	1.9 θ	0.48 θ
	PID_i	108 θ/τ_1	1.6 θ	0.58 θ
Distributed lags	PI	20K	0.50 $\sum \tau$	
	PID_n	10K	0.30 $\sum \tau$	0.09 $\sum \tau$
	PID_i	15K	0.25 $\sum \tau$	0.10 $\sum \tau$

NOTE: n = noninteracting; i = interacting controller modes

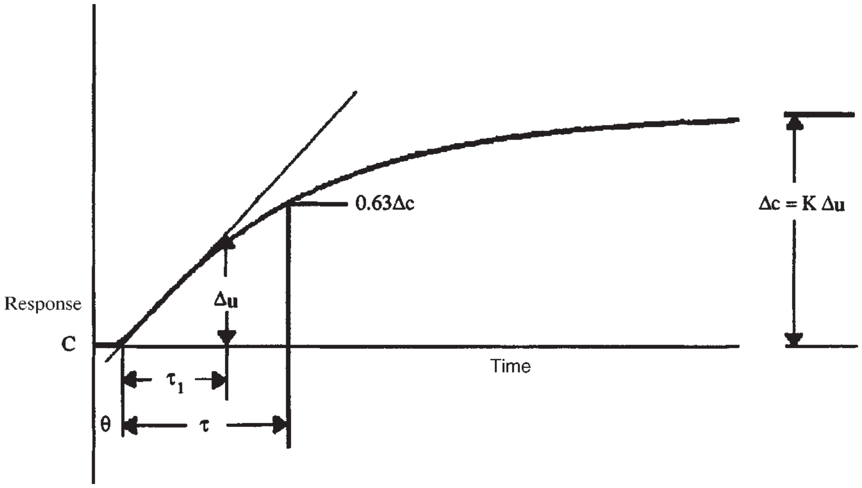


FIG. 8-28 If a steady state can be reached, gain K and time constant τ can be estimated from a step response; if not, use τ_1 instead.

minutes, since five time constants must elapse to approach a steady state within one percent, and unrequested disturbances may intervene. Estimated dead-time θ is the time from the step to the intercept of a straight line tangent to the steepest part of the response curve. The estimated time constant τ is the time from that point to 63 percent of the complete response. In the presence of a secondary lag, these results will not be completely accurate, however. The time for 63 percent response may be more accurately calculated as the residence time of the process: its volume divided by current volumetric flow rate.

Tuning Methods When Process Model Is Unknown Ziegler and Nichols developed two tuning methods for processes with unknown parameters. The open-loop method uses a step test without waiting for a steady state to be reached and is therefore applicable to very slow processes. Deadtime is estimated from the intercept of the tangent in Fig. 8-28, whose slope is also used. If the process is non-self-regulating, the controlled variable will continue to follow this slope, changing an amount equal to Δu in a time equal to its time constant. This time estimate τ_1 is used along with θ to tune controllers according to Table 8-3, applicable to lag-dominant processes.

A more recent tuning approach uses integral criteria such as the integral of the squared error (ISE), integral of the absolute error (IAE), and the time-weighted IAE (ITAE) of Seborg, Edgar, and Mellichamp (*Process Dynamics and Control,* Wiley, New York, 1989). The controller parameters are selected to minimize various integrals. Power-law correlations for PID controller settings have been tabulated for a range of first-order model parameters. The best tuning parameters have been fitted using a general equation, $Y = A(\theta/\tau)^B$, where Y depends on the particular controller mode to be evaluated (K_C, τ_I, τ_D).

There are several features of the correlations that should be noted:

1. The controller gain is inversely proportional to the process gain for constant dead time and time constant.

2. The allowable controller gain is higher when the ratio of dead time to time constant becomes smaller. This is because dead time has a destabilizing effect on the control system, limiting the controller gain, while a larger time constant generally demands a higher controller gain.

A recent addition to the model-based tuning correlations is Internal Model Control (Rivera, Morari, and Skogestad, "Internal Model Control 4: PID Controller Design," *IEC Proc. Des. Dev.,* **25,** 252, 1986), which offers some advantages over the other methods described here. However, the correlations are similar to the ones discussed above. Other plant testing and controller design approaches such as frequency response can be used for more complicated models.

The Ziegler and Nichols closed-loop method requires forcing the loop to cycle uniformly under proportional control. The natural period τ_n of the cycle—the proportional controller contributes no phase shift to alter it—is used to set the optimum integral and derivative time constants. The optimum proportional band is set relative to the undamped proportional band P_u, which produced the uniform oscillation. Table 8-4 lists the tuning rules for a lag-dominant process. A uniform cycle can also be forced using on/off control to cycle the manipulated variable between two limits. The period of the cycle will be close to τ_n if the cycle is symmetrical; the peak-to-peak amplitude of the controlled variable A_c divided by the difference between the output limits A_m is a measure of process gain at that period and is therefore related to P_u for the proportional cycle:

$$P_u = 100\,\frac{\pi}{4}\,\frac{A_c}{A_m} \tag{8-37}$$

The factor $\pi/4$ compensates for the square wave in the output. Tuning rules are given in Table 8-4.

TABLE 8-3 Tuning Rules Using Slope and Intercept

Controller	P	τ_I	τ_D
PI	150 θ/τ	3.5 θ	—
PID$_n$	75 θ/τ	2.1 θ	0.63 θ
PID$_i$	113 θ/τ	1.8 θ	0.70 θ

NOTE: n = noninteracting, i = interacting controller modes

TABLE 8-4 Tuning Rules Using Proportional Cycle

Controller	P	τ_I	τ_D
PI	1.70 P_u	0.81 τ_n	—
PID$_n$	1.30 P_u	0.48 τ_n	0.11 τ_n
PID$_i$	1.80 P_u	0.39 τ_n	0.14 τ_n

ADVANCED CONTROL SYSTEMS

BENEFITS OF ADVANCED CONTROL

The economics of most processes are determined by the steady-state operating conditions. Excursions from these steady-state conditions generally average out and have an insignificant effect on the economics of the process, except when the excursions lead to off-specification products. In order to enhance the economic performance of a process, the steady-state operating conditions must be altered in a manner that leads to more efficient process operation.

The following hierarchy is used for process control:
Level 0: Measurement devices and actuators
Level 1: Regulatory control
Level 2: Supervisory control
Level 3: Production control
Level 4: Information technology

Levels 2, 3, and 4 clearly affect the process economics, as all three levels are directed to optimizing the process in some manner. However, level 0 (measurement devices and actuators) and level 1 (regulatory control) would appear to have no effect on process economics. Their direct effect is indeed minimal, but indirectly, they have a major effect. Basically, these levels provide the foundation for all higher levels. A process cannot be optimized until it can be operated consistently at the prescribed targets. Thus, a high degree of regulatory control must be the first goal of any automation effort. In turn, the measurements and actuators provide the process interface for regulatory control.

For most processes, the optimum operating point is determined by a constraint. The constraint might be a product specification (a product stream can contain no more than 2 percent ethane); violation of this constraint causes off-specification product. The constraint might be an equipment limit (vessel pressure rating is 300 psig); violation of this constraint causes the equipment protection mechanism (pressure relief device) to activate. As the penalties are serious, violation of such constraints must be very infrequent.

If the regulatory control system were perfect, the target could be set exactly equal to the constraint (that is, the target for the pressure controller could be set at the vessel relief pressure). However, no regulatory control system is perfect. Therefore, the value specified for the target must be on the safe side of the constraint, thus giving the control system some “elbow room.” How much depends on the following:

1. *The performance of the control system* (*i.e., how effectively it responds to disturbances*). The faster the control system reacts to a disturbance, the closer the process can be operated to the constraint.
2. *The magnitude of the disturbances to which the control system must respond.* If the magnitude of the major disturbances can be reduced, the process can be operated closer to the constraint.

One measure of the performance of a control system is the variance of the controlled variable from the target. Both improving the control system and reducing the disturbances will lead to a lower variance in the controlled variable.

In a few applications, improving the control system leads to a reduction in off-specification product and thus improved process economics. However, in most situations, the process is operated sufficiently far from the constraint that very little, if any, off-specification product results from control system deficiencies. Management often places considerable emphasis on avoiding off-spec production, so consequently the target is actually set far more conservatively than it should be.

In most applications, simply improving the control system does not directly lead to improved process economics. Instead, the control system improvement must be accompanied by shifting the target closer to the constraint. There is always a cost of operating a process in a conservative manner. The cost may be a lower production rate, a lower process efficiency, a product giveaway, or otherwise. When management places extreme emphasis on avoiding off-spec production, the natural reaction is to operate very conservatively, thus incurring other costs.

The immediate objective of an advanced control effort is to reduce the variance in an important controlled variable. However, this effort must be coupled with a commitment to adjust the target for this controlled variable so that the process is operated closer to the constraint. In large throughput (commodity) processes, very small shifts in operating targets can lead to large economic returns.

ADVANCED CONTROL TECHNIQUES

General References: Seborg, Edgar, and Mellichamp, *Process Dynamics and Control,* John Wiley and Sons, New York, 1989. Stephanopoulos, *Chemical Process Control: An Introduction to Theory and Practice,* Prentice Hall, Englewood Cliffs, New Jersey, 1984. Shinskey, *Process Control Systems,* 3d ed., McGraw-Hill, New York, 1988. Ogunnaike and Ray, *Process Dynamics, Modeling, and Control,* Oxford University Press, New York, 1994.

While the single-loop PID controller is satisfactory in many process applications, it does not perform well for processes with slow dynamics, time delays, frequent disturbances, or multivariable interactions. We discuss several advanced control methods hereafter that can be implemented via computer control, namely feedforward control, cascade control, time-delay compensation, selective and override control, adaptive control, fuzzy logic control, and statistical process control.

Feedforward Control If the process exhibits slow dynamic response and disturbances are frequent, then the application of feedforward control may be advantageous. Feedforward (FF) control differs from feedback (FB) control in that the primary disturbance or load (L) is measured via a sensor and the manipulated variable (*m*) is adjusted so that deviations in the controlled variable from the set point are minimized or eliminated (see Fig. 8-29). By taking control action based on measured disturbances rather than controlled variable error, the controller can reject disturbances before they affect the controlled variable *c*. In order to determine the appropriate settings for the manipulated variable, one must develop mathematical models that relate:

1. The effect of the manipulated variable on the controlled variable
2. The effect of the disturbance on the controlled variable

These models can be based on steady-state or dynamic analysis. The performance of the feedforward controller depends on the accuracy of both models. If the models are exact, then feedforward control offers the potential of perfect control (i.e., holding the controlled variable precisely at the set point at all times because of the ability to predict the appropriate control action). However, since most mathematical models are only approximate and since not all disturbances are measurable, it is standard practice to utilize feedforward control in conjunction with feedback control. Table 8-5 lists the relative advantages and disadvantages of feedforward and feedback control. By combining the two control methods, the strengths of both schemes can be utilized.

FF control therefore attempts to eliminate the effects of measurable disturbances, while FB control would correct for unmeasurable

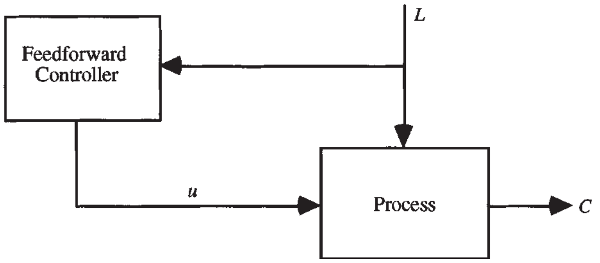


FIG. 8-29 Block diagram for feedforward control configuration.

TABLE 8-5 Relative Advantages and Disadvantages of Feedback and Feedforward

Advantages	Disadvantages
Feedforward	
• Acts before the effect of a disturbance has been felt by the system • Good for systems with large time constant or deadtime • Does not introduce instability in the closed-loop response	• Requires measurement of all possible disturbances and their direct measurement • Cannot cope with unmeasured disturbances • Sensitive to process/model error
Feedback	
• Does not require identification and measurement of any disturbance for corrective action • Does not require an explicit process model • Controller can be robust to process/ model errors	• Control action not taken until the effect of the disturbance has been felt by the system • Unsatisfactory for processes with large time constants and frequent disturbances • May cause instability in the closed-loop response

disturbances and modeling errors. This is often referred to as feedback trim. These controllers have become widely accepted in the chemical process industries since the 1960s.

Design Based on Material and Energy Balances Consider a heat exchanger example (see Fig. 8-30) to illustrate the use of FF and FB control. The control objective is to maintain T_2, the exit liquid temperature, at the desired value (or set point) T_{set} despite variations in inlet liquid flow rate F and inlet liquid temperature T_1. This is done by manipulating W, the steam flow rate. A feedback control scheme would entail measuring T_2, comparing T_2 to T_{set}, and then adjusting W. A feedforward control scheme requires measuring F and T_1, and adjusting W (knowing T_{set}), in order to control exit temperature, T_2.

Figure 8-31 shows the control system diagrams for FB and FF control. A feedforward control algorithm can be designed for the heat exchanger in the following manner. Using a steady-state energy balance and assuming no heat loss from the heat exchanger,

$$WH = FC(T_2 - T_1) \tag{8-38}$$

where H = latent heat of vaporization
C_L = specific heat of liquid.

Rearranging Eq. (8-38),

$$W = \frac{C_L}{H} F(T_2 - T_1) \tag{8-39}$$

or

$$W = K_1F(T_2 - T_1) \tag{8-40}$$

with

$$K_1 = \frac{C_L}{H} \tag{8-41}$$

Replace T_2 by T_{set}:

$$W = K_1F(T_{set} - T_1) \tag{8-42}$$

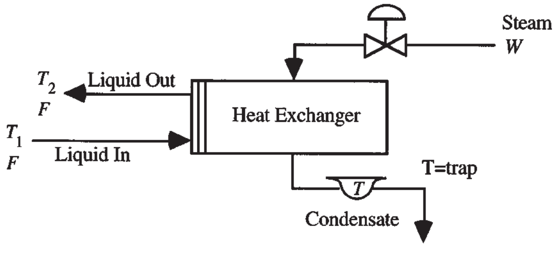
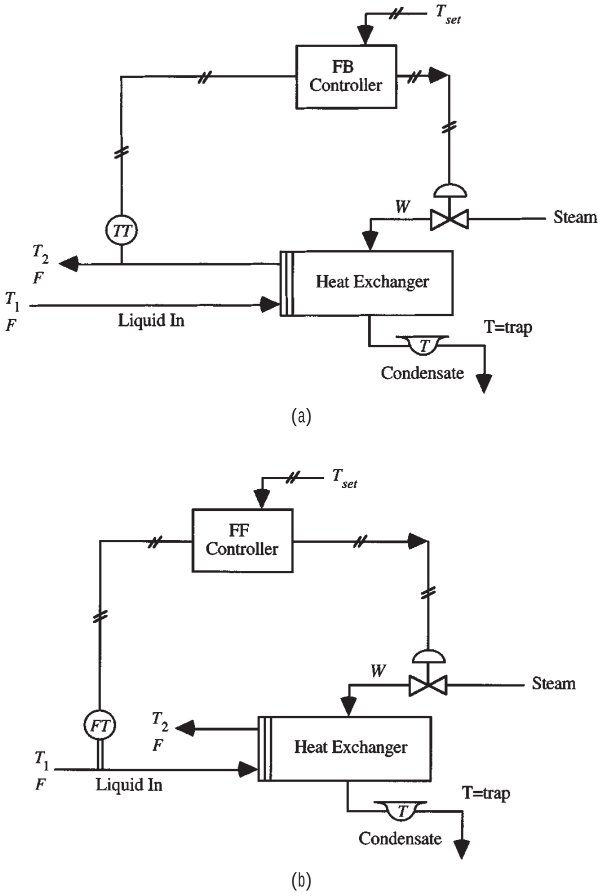


FIG. 8-30 A heat exchanger diagram.



(a)

(b)

FIG. 8-31 (*a*) Feedback control of a heat exchanger. (*b*) Feedforward control of a heat exchanger.

Equation (8-42) can be used in the FF calculation, assuming one knows the physical properties C_L and H. Of course, it is probable that the model will contain errors (e.g., unmeasured heat losses, incorrect C_L or H). Therefore, K_1 can be designated as an adjustable parameter that can be tuned. The use of a physical model for FF control is desirable since it provides a physical basis for the control law and gives an a priori estimate of what the tuning parameters are. Note that such a model could be nonlinear [e.g., in Eq. (8-42), F and T_{set} are multiplied].

Block Diagram Analysis One shortcoming of this feedforward design procedure is that it is based on the steady-state characteristics of the process and as such, neglects process dynamics (i.e., how fast the controlled variable responds to changes in the load and manipulated variables). Thus, it is often necessary to include "dynamic compensation" in the feedforward controller. The most direct method of designing the FF dynamic compensator is to use a block diagram of a general process, as shown in Fig. 8-32. G_t represents the disturbance transmitter, G_f is the feedforward controller, G_L relates the load to the controlled variable, G_v is the valve, and G_p is the process. G_m is the output transmitter and G_c is the feedback controller. All blocks correspond to transfer functions (via Laplace transforms).

Using block diagram algebra and Laplace transform variables, the controlled variable $C(s)$ is given by

$$C(s) = \frac{G_tG_fL(s) + G_LL(s)}{1 + G_mG_cG_vG_p} \tag{8-43}$$

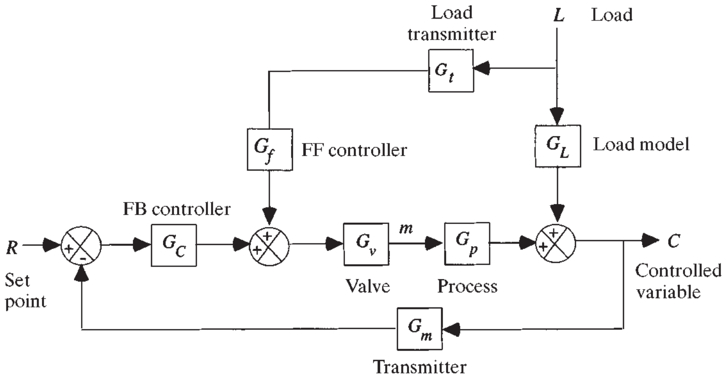


FIG. 8-32 Block diagram for feedback-feedforward control.

For disturbance rejection [$L(s) \neq 0$] we require that $C(s) = 0$, or zero error. Solving Eq. (8-43) for G_f,

$$G_f = \frac{-G_L}{G_tG_vG_p} \tag{8-44}$$

Suppose the dynamics of G_L and G_p are first order; in addition, assume that $G_v = K_v$ and $G_t = K_t$ (constant gains for simplicity).

$$G_L(s) = \frac{K_L}{\tau_L s + 1} = \frac{C(s)}{L(s)} \tag{8-45}$$

$$G_p(s) = \frac{K_p}{\tau_p s + 1} = \frac{C(s)}{U(s)} \tag{8-46}$$

Using Eq. (8-44),

$$G_f(s) = -\frac{\tau_p s + 1}{\tau_L s + 1} \cdot \frac{K_L}{K_pK_vK_t} = \frac{-K(\tau_p s + 1)}{\tau_L s + 1} \tag{8-47}$$

The above FF controller can be implemented using analog elements or more commonly by a digital computer. Figure 8-33 compares typical responses for PID FB control, steady-state FF control ($s = 0$), dynamic FF control, and combined FF/FB control. In practice, the engineer can tune K, τ_p and τ_L in the field to improve the performance of the FF controller. The feedforward controller can also be simplified to provide steady-state feedforward control. This is done by setting $s = 0$ in $G_f(s)$. This might be appropriate if there is uncertainty in the dynamic models for G_L and G_p.

Other Considerations in Feedforward Control The tuning of feedforward and feedback control systems can be performed independently. In analyzing the block diagram in Fig. 8-32, note that G_f is chosen to cancel out the effects of the disturbance $L(s)$ as long as there are no model errors. For the feedback loop, therefore, the effects of $L(s)$ can also be ignored, which for the servo case is:

$$\frac{C(s)}{R(s)} = \frac{G_cG_vG_pK_m}{1 + G_cG_vG_pG_m} \tag{8-48}$$

Note that the characteristic equation will be unchanged for the FF + FB system, hence system stability will be unaffected by the presence of the FF controller. In general, the tuning of the FB controller can be less conservative than for the case of FB alone, since smaller excursions from the set point will result. This in turn would make the dynamic model $G_p(s)$ more accurate.

The tuning of the controller in the feedback loop can be theoretically performed independent of the feedforward loop (i.e., the feedforward loop does not introduce instability in the closed-loop response). For more information on feedforward/feedback control applications and design of such controllers, refer to the general references.

Cascade Control One of the disadvantages of using conventional feedback control for processes with large time lags or delays is that disturbances are not recognized until after the controlled variable deviates from its set point. In these processes, correction by feedback control is generally slow and results in long-term deviation from set point. One way to improve the dynamic response to load changes is by using a secondary measurement point and a secondary controller; the secondary measurement point is located so that it recognizes the upset condition before the primary controlled variable is affected.

One such approach is called cascade control, which is routinely used in most modern computer control systems. Consider a chemical reactor, where reactor temperature is to be controlled by coolant flow to the jacket of the reactor (Fig. 8-34). The reactor temperature can be influenced by changes in disturbance variables such as feed rate or feed temperature; a feedback controller could be employed to compensate for such disturbances by adjusting a valve on the coolant flow to the reactor jacket. However, suppose an increase occurs in the

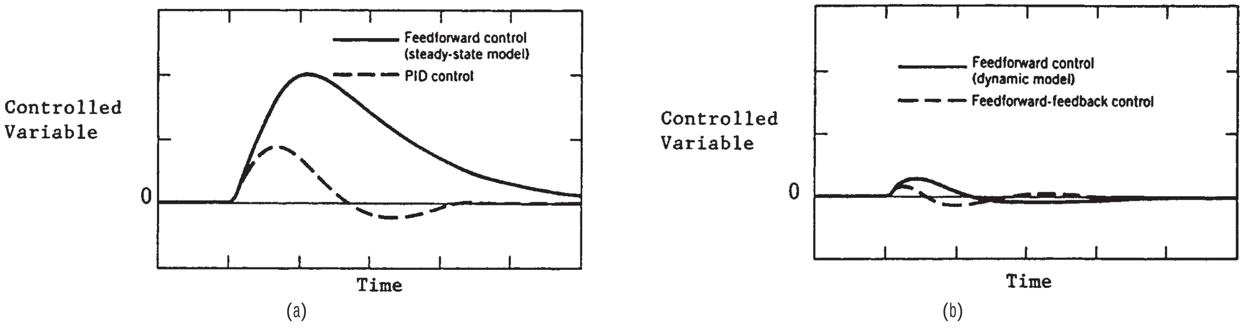


FIG. 8-33 (*a*) Comparison of FF (steady state model) and PID FB control for load change; (*b*) comparison of FF (dynamic model) and combined FF/FB control.

coolant temperature as a result of changes in the plant coolant system. This will cause a change in the reactor temperature measurement, although such a change will not occur quickly, and the corrective action taken by the controller will be delayed.

Cascade control is one solution to this problem (see Fig. 8-35). Here the jacket temperature is measured, and an error signal is sent from this point to the coolant control valve; this reduces coolant flow, maintaining the heat transfer rate to the reactor at a constant level and rejecting the disturbance. The cascade control configuration will also adjust the setting of the coolant control valve when an error occurs in reactor temperature. The cascade control scheme shown in Fig. 8-35 contains two controllers. The primary controller is the reactor temperature coolant temperature controller. It measures the reactor temperature, compares it to the set point, and computes an output, which is the set point for the coolant flow rate controller. This secondary controller compares the set point to the coolant temperature measurement and adjusts the valve. The principal advantage of cascade control is that the secondary measurement (jacket temperature) is located closer to a potential disturbance in order to improve the closed-loop response.

Figure 8-36 shows the block diagram for a general cascade control system. In tuning of a cascade control system, the secondary controller (in the inner loop) is tuned first with the primary controller in manual. Often only a proportional controller is needed for the secondary loop, since offset in the secondary loop can be treated by using proportional plus integral action in the primary loop. When the primary controller is transferred to automatic, it can be tuned using the techniques described earlier in this section. For more information on theoretical analysis of cascade control systems, see the general references for a discussion of applications of cascade control.

Time-Delay Compensation Time delays are a common occurrence in the process industries because of the presence of recycle loops, fluid-flow distance lags, and "dead time" in composition measurements resulting from use of chromatographic analysis. The presence of a time delay in a process severely limits the performance of a conventional PID control system, reducing the stability margin of the closed-loop control system. Consequently, the controller gain must be reduced below that which could be used for a process without delay. Thus, the response of the closed-loop system will be sluggish compared to that of the system with no time delay.

In order to improve the performance of time-delay systems, special control algorithms have been developed to provide time-delay compensation. The Smith predictor technique is the best known algorithm; a related method is called the analytical predictor. Various investigators have found that based on integral squared error, the performance of the Smith predictor can be as much as 30 percent better than for a conventional controller.

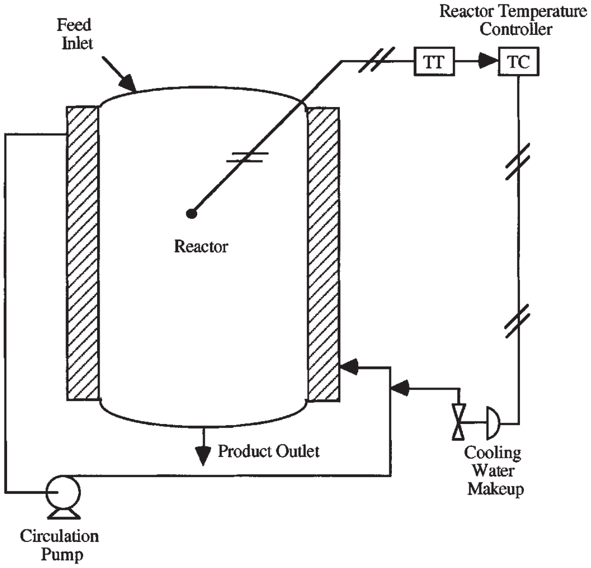


FIG. 8-34 Conventional control of an exothermic chemical reactor.

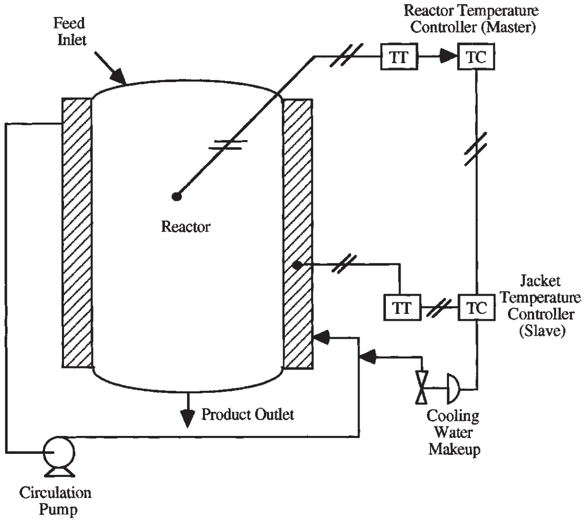


FIG. 8-35 Cascade control of an exothermic chemical reactor.

The Smith predictor is a model-based control strategy that involves a more complicated block diagram than that for a conventional feedback controller, although a PID controller is still central to the control strategy (see Fig. 8-37). The key concept is based on better coordination of the timing of manipulated variable action. The loop configuration takes into account the fact that the current controlled variable measurement is not a result of the current manipulated variable action, but the value taken 0 time units earlier. Time-delay compensation can yield excellent performance; however, if the process model parameters change (especially the time delay), the Smith predictor performance will deteriorate and is not recommended unless other precautions are taken.

Selective and Override Control When there are more controlled variables than manipulated variables, a common solution to this problem is to use a selector to choose the appropriate process variable from among a number of available measurements. Selectors can be based on either multiple measurement points, multiple final control elements, or multiple controllers, as discussed below. Selectors are used to improve the control system performance as well as to protect equipment from unsafe operating conditions.

One type of selector device chooses as its output signal the highest (or lowest) of two or more input signals. This approach is often referred to as auctioneering. On instrumentation diagrams, the symbol HS denotes high selector and LS a low selector. For example, a high selector can be used to determine the hot-spot temperature in a fixed-bed chemical reactor. In this case, the output from the high selector is the input to the temperature controller. In an exothermic catalytic reaction, the process may run away due to disturbances or changes in the reactor. Immediate action should be taken to prevent a dangerous rise in temperature. Because a hot spot may potentially develop at one of several possible locations in the reactor, multiple (redundant) measurement points should be employed. This approach minimizes the time required to identify when a temperature has risen too high at some point in the bed.

The use of high or low limits for process variables is another type of selective control, called an override. The feature of anti-reset windup in feedback controllers is a type of override. Another example is a distillation column with lower and upper limits on the heat input to the column reboiler. The minimum level ensures that liquid will remain

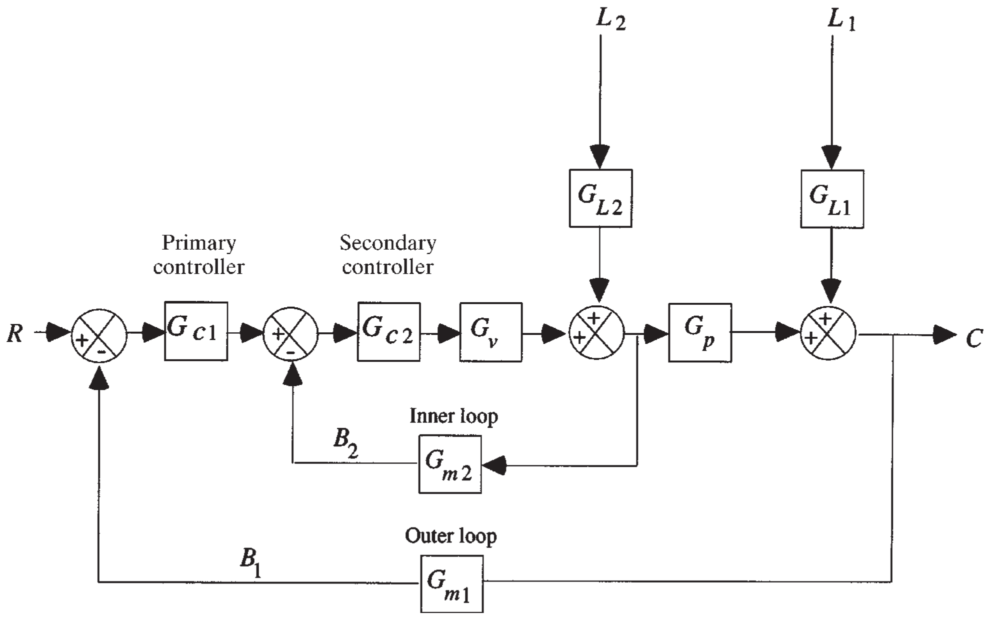


FIG. 8-36 Block diagram of the cascade control system. For a chemical reactor, L_1 would correspond to a feed temperature or composition disturbance, while L_2 would be a change in the cooling water temperature.

on the trays, while the upper limit is determined by the onset of flooding. Overrides are also used in forced-draft combustion-control systems to prevent an imbalance between air flow and fuel flow, which could result in unsafe operating conditions.

Other types of selective systems employ multiple final control elements or multiple controllers. In some applications, several manipulated variables are used to control a single process variable (also called split-range control). Typical examples include the adjustment of both inflow and outflow from a chemical reactor in order to control reactor pressure or the use of both acid and base to control pH in waste-water treatment. In this approach, the selector chooses from several controller outputs which final control element should be adjusted (Marlin, *Process Control,* McGraw-Hill, New York, 1995).

Adaptive Control Process control problems inevitably require on-line tuning of the controller constants to achieve a satisfactory degree of control. If the process operating conditions or the environment changes significantly, the controller may have to be retuned. If these changes occur quite frequently, then adaptive control techniques should be considered. An adaptive control system is one in which the controller parameters are adjusted automatically to compensate for changing process conditions.

During the 1980s, several adaptive controllers were field-tested and commercialized in the U.S. and abroad, including products by ASEA (Sweden), Leeds and Northrup, Foxboro, and Sattcontrol. At the present time, some form of adaptive tuning is available on almost all PID controllers. The ASEA adaptive controller, Novatune, was

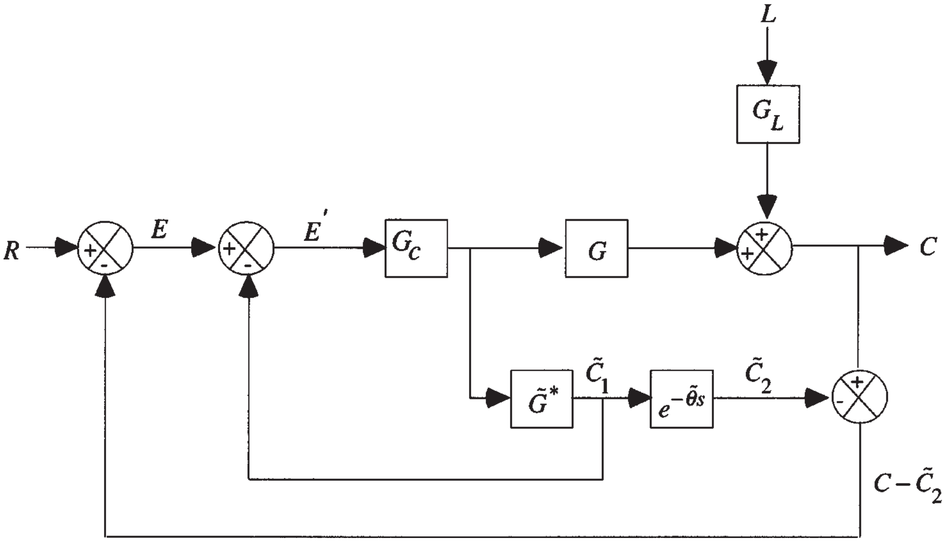


FIG. 8-37 Block diagram of the Smith predictor. The process model used in the controller is $\tilde{G} = \tilde{G}^\circ e^{-\tilde{\theta}s}$ ($\tilde{G}^\circ$ = model without delay; $e^{-\tilde{\theta}s}$ = time delay element).

announced in 1983 and is generally based on minimum-variance-control algorithms. Both feedforward and feedback control capabilities reside in the hardware. The unit has been tested successfully in reactor and paper machine control applications in Europe and in pH control of wastewater in the United States.

Foxboro developed a self-tuning PID controller that is based on a so-called "expert system" approach for adjustment of the controller parameters. The on-line tuning of K_c, τ_I, and τ_D is based on the closed-loop transient response to a step change in set point. By evaluating the salient characteristics of the response (e.g., the decay ratio, overshoot, and closed-loop period), the controller parameters can be updated without actually finding a new process model. The details of the algorithm, however, are proprietary.

The Sattcontroller (also marketed by Fisher-Rosemount) has an autotuning function that is based on placing the process in a controlled oscillation at very low amplitude, comparable with that of the noise level of the process. This is done via a relay-type step function with hysteresis. The autotuner identifies the dynamic parameters of the process (the ultimate gain and period) and automatically calculates K_c, τ_I, and τ_D using empirical tuning rules. Gain scheduling can also be implemented with this controller, using up to three sets of PID controller parameters.

The subject of adaptive control is one of current interest. New algorithms are presently under development, but these need to be field-tested before industrial acceptance can be expected. It is clear, however, that digital computers will be required for implementation of self-adaptive controllers due to their complexity. An adaptive controller is inherently nonlinear and therefore more complicated than the conventional PID controller.

Fuzzy Logic Control The application of fuzzy logic to process control requires the concepts of fuzzy rules and fuzzy inference. A fuzzy rule, also known as a fuzzy IF-THEN statement, has the form:

$$\text{If } x \text{ then } y$$

where x specifies a vector of input variables and corresponding membership values and y specifies an output variable and its corresponding membership value. For example,

if input1 = high
and input2 = low,
then output = medium.

Three functions are required to perform logical inferencing with the fuzzy rules. The fuzzy AND is the product of a rule's input membership values, generating a weight for the rule's output. The fuzzy OR is a normalized sum of the weights assigned to each rule that contributes to a particular decision. The third function used is defuzzification, which generates a crisp final output. In one approach, the crisp output is the weighted average of the peak element values: $\sum [w(i)\, p(i)]/\sum [w(i)]$.

With a single feedback control architecture, information that is readily available to the algorithm includes the error signal, the difference between the process variable and the set point variable, change in error from previous cycles to the current cycle, changes to the set point variable, change of the manipulated variable from cycle to cycle, and the change in the process variable from past to present. In addition, multiple combinations of the system response data are available. As long as the irregularity lies in that dimension wherein fuzzy decisions are being based or associated, the result should be enhanced performance. This enhanced performance should be demonstrated in both the transient and steady-state response. If the system tends to have changing dynamic characteristics or exhibits nonlinearities, fuzzy logic control should offer a better alternative to using constant PID settings. Most fuzzy logic software begins building its information base during the autotune function. In fact, the majority of the information used in the early stages of system startup comes from the autotune solutions.

In addition to single-loop process controllers, products that have benefited from the implementation of fuzzy logic are:

- Camcorders with automatic compensation for operator-injected noise such as shaking and moving
- Elevators with decreased wait time, making intelligent floor decisions and minimizing travel and power consumption
- Antilock braking systems with quickly reacting independent wheel decisions based on current and acquired knowledge
- Television with automatic color, brightness, and acoustic control based on signal and environmental conditions

Sometimes fuzzy logic controllers are combined with pattern recognition software such as artificial neural networks (Kosko, *Neural Networks and Fuzzy Systems,* Prentice Hall, Englewood Cliffs, New Jersey, 1992).

Statistical Process Control Statistical process control (SPC), also called statistical quality control (SQC), involves the application of statistical concepts to determine whether a process is operating satisfactorily. The ideas involved in statistical quality control are over fifty years old, but only recently with the growing worldwide focus on increased productivity have applications of SPC become widespread. If a process is operating satisfactorily (or "in control"), then the variation of product quality falls within acceptable bounds, usually the minimum and maximum values of a specified composition or property (product specification).

Figure 8-38 illustrates the typical spread of values of the controlled variable that might be expected to occur under steady-state operating conditions. The mean and root mean square (RMS) deviation are identified in Fig. 8-38 and can be computed from a series of n observations $c_1, c_2, \ldots c_n$ as follows:

$$\text{mean: } \bar{c} = \frac{1}{n} \sum_{i=1}^{n} c_i \tag{8-49}$$

RMS deviation:

$$\sigma = \left[\frac{1}{n} \sum_{i=1}^{n} (c_i - \bar{c})^2 \right]^{1/2} \tag{8-50}$$

The RMS deviation is a measure of the spread of values for c around the mean. A large value of σ indicates that wide variations in c occur. The probability that the controlled variable lies between the values of c_1 and c_2 is given by the area under the distribution between c_1 and c_2 (histogram). If the histogram follows a normal probability distribution, then 99.7 percent of all observations should lie with $\pm 3\sigma$ of the mean (between the lower and upper control limits). These limits are used to determine the quality of control.

If all data from a process lie within the $\pm 3\sigma$ limits, then we conclude that nothing unusual has happened during the recorded time period. The process environment is relatively unchanged, and the product quality lies within specification. On the other hand, if repeated violations of the $\pm 3\sigma$ limits occur, then the process environment has changed and the process is out of control.

One way to codify abnormal behavior is the so-called Western Electric rules, which identify cases where a process is out of control:

1. One point that occurs outside the upper or lower control limits
2. Any seven consecutive points lying on the same side of the center line (mean)
3. Any seven consecutive points that increase or decrease
4. Any nonrandom pattern

In the above list, one assumes that sample values are independent (i.e., not correlated).

There are important economic consequences of a process being out of control; for example, product waste and customer dissatisfaction. Hence, statistical process control does provide a way to continuously monitor process performance and improve product quality. A typical process may go out of control due to several reasons, including

- Persistent disturbances from the weather
- An undetected grade change in raw materials
- A malfunctioning instrument or control system

Statistical quality control is a diagnostic tool—that is, an indicator of quality problems—but it does not identify the source of the problem or the corrective action to be taken. The Shewhart chart provides a way to analyze variability of a single measurement, as discussed in the following example. The data in Fig. 8-39 were obtained from the monitoring of pH in a yarn-soaking kettle used in textile manufacturing (Seborg, Edgar, and Mellichamp, *Process Dynamics and Control,* Wiley, New York, 1989). Because pH has a crucial influence on color and durability of the yarn, it is important to maintain pH within a range that gives the best results for both characteristics. The pH is considered to be in control between values of 4.25 and 4.64. At the

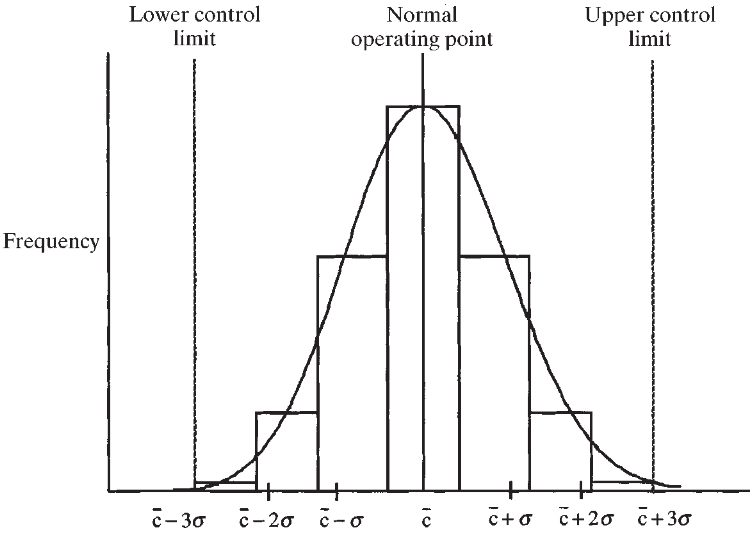


FIG. 8-38 Histogram plotting frequency of occurrence. c = mean, σ = rms deviation. Also shown is fit by normal probability distribution.

25th day, the data show that pH is out of control; this might imply that a property change in the raw material has occurred and must be corrected with the supplier. However, a real-time correction would be preferable. In Fig. 8-39, the pH was adjusted by slowly adding more acid to the vats until it came back into control (on day 29).

In continuous processes where automatic feedback control has been implemented, the feedback mechanism theoretically ensures that product quality is at or near the set point regardless of process disturbances. This, of course, requires that an appropriate manipulated variable has been identified for adjusting the product quality. However, even under feedback control, there may be daily variations of product quality because of disturbances or equipment or instrument malfunctions. These occurrences can be analyzed using the concepts of statistical quality control.

More details on statistical process control are available in several textbooks (Grant and Leavenworth, *Statistical Quality Control,* McGraw-Hill, New York, 1980; Montgomery, *Introduction to Statistical Quality Control,* Wiley, New York, 1985).

MULTIVARIABLE CONTROL PROBLEMS

General References: Shinskey, F. G., *Process Control Systems,* 3d ed., McGraw-Hill, New York, 1988. Seborg, D. E., T. F. Edgar, and D. A. Mellichamp, *Process Dynamics and Control,* Wiley, New York, 1989. McAvoy, T. J., *Interaction Analysis,* ISA, Research Triangle Park, North Carolina, 1983.

Process control books and journal articles tend to emphasize problems with a single controlled variable. In contrast, most practical problems are *multivariable* control problems because many process variables must be controlled. In fact, for virtually any important industrial process, at least two variables must be controlled: product quality and throughput. In this section, strategies for multivariable control problems are considered.

Three examples of simple multivariable control problems are shown in Fig. 8-40. The in-line blending system blends pure components *A* and *B* to produce a product stream with flow rate *w* and mass fraction of *A, x.* Adjusting either inlet flow rate w_A or w_B affects *both* of the controlled variables *w* and *x.* For the pH neutralization process in Figure 8-40(*b*), liquid level *h* and the pH of the exit stream are to be controlled by adjusting the acid and base flow rates w_a and w_b. Each of the manipulated variables affects both of the controlled variables. Thus, both the blending system and the pH neutralization process are said to exhibit strong process interactions. In contrast, the process interactions for the gas-liquid separator in Fig. 8-40(*c*) are not as strong because one manipulated variable, liquid flow rate *L,* has only a small and indirect effect on one controlled variable, pressure *P.*

Strong process interactions can cause serious problems if a conventional multiloop feedback control scheme (e.g., PI or PID controllers) is employed. The process interactions can produce undesirable control loop interactions where the controllers fight each other. Also, it may be difficult to determine the best pairing of controlled and manipulated variables. For example, in the in-line blending process in Fig. 8-40(*a*), should *w* be controlled with w_A and *x* with w_B, or vice versa?

Control Strategies for Multivariable Control Problems If a conventional multiloop control strategy performs poorly due to control loop interactions, a number of solutions are available:

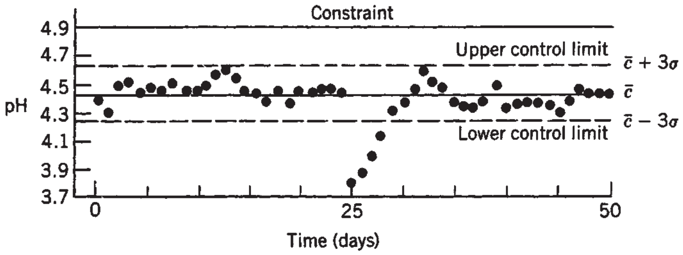
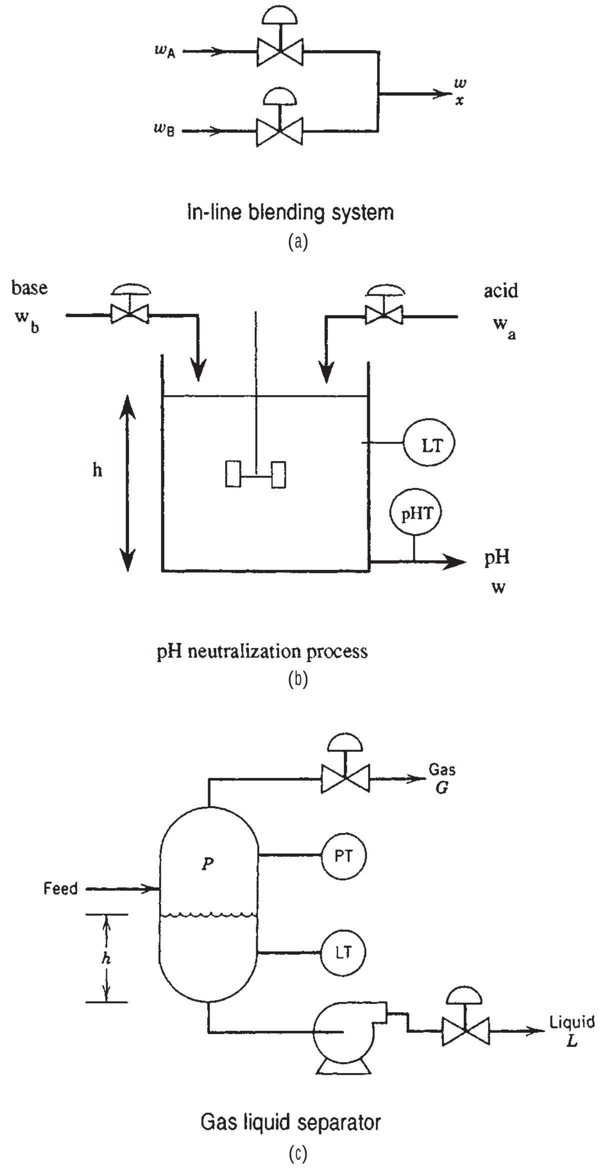


FIG. 8-39 Process control chart for the average daily pH readings.



In-line blending system
(a)

pH neutralization process
(b)

Gas liquid separator
(c)

FIG. 8-40 Physical examples of multivariable control problems.

a. Detune one or more of the control loops

b. Choose different controlled or manipulated variables (or pairings)

c. Use a decoupling control system

d. Use a multivariable control scheme (e.g., model predictive control)

Detuning a controller (e.g., using a smaller controller gain or a larger reset time) tends to reduce control loop interactions by sacrificing the performance for the detuned loops. This approach may be acceptable if some of the controlled variables are faster or less important than others.

The selection of controlled and manipulated variables is of crucial importance in designing a control system. In particular, a judicious choice may significantly reduce control loop interactions. For the blending process in Fig. 8-40(*a*), a straightforward control strategy would be to control x by adjusting w_A, and w by adjusting w_B. But physical intuition suggests that it would be better to control x by adjusting the ratio $w_A/(w_A + w_B)$ and to control product flow rate w by the sum $w_A + w_B$. Thus, the new manipulated variables would be: $M_1 = w_A/(w_A + w_B)$ and $M_2 = w_A + w_B$. In this control scheme, M_1 only affects x and M_2 only affects w. Thus, the control loop interactions have been eliminated. Similarly, for the pH neutralization process in Fig. 8-40(*b*), the control loop interactions would be greatly reduced if pH is controlled by $M_1 = w_a/(w_a + w_b)$ and liquid level h is controlled by $M_2 = w_a + w_b$.

Decoupling Control Systems Decoupling control systems provide an alternative approach for reducing control loop interactions. The basic idea is to use additional controllers called "decouplers" to compensate for undesirable process interactions.

As an illustrative example, consider the simplified block diagram for a representative decoupling control system shown in Fig. 8-41. The two controlled variables C_1 and C_2 and two manipulated variables M_1 and M_2 are related by four process transfer functions, G_{p11}, G_{p12}, and so on. For example, G_{p11} denotes the transfer function between M_1 and C_1:

$$\frac{C_1(s)}{M_1(s)} = G_{p11}(s) \tag{8-51}$$

Figure 8-41 includes two conventional feedback controllers: G_{c1} controls C_1 by manipulating M_1, and G_{c2} controls C_2 by manipulating M_2. The output signals from the feedback controllers serve as input signals to the two decouplers D_{12} and D_{21}. The block diagram is in a simplified form because the load variables and transfer functions for the final control elements and sensors have been omitted.

The function of the decouplers is to compensate for the undesirable process interactions represented by G_{p12} and G_{p21}. Suppose that the process transfer functions are all known. Then the ideal design equations are:

$$D_{12}(s) = -\frac{G_{p12}(s)}{G_{p11}(s)} \tag{8-52}$$

$$D_{21}(s) = -\frac{G_{p21}(s)}{G_{p22}(s)} \tag{8-53}$$

These decoupler design equations are very similar to the ones for feedforward control in an earlier section. In fact, decoupling can be interpreted as a type of feedforward control where the input signal is the output of a feedback controller rather than a measured load variable.

In principle, ideal decoupling eliminates control loop interactions and allows the closed-loop system to behave as a set of independent control loops. But in practice, this ideal behavior is not attained for a variety of reasons, including imperfect process models and the presence of saturation constraints on controller outputs and manipulated variables. Furthermore, the ideal decoupler design equations in (8-52) and (8-53) may not be physically realizable and thus would have to be approximated.

In practice, other types of decouplers and decoupling control configurations have been employed. For example, in *partial decoupling,* only a single decoupler is employed (i.e., either D_{12} or D_{21} in Fig. 8-41 is set equal to zero). This approach tends to be more robust than complete decoupling and is preferred when one of the controlled variables is more important than the other. *Static decouplers* can be used to reduce the steady-state interactions between control loops. They can be designed by replacing the transfer functions in Eqs. (8-52) and (8-53) with the corresponding steady-state gains,

$$D_{12}(s) = -\frac{K_{p12}}{K_{p11}} \tag{8-54}$$

$$D_{21}(s) = -\frac{K_{p21}}{K_{p22}} \tag{8-55}$$

The advantage of static decoupling is that less process information is required: namely, only steady-state gains. *Nonlinear decouplers* can be used when the process behavior is nonlinear.

Pairing of Controlled and Manipulated Variables A key decision in multiloop-control-system design is the pairing of manipu-

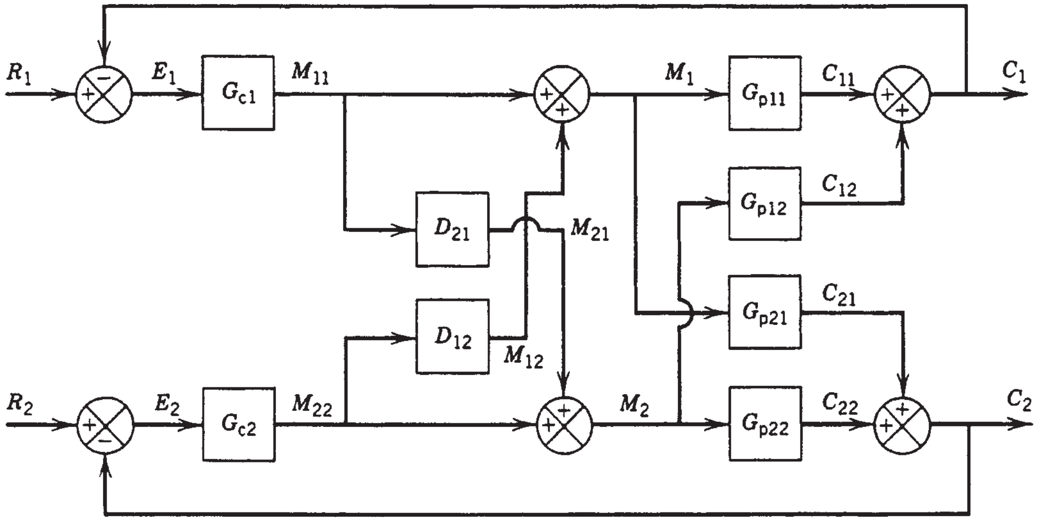


FIG. 8-41 Decoupling control system.

lated and controlled variables. This is referred to as the controller-pairing problem. Suppose there are N controlled variables and N manipulated variables. Then $N!$ distinct control configurations exist. For example, if $N = 5$, then there are 120 different multiloop control schemes. In practice, many of them would be rejected based on physical insight or previous experience. But a smaller number (e.g., 5–15) may appear to be feasible and further analysis would be warranted. Thus, it is very useful to have a simple method for choosing the most promising control configuration.

The most popular and widely used technique for determining the best controller pairing is the relative gain array (RGA) method (Bristol, "On a New Measure of Process Interaction," *IEEE Trans. Auto. Control,* AC-11, 133, 1966). The RGA method provides two important items of information:

1. A measure of the degree of process interactions between the manipulated and controlled variables
2. A recommended controller pairing

An important advantage of the RGA method is that it requires minimal process information: namely, steady-state gains. Another advantage is that the results are independent of both the physical units used and the scaling of the process variables. The chief disadvantage of the RGA method is that it neglects process dynamics, which can be an important factor in the pairing decision. Thus, the RGA analysis should be supplemented with an evaluation of process dynamics. Although extensions of the RGA method that incorporate process dynamics have been reported, these extensions have not been widely applied.

RGA Method for 2 × 2 Control Problems To illustrate the use of the RGA method, consider a control problem with two inputs and two outputs. The more general case of $N \times N$ control problems is considered elsewhere (McAvoy, *Interaction Analysis,* ISA, Research Triangle Park, North Carolina, 1983). As a starting point, it is assumed that a linear, steady-state process model is available,

$$C_1 = K_{11} M_1 + K_{12} M_2 \tag{8-56}$$

$$C_2 = K_{21} M_1 + K_{22} M_2 \tag{8-57}$$

where M_1 and M_2 are steady-state values of the manipulated inputs; C_1 and C_2 are steady-state values of the controlled outputs; and the values K are steady-state gains. The C and M variables are deviation variables from nominal steady-state values. This process model could be obtained in a variety of ways, such as by linearizing a theoretical model or by calculating steady-state gains from experimental data or a steady-state simulation.

By definition, the relative gain λ_{ij} between the ith manipulated variable and the jth controlled variable is defined as:

$$\lambda_{ij} = \frac{\text{open-loop gain between } C_i \text{ and } M_j}{\text{closed-loop gain between } C_i \text{ and } M_j} \tag{8-58}$$

where the open-loop gain is simply K_{ij} from Eqs. (8-56) and (8-57). The closed-loop gain is defined to be the steady-state gain between M_j and C_i when the other control loop is closed and no offset occurs due to the presence of integral control action. The RGA for the 2 × 2 process is denoted by

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} \tag{8-59}$$

The RGA has the important normalization property that the sum of the elements in each row and each column is exactly one. Consequently, the RGA in Eq. (8-59) can be written as

$$\Lambda = \begin{pmatrix} \lambda & 1-\lambda \\ 1-\lambda & \lambda \end{pmatrix} \tag{8-60}$$

where λ can be calculated from the following formula:

$$\lambda = \frac{1}{1 - \dfrac{K_{12} K_{21}}{K_{11} K_{22}}} \tag{8-61}$$

Ideally, the relative gains that correspond to the proposed controller pairing should have a value of one since Eq. (8-58) implies that the open and closed-loop gains are then identical. If a relative gain equals one, the steady-state operation of this loop will not be affected when the other control loop is changed from manual to automatic, or vice versa. Consequently, the recommendation for the best controller pairing is to pair the controlled and manipulated variables so that the corresponding relative gains are positive and close to one.

RGA Example In order to illustrate use of the RGA method, consider the following steady-state version of a transfer function model for a pilot-scale, methanol-water distillation column (Wood and Berry, "Terminal Composition Control of a Binary Distillation Column," *Chem. Eng. Sci.,* **28,** 1707, 1973): $K_{11} = 12.8$, $K_{12} = -18.9$, $K_{21} = 6.6$, and $K_{22} = -19.4$. It follows that $\lambda = 2$ and

$$\Lambda = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \tag{8-62}$$

Thus it is concluded that the column is fairly interacting and the recommended controller pairing is to pair C_1 with M_1 and C_2 with M_2.

MODEL PREDICTIVE CONTROL

Introduction The model-based control strategy that has been most widely applied in the process industries is *model predictive control* (MPC). It is a general method that is especially well-suited for difficult multiinput, multioutput (MIMO) control problems where there are significant interactions between the manipulated inputs and the controlled outputs. Unlike other model-based control strategies, MPC can easily accommodate inequality constraints on input and output variables such as upper and lower limits or rate-of-change limits.

A key feature of MPC is that future process behavior is predicted using a dynamic model and available measurements. The controller outputs are calculated so as to minimize the difference between the predicted process response and the desired response. At each sampling instant, the control calculations are repeated and the predictions updated based on current measurements. In typical industrial applications, the set point and target values for the MPC calculations are updated using on-line optimization based on a steady-state model of the process. Constraints on the controlled and manipulated variables can be routinely included in both the MPC and optimization calculations. The extensive MPC literature includes survey articles (Garcia, Prett, and Morari, *Automatica,* **25,** 335, 1989; Richalet, *Automatica,* **29,** 1251, 1993) and books (Prett and Garcia, *Fundamental Process Control,* Butterworths, Stoneham, Massachusetts, 1988; Soeterboek, *Predictive Control—A Unified Approach,* Prentice Hall, Englewood Cliffs, New Jersey, 1991).

The current widespread interest in MPC techniques was initiated by pioneering research performed by two industrial groups in the 1970s. Shell Oil (Houston, TX) reported their Dynamic Matrix Control (DMC) approach in 1979, while a similar technique, marketed as IDCOM, was published by a small French company, ADERSA, in 1978. Since then, there have been over one thousand applications of these and related MPC techniques in oil refineries and petrochemical plants around the world. Thus, MPC has had a substantial impact and is currently the method of choice for difficult multivariable control problems in these industries. However, relatively few applications have been reported in other process industries, even though MPC is a very general approach that is not limited to a particular industry.

Advantages and Disadvantages of MPC Model Predictive Control offers a number of important advantages:

1. It is a general control strategy for MIMO processes with inequality constraints on input and output variables.
2. It can easily accommodate difficult or unusual dynamic behavior such as large time delays and inverse responses.
3. Since the control calculations are based on optimizing control system performance, MPC can be readily integrated with on-line optimization strategies to optimize plant performance.
4. The control strategy can be easily updated on-line to compensate for changes in process conditions, constraints, or performance criteria.

But current versions of MPC have significant disadvantages:

1. The MPC strategy is very different from conventional multiloop control strategies and thus initially unfamiliar to plant personnel.
2. The MPC calculations can be relatively complicated (e.g., solving an LP or QP problem at each sampling instant) and thus require a significant amount of computer resources and effort.
3. The development of a dynamic model from plant data is time consuming, typically requiring one to three weeks of around-the-clock plant tests.
4. Since empirical models are generally used, they are only valid over the range of conditions that were considered during the plant tests.
5. Theoretical studies have demonstrated that MPC can perform poorly for some types of process disturbances, especially when output constraints are employed (Lundstrom, Lee, Morari, and Skogestad, *Computers Chem. Eng.,* **19,** 409, 1995).

Since MPC has been widely used and has had considerable impact, there is a broad consensus that its advantages far outweigh its disadvantages.

Economic Incentives for Automation Projects Industrial applications of advanced process control strategies such as MPC are motivated by the need for improvements regarding safety, product quality, environmental standards, and economic operation of the process. One view of the economics incentives for advanced automation techniques is illustrated in Fig. 8-42. Distributed control systems (DCS) are widely used for data acquisition and conventional single-loop (PID) control. Usually, they are the most expensive part of the entire control system. The addition of advanced regulatory control systems such as decouplers, selective controls, and time-delay compensation can provide additional benefits for a modest incremental cost. But experience has indicated that the major benefits can be obtained for relatively small incremental costs through a combination of MPC and on-line optimization. The results in Fig. 8-42 are shown qualitatively, rather than quantitatively, because the actual costs and benefits are application-dependent.

A key reason why MPC has become a major commercial and technical success is that there are numerous vendors who are licensed to market MPC products and install them on a turnkey basis. Consequently, even medium-sized companies are able to take advantage of this new technology. Payout times of 3–12 months have been reported.

Basic Features of MPC Model predictive control strategies have a number of distinguishing features:

1. A dynamic model of the process is used to predict the future outputs over a *prediction horizon* consisting of the next p sampling periods.
2. A reference trajectory is used to represent the desired output response over the prediction horizon.
3. Inequality constraints on the input and output variables can be included as an option.
4. At each sampling instant, a control policy consisting of the next m control moves is calculated. The control calculations are based on minimizing a quadratic or linear performance index over the prediction horizon while satisfying the constraints.
5. The performance index is expressed in terms of future control moves and the predicted deviations from the reference trajectory.
6. A *receding horizon approach* is employed. At each sampling instant, only the first control move (of the m moves that were calculated) is actually implemented. Then the predictions and control calculations are repeated at the next sampling instant.

These distinguishing features of MPC will now be described in more detail.

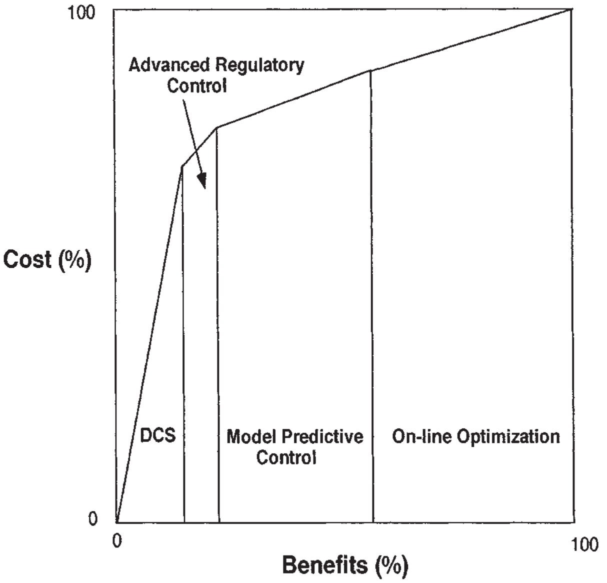


FIG. 8-42 Economic incentives for automation projects in the process industries.

A key feature of MPC is that a dynamic model of the process is used to predict future values of the controlled outputs. There is considerable flexibility concerning the choice of the dynamic model. For example, a physical model based on first principles (e.g., mass and energy balances) or an empirical model could be selected. Also, the empirical model could be a linear model (e.g., transfer function, step response model, or state space model) or a nonlinear model (e.g., neural net model). However, most industrial applications of MPC have relied on linear empirical models, which may include simple nonlinear transformations of process variables.

The original formulations of MPC (i.e., DMC and IDCOM) were based on empirical linear models expressed in either step-response or impulse-response form. For simplicity, we will consider only a single-input, single-output (SISO) model. However, the SISO model can be easily generalized to the MIMO models that are used in industrial applications. The step response model relating a single controlled variable y and a single manipulated variable u can be expressed as

$$\hat{y}(k) = \sum_{i=1}^{N} S_i \Delta u(k-i) + y(0) \tag{8-63}$$

where $\hat{y}(k)$ is the predicted value of y at the k-sampling instant; $u(k)$ is the value of the manipulated input at time k; and the model parameters S_i are referred to as the *step-response coefficients.* The initial value $y(0)$ is assumed to be known. The change in the manipulated input from one sampling instant to the next is denoted by

$$\Delta u(k) \overset{\triangle}{=} u(k) - u(k-1) \tag{8-64}$$

The step-response model is also referred to as a finite impulse response (FIR) model or a discrete convolution model.

In principle, the step-response coefficients can be determined from the output response to a step change in the input. A typical response to a unit step change in input u is shown in Fig. 8-43. The step response coefficients S_i are simply the values of the output variable at the sampling instants, after the initial value $y(0)$ has been subtracted. Theoretically, they can be determined from a single-step response, but, in practice, a number of "bump tests" are required to compensate for unanticipated disturbances, process nonlinearities, and noisy measurements.

The step-response model in Eq. (8-63) is equivalent to the following impulse response model:

$$\hat{y}(k) = \sum_{i=1}^{N} I_i u(k-i) + y(0) \tag{8-65}$$

where the impulse response coefficients I_i are related to the step-response coefficients by $I_i = S_i - S_{i-1}$. Step- and impulse-response models typically contain a large number of parameters because the model horizon N is usually quite large ($30 < N < 70$). In fact, these models are often referred to as nonparametric models. The DMC version of MPC is based on step-response models, while IDCOM utilizes impulse response models.

The receding horizon feature of MPC is shown in Fig. 8-44 with the current sampling instant denoted by k. Past input signals [$u(i)$ for $i < k$] are used to predict the output at the next p sampling instants [$\hat{y}(k+i)$ for $i = 1, 2, \ldots, p$]. The control calculations are performed to generate an m-step control policy [$u(k), u(k+1), \ldots, u(k+m)$], which optimizes the performance index. The first control action, $u(k)$, is implemented. Then at the next sampling instant $(k+1)$, the prediction and control calculations are repeated in order to determine $u(k+1)$. In Fig. 8-44, the reference trajectory (or target) is considered to be constant. Other possibilities include a gradual or step set point change that can be generated by on-line optimization.

The performance index for MPC applications is usually a linear or quadratic function of the predicted errors and calculated future control moves. For example, the following quadratic performance index has been widely used:

$$\min_{\Delta \boldsymbol{u}(k)} J = \sum_{i=1}^{p} w_i e^2(k+i) + \delta \sum_{i=1}^{m} \Delta u^2(k+i-1) \tag{8-66}$$

The value $e(k+i)$ denotes the predicted error at time $(k+i)$,

$$e(k+i) \overset{\triangle}{=} r(k+i) - \hat{y}(k+i) \tag{8-67}$$

where $r(k+i)$ is the reference value at time $k+i$, and $\Delta \boldsymbol{u}(k)$ denotes the vector of current and future control moves over the next m sampling instants:

$$\Delta \boldsymbol{u}(k) = [\Delta \boldsymbol{u}(k), \Delta \boldsymbol{u}(k+1), \ldots, \Delta \boldsymbol{u}(k+m-1)]^T \tag{8-68}$$

Equation (8-66) contains two types of design parameters that can also be used for tuning purposes. The move suppression factor δ penalizes large control moves, while the weighting factors w_i allow the predicted errors to be weighed differently at each time step, if desired.

Inequality constraints on the future inputs or their rates of change are widely used in the MPC calculations. For example, if both upper and lower limits are required, the constraints could be expressed as:

$$B_{i_*} \le u(k+i) \le B_i^{*} \qquad \text{for } i = 1, 2, \ldots, m \tag{8-69}$$

$$C_{i_*} \le \Delta u(k+i) \le C_i^{*} \qquad \text{for } i = 1, 2, \ldots, m \tag{8-70}$$

where B_i and C_i are constants. Constraints on the predicted outputs are sometimes included as well:

$$D_{i_*} \le \hat{y}(k+i) \le D_i^{*} \qquad \text{for } i = 1, 2, \ldots, p \tag{8-71}$$

The minimization of the quadratic performance index in Eq. (8-66), subject to the constraints in Eq. (8-69) to (8-71) and the step-response model in Eq. (8-63), can be formulated as a standard quadratic pro-

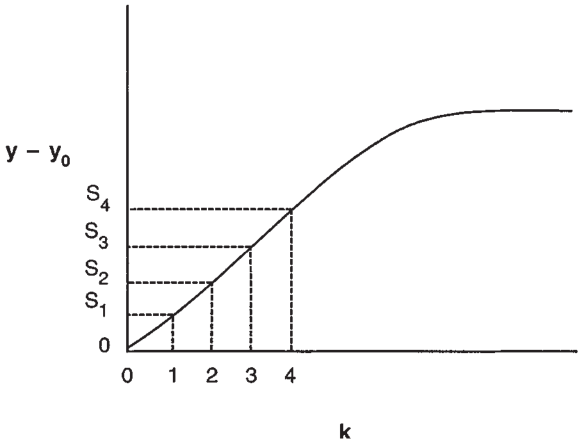


FIG. 8-43 Step response for u, a unit step change in the input.

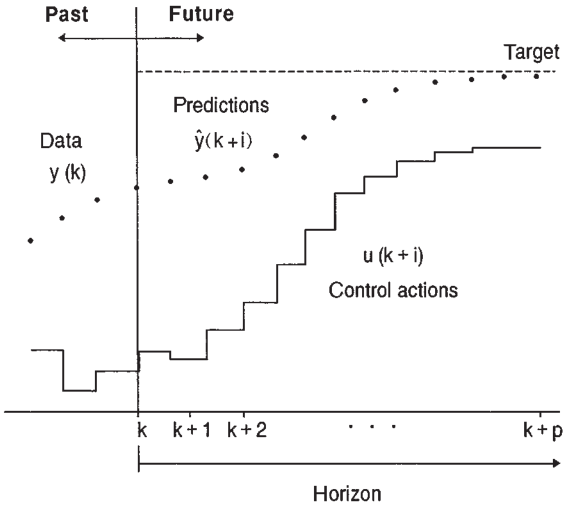


FIG. 8-44 The "moving horizon" approach of model predictive control.

gramming (QP) problem. Consequently, efficient QP solution techniques can be employed. When the inequality constraints in Eqs. (8-69) to (8-71) are omitted, the optimization problem has an analytical solution (Prett and Garcia, *Fundamental Process Control,* Butterworths, Stoneham, Massachusetts, 1988; Soeterboek, *Predictive Control—A Unified Approach,* Prentice Hall, Englewood Cliffs, New Jersey, 1991). If the quadratic terms in Eq. (8-66) are replaced by linear terms, a linear programming program (LP) problem results that can also be solved using standard methods.

This MPC formulation for SISO control problems can easily be extended to MIMO problems.

Implementation Issues A critical factor in the successful application of any model-based technique is the availability of a suitable dynamic model. In typical MPC applications, an empirical model is identified from data acquired during extensive plant tests. The experiments generally consist of a series of bump tests in the manipulated variables. Typically, the manipulated variables are adjusted one at a time and the plant tests require a period of one to three weeks. The step or impulse response coefficients are then calculated using linear-regression techniques such as least-squares methods. However, details concerning the procedures utilized in the plant tests and subsequent model identification are considered to be proprietary information. The scaling and conditioning of plant data for use in model identification and control calculations can be key factors in the success of the application.

The MPC control problem illustrated in Eqs. (8-66) to (8-71) contains a variety of design parameters: model horizon N, prediction horizon p, control horizon m, weighting factors w_i, move suppression factor δ, the constraint limits B_i, C_i, and D_i, and the sampling period Δt. Some of these parameters can be used to tune the MPC strategy, notably the move suppression factor δ, but details remain largely proprietary. One commercial controller, Honeywell's RMPCT® (Robust Multivariable Predictive Control Technology), provides default tuning parameters based on the dynamic process model and the model uncertainty.

Integration of MPC and On-Line Optimization As indicated in Fig. 8-42, significant potential benefits can be realized by using a combination of MPC and on-line optimization. At the present time, most commercial MPC packages integrate the two methodologies in a hierarchical configuration such as the one shown in Fig. 8-45. The MPC calculations are performed quite often (e.g., every 1–10 min) and implemented as set points for PID control loops at the DCS level. The targets and constraints for the MPC calculations are generated by solving a steady-state optimization problem (LP or QP) based on a linear process model. These calculations may be performed as often as the MPC calculations. As an option, the targets and constraints for the LP or QP optimization can be generated from a nonlinear process model using a nonlinear optimization technique. These calculations tend to be performed less frequently (e.g., every 1–24 hours) due to the complexity of the calculations and the process models.

The combination of MPC and frequent on-line optimization has been successfully applied in oil refineries and petrochemical plants around the world.

REAL-TIME PROCESS OPTIMIZATION

General References: Biles and Swain, *Optimization and Industrial Experimentation,* Wiley—Interscience, New York, 1980. Dantzig, *Linear Programming and Extensions,* Princeton University Press, Princeton, New Jersey, 1963. Edgar and Himmelblau, *Optimization of Chemical Processes,* McGraw-Hill, New York, 1987. Fletcher, *Practical Methods of Optimization.* Wiley, New York, 1980. Gill, Murray, and Wright, *Practical Optimization,* Academic Press, New York, 1981. Murtagh, *Advanced Linear Programming,* McGraw-Hill, New York, 1983. Murty, *Linear Programming,* Wiley, New York, 1983. Reklaitis, Ravindran, and Ragsdell, *Engineering Optimization,* Wiley—Interscience, New York, 1984.

The chemical industry has undergone significant changes during the past 20 years due to the increased cost of energy and raw materials, more stringent environmental regulations, and intense worldwide competition. Modifications of both plant-design procedures and plant operating conditions have been implemented in order to reduce costs and meet constraints. One of the most important engineering tools that can be employed in such activities is optimization. As plant computers have become more powerful, the size and complexity of problems that can be solved by optimization techniques have correspondingly expanded. A wide variety of problems in the operation and analysis of chemical plants (as well as many other industrial processes) can be solved by optimization. Real-time optimization means that the process-operating conditions (set points) are evaluated on a regular basis and optimized. Sometimes this is called steady-state optimization or supervisory control. This section examines the basic characteristics of optimization problems and their solution techniques and describes some representative benefits and applications in the chemical and petroleum industries.

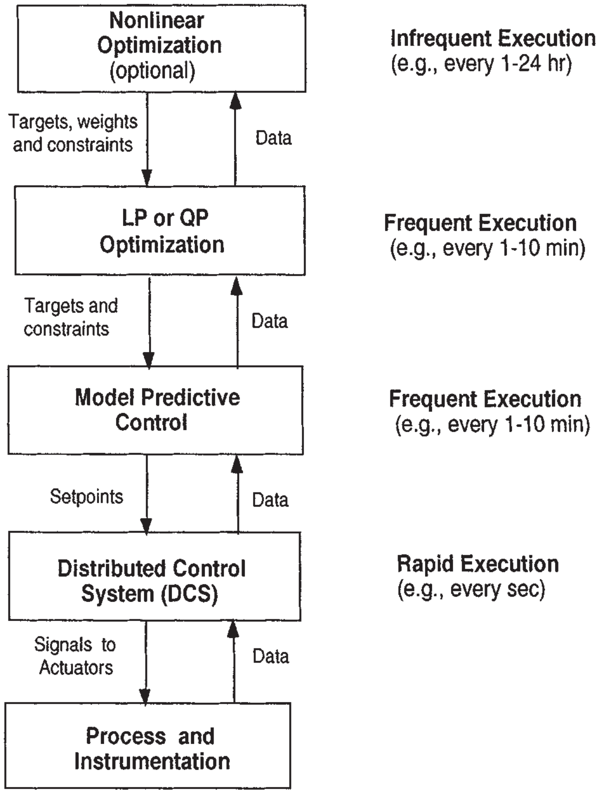

Nonlinear Optimization (optional)
Infrequent Execution (e.g., every 1-24 hr)
Targets, weights and constraints
Data
LP or QP Optimization
Frequent Execution (e.g., every 1-10 min)
Targets and constraints
Data
Model Predictive Control
Frequent Execution (e.g., every 1-10 min)
Setpoints
Data
Distributed Control System (DCS)
Rapid Execution (e.g., every sec)
Signals to Actuators
Data
Process and Instrumentation

FIG. 8-45 Hierarchical control configuration for MPC and on-line optimization.

Typical problems in chemical engineering process design or plant operation have many possible solutions. Optimization is concerned with selecting the best among the entire set by efficient quantitative methods. Computers and associated software make the computations involved in the selection feasible and cost-effective. Engineers work to improve the initial design of equipment and strive for enhancements in the operation of the equipment once it is installed in order to realize the most production, the greatest profit, the maximum cost, the least energy usage, and so on. In plant operations, benefits arise from improved plant performance, such as improved yields of valuable products (or reduced yields of contaminants), reduced energy consumption, higher processing rates, and longer times between shutdowns. Optimization can also lead to reduced maintenance costs, less equipment wear, and better staff utilization. It is helpful to systematically identify the objective, constraints, and degrees of freedom in a process or a plant if such benefits as improved quality of designs, faster and more reliable troubleshooting, and faster decision making are to be achieved.

Optimization can take place at many levels in a company, ranging from a complex combination of plants and distribution facilities down through individual plants, combinations of units, individual pieces of equipment, subsystems in a piece of equipment, or even smaller entities. Problems that can be solved by optimization can be found at all these levels.

While process design and equipment specification are usually performed prior to the implementation of the process, optimization of operating conditions is carried out monthly, weekly, daily, hourly, or even every minute. Optimization of plant operations determines the set points for each unit at the temperatures, pressures, and flow rates that are the best in some sense. For example, the selection of the percentage of excess air in a process heater is quite critical and involves a balance on the fuel-air ratio to assure complete combustion and at the same time make the maximum use of the heating potential of the fuel. Typical day-to-day optimization in a plant minimizes steam consumption or cooling water consumption, optimizes the reflux ratio in a distillation column, or allocates raw materials on an economic basis [Latour, *Hydro Proc.,* **58**(6), 73, 1979, and *Hydro. Proc.,* **58**(7), 219, 1979].

A real-time optimization (RTO) system determines set point changes and implements them via the computer control system without intervention from unit operators. The RTO system completes all data transfer, optimization calculations, and set point implementation before unit conditions change and invalidate the computed optimum. In addition, the RTO system should perform all tasks without upsetting plant operations. Several steps are necessary for implementation of RTO, including determination of the plant steady state, data gathering and validation, updating of model parameters (if necessary) to match current operations, calculation of the new (optimized) set points, and the implementation of these set points.

To determine if a process unit is at steady state, a program monitors key plant measurements (e.g., compositions, product rates, feed rates, and so on) and determines if the plant is steady enough to start the sequence. Only when all of the key measurements are within the allowable tolerances is the plant considered steady and the optimization sequence started. Tolerances for each measurement can be tuned separately. Measured data are then collected by the optimization computer. The optimization system runs a program to screen the measurements for unreasonable data (gross error detection). This validity checking automatically modifies the model updating calculation to reflect any bad data or when equipment is taken out of service. Data validation and reconciliation (on-line or off-line) is an extremely critical part of any optimization system.

The optimization system then may run a parameter-fitting case that updates model parameters to match current plant operation. The integrated process model calculates such items as exchanger heat transfer coefficients, reactor performance parameters, furnace efficiencies, and heat and material balances for the entire plant. Parameter fitting allows for continual updating of the model to account for plant deviations and degradation of process equipment. After completion of the parameter fitting, the information regarding the current plant constraints, the control status data, and the economic values for feed products, utilities, and other operating costs are collected. The economic values are updated by the planning and scheduling department on a regular basis. The optimization system then calculates the optimized set points. The steady-state condition of the plant is rechecked after the optimization case is successfully completed. If the plant is still steady, then the values of the optimization targets are transferred to the process-control system for implementation. After a line-out period, the process-control computer resumes the steady-state detection calculations, restarting the cycle.

Essential Features of Optimization Problems The solution of optimization problems involves the use of various tools of mathematics. Consequently, the formulation of an optimization problem requires the use of mathematical expressions. From a practical viewpoint, it is important to mesh properly the problem statement with the anticipated solution technique. Every optimization problem contains three essential categories:

1. An objective function to be optimized (revenue function, cost function, etc.)
2. Equality constraints (equations)
3. Inequality constraints (inequalities)

Categories 2 and 3 comprise the model of the process or equipment; category 1 is sometimes called the economic model.

No single method or algorithm of optimization exists that can be applied efficiently to all problems. The method chosen for any particular case will depend primarily on (1) the character of the objective function, (2) the nature of the constraints, and (3) the number of independent and dependent variables. Table 8-6 summarizes the six general steps for the analysis and solution of optimization problems (Edgar and Himmelblau, *Optimization of Chemical Processes,* McGraw-Hill, New York, 1988). You do not have to follow the cited order exactly, but you should cover all of the steps eventually. Shortcuts in the procedure are allowable, and the easy steps can be performed first. Steps 1, 2, and 3 deal with the mathematical definition of the problem: identification of variables and specification of the objective function and statement of the constraints. If the process to be optimized is very complex, it may be necessary to reformulate the problem so that it can be solved with reasonable effort. Later in this section, we discuss the development of mathematical models for the process and the objective function (the economic model).

Step 5 in Table 8-6 involves the computation of the optimum point. Quite a few techniques exist to obtain the optimal solution for a problem. We describe several classes of methods below; Fig. 8-46 is a diagram for selection of individual optimization techniques. In general, the solution of most optimization problems involves the use of a digital computer to obtain numerical answers. Over the past 15 years, substantial progress has been made in developing efficient and robust digital methods for optimization calculations. Much is known about which methods are most successful. Virtually all numerical optimization methods involve iteration, and the effectiveness of a given technique often depends on a good first guess for the values of the variables at the optimal solution. After the optimum is computed, a sensitivity analysis for the objective function value should be performed to determine the effects of errors or uncertainty in the objective function, mathematical model, or other constraints.

Development of Process (Mathematical) Models Constraints in optimization problems arise from physical bounds on the variables, empirical relations, physical laws, and so on. The mathematical relations describing the process also comprise constraints. Two general categories of models exist:

1. Those based on physical theory
2. Those based on strictly empirical descriptions

Mathematical models based on physical and chemical laws (e.g., mass and energy balances, thermodynamics, chemical reaction kinetics) are frequently employed in optimization applications. These models are conceptually attractive because a general model for any system size can be developed before the system is constructed. On the other hand, an empirical model can be devised that simply correlates input-output data without any physiochemical analysis of the process. For

TABLE 8-6 The Six Steps Used to Solve Optimization Problems

1. Analyze the process itself so that the process variables and specific characteristics of interest are defined (i.e., make a list of all of the variables).
2. Determine the criterion for optimization and specify the objective function in terms of the above variables together with coefficients. This step provides the performance model (sometimes called the economic model when appropriate).
3. Develop via mathematical expressions a valid process or equipment model that relates the input-output variables of the process and associated coefficients. Include both equality and inequality constraints. Use well-known physical principles (mass balances, energy balances), empirical relations, implicit concepts, and external restrictions. Identify the independent and dependent variables (number of degrees of freedom).
4. If the problem formulation is too large in scope, (a) break it up into manageable parts and/or (b) simplify the objective function and model.
5. Apply a suitable optimization technique to the mathematical statement of the problem.
6. Check the answers and examine the sensitivity of the result to changes in the coefficients in the problem and the assumptions.

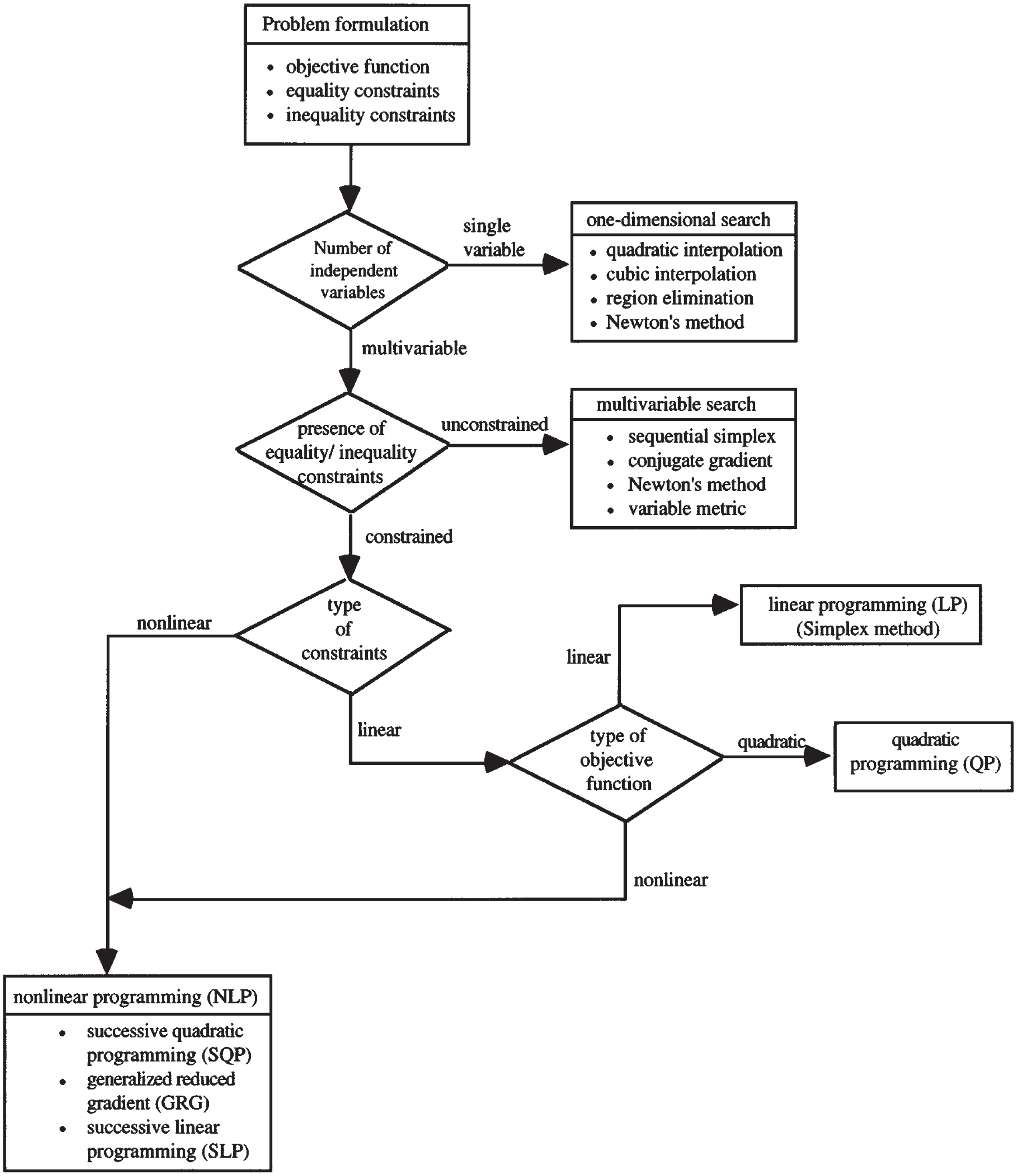


FIG. 8-46 Diagram for selection of optimization techniques with algebraic constraints and objective function.

these models, optimization is often used to fit a model to process data, using a procedure called parameter estimation. The well-known least squares curve-fitting procedure is based on optimization theory, assuming that the model parameters are contained linearly in the model. One example is the yield matrix, where the percentage yield of each product in a unit operation is estimated for each feed component using process data rather than employing a mechanistic set of chemical reactions.

Formulation of the Objective Function The formulation of objective functions is one of the crucial steps in the application of optimization to a practical problem. You must be able to translate the desired objective into mathematical terms. In the chemical process industries, the objective function often is expressed in units of currency (e.g., U.S. dollars) because the normal industrial goal is to minimize costs or maximize profits subject to a variety of constraints.

A typical economic model involves the costs of raw materials, values of products, costs of production, as functions of operating conditions, projected sales figures, and the like. An objective function can be expressed in terms of these quantities; for example, annual operating profit ($/yr) might be expressed as:

$$J = \sum_s F_s V_s - \sum_r F_r C_r - \text{OC} \tag{8-72}$$

where J = profit/time

$\sum_s F_s V_s$ = sum of product flow rates times respective product values (income)

$\sum_r F_r C_r$ = sum of feed flows times respective unit costs

OC = operating costs/time

Unconstrained Optimization Unconstrained optimization refers to the case where no inequality constraints are present and all equality constraints can be eliminated by solving for selected dependent variables followed by substitution for them in the objective function. Very few realistic problems in process optimization are unconstrained. However, it is desirable to have efficient unconstrained optimization techniques available since these techniques must be applied in real time and iterative calculations cost computer time. The two classes of unconstrained techniques are single-variable optimization and multivariable optimization.

Single Variable Optimization Many process optimization problems can be reduced to the variation of a single variable so as to maximize profit or some other overall process objective function. Some examples of single variable optimization include optimizing the reflux ratio in a distillation column or the air/fuel ratio in a furnace (Martin, Latour, and Richard, *Chem. Engr. Prog.,* **77,** September, 1981). While most processes actually are multivariable processes with several operating degrees of freedom, often we choose to optimize only the most important variable in order to keep the strategy uncomplicated. One characteristic implicitly required in a single variable optimization problem is that the objective function J be unimodal in the variable x.

The selection of a method for one-dimensional search is based on the trade-off between number of function evaluations versus computer time. We can find the optimum by evaluating the objective function at many values of x using a small grid spacing (Δx) over the allowable range of x values, but this method is generally inefficient. There are three classes of techniques that can be used efficiently for one-dimensional search:

1. Indirect
2. Region elimination
3. Interpolation

Indirect methods seek to solve the necessary condition $dJ/dx = 0$ by iteration, but these methods are not as popular as the second two classes. Region elimination methods include equal interval search, dichotomous search (or bisecting), Fibonacci search, and golden section. These methods do not use information on the shape of the function (other than being unimodal) and thus tend to be rather conservative. The third class of techniques uses repeated polynomial fitting to predict the optimum. These interpolation methods tend to converge rapidly to the optimum without being very complicated. Two interpolation methods, quadratic and cubic interpolation, have been used in many optimization packages.

Multivariable Optimization In multivariable optimization problems, there is no guarantee that the optimum can be reached in a reasonable amount of computer time. The numerical optimization of general nonlinear multivariable objective functions requires that efficient and robust techniques be employed. Efficiency is important since iteration is employed. For example, in multivariable "grid" search for a problem with four independent variables, an equally spaced grid for each variable is prescribed. For ten values of each of the four variables, there would be 10^4 total function evaluations required to find the best answer for the grid intersections. However, this computational effort still may not yield a result close enough to the true optimum, requiring further search. Therefore, grid search is a very inefficient method for most problems involving many variables.

In multivariable optimization, the difficulty of dealing with multivariable functions is usually resolved by treating the problem as a series of one-dimensional searches. For a given starting point, a search direction **s** is specified, and the optimum is found by searching along that direction. The step size ε is the distance moved along **s.** Then a new search direction is determined, followed by another one-dimensional search. The algorithm used to specify the search direction depends on the optimization method.

There are two basic types of unconstrained optimization algorithms: (1) those requiring function derivatives and (2) those that do not. The nonderivative methods are of interest in optimization applications because these methods can be readily adapted to the case in which experiments are carried out directly on the process. In such cases, an actual process measurement (such as yield) can be the objective function, and no mathematical model for the process is required. Methods that do not require derivatives are called direct methods and include sequential simplex (Nelder-Meade) and Powell's method. The sequential simplex method is quite satisfactory for optimization with two or three independent variables, is simple to understand, and is fairly easy to execute. Powell's method is more efficient than the simplex method and is based on the concept of conjugate search directions.

The second class of multivariable optimization techniques in principle requires the use of partial derivatives, although finite difference formulas can be substituted for derivatives; such techniques are called indirect methods and include the following classes:

1. Steepest descent (gradient) method
2. Conjugate gradient (Fletcher-Reeves) method
3. Newton's method
4. Quasi-Newton methods

The steepest descent method is quite old and utilizes the intuitive concept of moving in the direction where the objective function changes the most. However, it is clearly not as efficient as the other three. Conjugate gradient utilizes only first-derivative information, as does steepest descent, but generates improved search directions. Newton's method requires second derivative information but is very efficient, while quasi-Newton retains most of the benefits of Newton's method but utilizes only first derivative information. All of these techniques are also used with constrained optimization.

Constrained Optimization When constraints exist and cannot be eliminated in an optimization problem, more general methods must be employed than those described above, since the unconstrained optimum may correspond to unrealistic values of the operating variables. The general form of a nonlinear programming problem allows for a nonlinear objective function and nonlinear constraints, or

$$\begin{aligned} &\text{Minimize} && J(x_1, x_2, \ldots, x_n) && \\ &\text{Subject to} && h_i(x_1, x_2, \ldots, x_n) = 0 && (i = 1, r_c) \\ & && g_i(x_1, x_2, \ldots, x_n) \geq 0 && (i = 1, m_c) \end{aligned} \tag{8-73}$$

In this case, there are n process variables with r_c equality constraints and m_c inequality constraints. Such problems pose a serious challenge to performing optimization calculations in a reasonable amount of time. Typical constraints in chemical process optimization include operating conditions (temperatures, pressures, and flows have limits), storage capacities, and product purity specifications.

An important class of constrained optimization problems is one in which both the objective function and constraints are linear. The solution of these problems is highly structured and can be obtained rapidly. The accepted procedure, linear programming (LP), has become quite popular in the past twenty years, solving a wide range of industrial problems. It is increasingly being used for on-line optimization. For processing plants, there are several different kinds of linear constraints that may arise, making the LP method of great utility.

1. Production limitation due to equipment throughput restrictions, storage limits, or market constraints.
2. Raw material (feedstock) limitation.
3. Safety restrictions on allowable operating temperatures and pressures.
4. Physical property specifications placed on the composition of the final product. For blends of various products, we usually assume that a composite property can be calculated through the averaging of pure component physical properties.

5. Material and energy balances of the steady-state model.

The optimum in linear programming lies at the constraint intersections, which was generalized to any number of variables and constraints by George Dantzig. The Simplex algorithm is a matrix-based numerical procedure for which many digital computer codes exist, both for mainframe and microcomputers (Edgar and Himmelblau, *Optimization of Chemical Processes,* McGraw-Hill, New York, 1987; Schrage, *Linear, Integer, and Quadratic Programming with LINDO,* Scientific Press, Palo Alto, California, 1983). The algorithm can handle virtually any number of inequality constraints and any number of variables in the objective function and utilizes the observation that only the constraint boundaries need to be examined to find the optimum. In some instances, nonlinear optimization problems even with nonlinear constraints can be linearized so that the LP algorithm can be employed to solve them (called successive linear programming or SLP). In the process industries, the Simplex algorithm has been applied to a wide range of problems, including refinery scheduling, olefins production, the optimal allocation of boiler fuel, and the optimization of a total plant.

Nonlinear Programming The most general case for optimization occurs when both the objective function and constraints are nonlinear, a case referred to as nonlinear programming. While the idea behind the search methods used for unconstrained multivariable problems are applicable, the presence of constraints complicates the solution procedure.

In practice, one of the best current general algorithms (best on the basis of many tests) using iterative linearization is the Generalized Reduced Gradient algorithm (GRG). The GRG algorithm employs linear or linearized constraints, defines new variables that are normal to the constraints, and expresses the gradient (or other search direction) in terms of this normal basis (Liebman, Lasdon, Schrage, and Waren, *GINO,* Scientific Press, Palo Alto, California, 1986). Other established types of constrained optimization methods include the following types of algorithms:

1. Penalty functions with augmented Lagrangian method (an enhancement of the classical Lagrange multiplier method)
2. Successive quadratic programming

All of these methods have been utilized to solve nonlinear programming problems in the field of chemical engineering design and operations (Lasdon and Waren, *Oper. Res.,* **5,** 34, 1980). Nonlinear programming is receiving increased usage in the area of real-time optimization.

One important class of nonlinear programming techniques is called quadratic programming (QP), where the objective function is quadratic and the constraints are linear. While the solution is iterative, it can be obtained quickly as in linear programming. This is the basis for the newest type of constrained multivariable control algorithms called model predictive control. The dominant method used in the refining industry utilizes the solution of a QP and is called dynamic matrix control or DMC. See the earlier subsection on model predictive control for more details.

EXPERT SYSTEMS

An expert system is a computer program that uses an expert's knowledge in a particular domain to solve a narrowly focused, complex problem. An off-line system uses information entered manually and produces results in visual form to guide the user in solving the problem at hand. An on-line system uses information taken directly from process measurements to perform tasks automatically or instruct or alert operating personnel to the status of the plant.

Each expert system has a rule base created by the expert to respond the way the expert would to sets of input information. Expert systems used for plant diagnostics and management usually have an open rule base, which can be changed and augmented as more experience accumulates and more tasks are to be automated. The system begins as an empty shell with an assortment of functions such as equation-solving, logic, and simulation, as well as input and display tools to allow an expert to construct a proprietary rule base. The "expert" in this case would be the person or persons having the deepest knowledge about the process, its problems, its symptoms, and remedies. Converting these inputs into meaningful outputs is the principal task in constructing a rule base. Skill at computer programming is especially helpful, although most shells allow rules to be entered in the vernacular. First-principles models (deep knowledge) produce the most accurate results, although heuristics are always required to establish limits.

A closed expert system is one designed by an expert to be sold in quantity for use by others (where open systems tend to be unique). It is closed to keep users from altering the rule base and thereby changing the product. Common examples in process control are autotuning and self-tuning controllers whose rule base is designed by one or more experts in that field. Once packaged and sold, its rule base cannot be changed in the field no matter how poorly it performs the task; revisions must be made by the manufacturer in later releases as for any software product.

The development vehicle used to create and test the rule base must be as flexible as possible, allowing easy alterations and expansion of the rule base with whatever displays can convey the most information. The delivery vehicle, however, should be virtually transparent to the user, conveying only as much information as needed to solve the problem at hand. Self-tuning controllers can perform their task without explicitly informing users, but their output and status is available on demand, and their operation may be easily limited or interrupted.

To be successful, the scope of an expert system must be limited to a narrow group of common problems that are readily solved by conventional means, and where the return on investment is greatest. Widening the scope usually requires more complex methods and treats less common problems having lower return.

UNIT OPERATIONS CONTROL

PROCESS AND INSTRUMENTATION DIAGRAMS

General References: Shinskey, *Process Control Systems,* 3d ed., McGraw-Hill, New York, 1988. Luyben, *Practical Distillation Control,* Van Nostrand Reinhold, New York, 1992.

The process and instrumentation (P&I) diagram provides a graphical representation of the control configuration for the process. The P&I diagrams illustrate the measurement devices that provide inputs to the control strategy, the actuators that will implement the results of the control calculations, and the function blocks that provide the control logic.

The symbology for drawing P&I diagrams generally follows standards developed by one of the following organizations:

1. International Society for Measurement and Control (ISA). The chemicals, refining, and foods industries generally follow this standard.
2. Scientific Apparatus Manufacturers Association (SAMA). The fossil-fuel electric utility industry generally follows this standard.

Both organizations update their standards from time to time, primarily because the continuing evolutions in control-system hardware provide additional possibilities for implementing control schemes.

Although arguments can be made for the advantages of each symbology, the practices within an industry seem to be mainly the result of historical practice with no indication of any significant shift. Most companies adopt one of the standards but then tailor or extend the symbology to best suit their internal practices. Such companies maintain an internal document and/or drawing that specifies the symbology used on their P&I diagrams. Their internal personnel and all contractors are instructed to adhere to this symbology when developing P&I diagrams.

Figure 8-47 presents a P&I diagram for a simple temperature control loop that adheres to the ISA symbology. The measurement

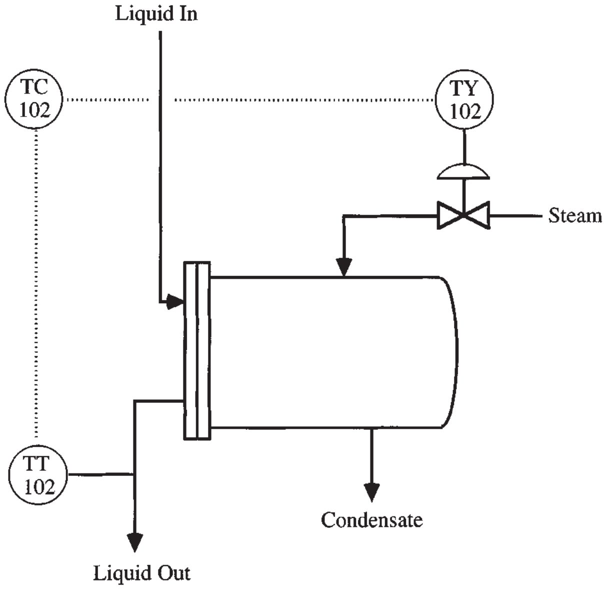


FIG. 8-47 Example of a process and instrument diagram.

devices and most elements of the control logic are represented by circles. In Figure 8-47, circles are used to designate the following:

1. TT102 is the temperature measurement device.
2. TC102 is the temperature controller.
3. TY102 is the current-to-pneumatic (I/P) transducer.

The symbol for the control valve in Fig. 8-47 is for a pneumatic positioning valve without a valve positioner.

Electronic signals (that is, 4–20 milliamp current loops) are represented by dashed lines. In Fig. 8-47, these include the following:

1. The signal from the measurement device to the controller.
2. The signal from the controller to the I/P transducer.

Pneumatic signals are represented by solid lines that are cross-hatched intermittently. The signal from the I/P transducer to the pneumatic positioning valve is pneumatic.

The ISA symbology provides different symbols for different types of actuators. Furthermore, variations for the controller symbol distinguish control algorithms implemented in DCS technology from panel-mounted single-loop controllers.

CONTROL OF HEAT EXCHANGERS

Steam-Heated Exchangers Steam, the most common heating medium, transfers its latent heat in condensing, causing heat flow to be proportional to steam flow. Thus, a measurement of steam flow is essentially a measure of heat transfer. Consider raising a liquid from temperature T_1 to T_2 by condensing steam:

$$Q = WH = FC_L(T_2 - T_1) \tag{8-74}$$

where W and H are the mass flow of steam and its latent heat, F and C_L are the mass flow and specific heat of the liquid, and Q is the rate of heat transfer. The response of controlled temperature to steam flow is linear:

$$\frac{dT_2}{dW} = \frac{H}{FC_L} \tag{8-75}$$

However, the steady-state process gain described by this derivative varies inversely with liquid flow: Adding a given increment of heat flow to a smaller flow of liquid produces a greater temperature rise.

Dynamically, the response of liquid temperature to a step in steam flow is that of a distributed lag, shown in Fig. 8-48. The time required to reach 63 percent complete response, $\sum \tau$, is essentially the residence time of the fluid in the exchanger, which is its volume divided by its flow. The residence time then varies inversely with flow. Table 8-2 gives optimum settings for PI and PID controllers for distributed lags, the proportional band varying directly with steady-state gain, and integral and derivative settings directly with $\sum \tau$. Since both these parameters vary inversely with liquid flow, fixed settings for the temperature controller are optimal at only one flow rate.

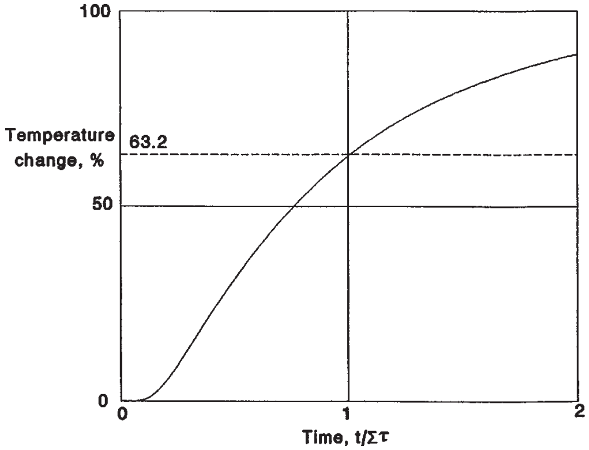


FIG. 8-48 Temperature leaving a heat exchanger responds as a distributed lag, the gain and time constant of which vary inversely with flow.

Undamped oscillations will be produced when the flow decreases by one-third from the value at which the controller was optimally tuned, whereas increasing flow rates produces an overdamped response. The stable operating range can be broadened to one-half the original flow by using an equal-percentage steam valve whose gain varies directly with flow. The best solution is to adapt the PID settings to change inversely with measured flow, thereby keeping the controller optimally tuned for all flow rates.

Feedforward control can also be applied by multiplying the liquid flow measurement—after dynamic compensation—by the output of the temperature controller, the result used to set steam flow in cascade. Feedforward is capable of a reduction in integrated error as much as a hundredfold but requires the use of a steam-flow loop and dynamic compensator to approach this.

Steam flow is sometimes controlled by manipulating a valve in the condensate line rather than the steam line, because it is smaller and hence less costly. Heat transfer, then, is changed by raising or lowering the level of condensate flooding the heat-transfer surface, an operation that is slower than manipulating a steam valve. Protection also needs to be provided against an open condensate valve blowing steam into the condensate system.

Exchange of Sensible Heat When there is no change in phase, heat transfer is no longer linear with flow of the manipulated stream, as illustrated by Fig. 8-49. Here again, an equal-percentage valve should be used on that stream to linearize the loop. The variable dynamics of the distributed lag apply, limiting the stable operating range in the same way as for the steam-heated exchanger. These heat exchangers are also sensitive to variations in the temperature of the manipulated stream, an increasingly common problem where heat is being recovered at variable temperatures for reuse in heat transfer.

Figure 8-50 shows a temperature controller (TC) setting a heat-flow controller (QC) in cascade. A measurement of the manipulated flow is multiplied by its temperature difference across the heat exchanger to calculate the current heat-transfer rate, using the right side of Eq. (8-74). Variations in supply temperature, then, appear as variations in calculated heat transfer, which the QC can quickly correct by adjusting the manipulated flow. An equal-percentage valve is still required to linearize the secondary loop, but the primary loop of temperature-setting heat flow is linear. Feedforward can be added by multiplying the dynamically compensated flow measurement of the other fluid by the output of the temperature controller.

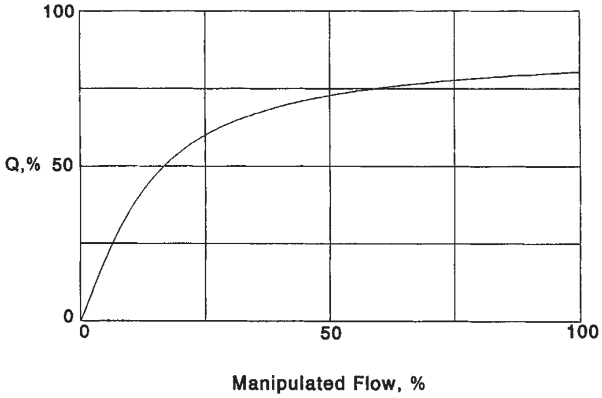


FIG. 8-49 Heat-transfer rate in sensible-heat exchange varies nonlinearly with flow of the manipulated fluid.

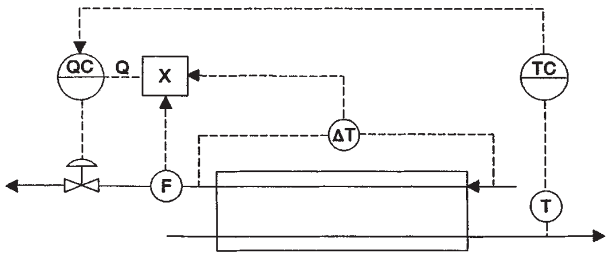


FIG. 8-50 Manipulating heat flow linearizes the loop and protects against variations in supply temperature.

When manipulating a stream whose flow is independently determined, such as flow of a product or a heat-transfer fluid from a fired heater, a three-way valve is used to divert the required flow to the heat exchanger. This does not alter the linearity of the process or its sensitivity to supply variations and even adds the possibility of independent flow variations. The three-way valve should have equal-percentage characteristics, and heat-flow control may be even more beneficial.

DISTILLATION COLUMN CONTROL

Distillation columns have four or more closed loops—increasing with the number of product streams and their specifications—all of which interact with each other to some extent. Because of this interaction, there are many possible ways to pair manipulated and controlled variables through controllers and other mathematical functions with widely differing degrees of effectiveness. Columns also differ from each other, so that no single rule of configuring control loops can be applied successfully to all. The following rules apply to the most common separations.

Controlling Quality of a Single Product If one of the products of a column is far more valuable than the others, its quality should be controlled to satisfy given specifications, and its recovery should be maximized by minimizing losses of its principal component in other streams. This is achieved by maximizing reflux ratio consistent with flooding limits on trays, which means maximizing the flow of reflux or vapor, whichever is limiting. The same rule should be followed when heating and cooling have little value. A typical example is the separation of high-purity propylene from much lower-valued propane, usually achieved with waste heat from quench water from the cracking reactors.

The most important factor affecting product quality is the material balance. In separating a feed stream F into distillate D and bottom B products, an overall mole-flow balance must be maintained:

$$F = D + B \tag{8-76}$$

as well as a balance on each component:

$$Fz_i = Dy_i + Bx_i \tag{8-77}$$

where z, y, and x are mol fractions of component i in the respective streams. Combining these equations gives a relationship between the composition of the products and their relative portion of the feed:

$$\frac{D}{F} = 1 - \frac{B}{F} = \frac{z_i - x_i}{y_i - x_i} \tag{8-78}$$

From the above, it can be seen that control of either x_i or y_i requires both product flow rates to change with feed rate and feed composition.

Figure 8-51 shows a propylene-propane fractionator controlled at maximum boilup by the differential pressure controller (DPC) across the trays. This loop is fast enough to reject upsets in the temperature of the quench water quite easily. Pressure is controlled by manipulating the heat-transfer surface in the condenser through flooding. If the condenser should become overloaded, pressure will rise above set point, but this has no significant effect on the other control loops. Temperature measurements on this column are not helpful, as the difference between the component boiling points is too small. Propane content in the propylene distillate is measured by a chromatographic analyzer sampling the overhead vapor for fast response and is controlled by the analyzer controller (AC) manipulating the ratio of distillate to feed rates. The feedforward signal from feed rate is dynamically compensated by $f(t)$ and nonlinearly characterized by $f(x)$ to account for variations in propylene recovery as feed rate changes. Distillate flow can be measured and controlled more accurately than reflux flow by a factor equal to the reflux ratio—in this column, typically between 10 and 20. Therefore, reflux flow is placed under accumulator level control (LC). Yet composition responds to the difference between boilup and reflux. To eliminate the lag inherent in the response of the level controller, reflux flow is driven by the subtractor in the direction opposite to distillate flow—this is essential to fast response of the composition loop.

Controlling Quality of Two Products Where the two products have similar value, or where heating and cooling costs are comparable to product losses, the compositions of both products should be controlled. This introduces the possibility of strong interaction between the two composition loops, as they tend to have the same speed of response. To minimize interaction, most columns should have distillate composition controlled by reflux ratio and bottom composition by boilup or preferably boilup-to-bottom ratio. These loops are insensitive to variations in feed rate, eliminating the need for feedforward control, and they also reject heat-balance upsets quite effectively. Figure 8-52 shows a depropanizer controlled by reflux and boilup ratios. The actual mechanism through which these ratios are manipulated is as $D/(L + D)$ and $B/(V + B)$, where L is reflux flow and V is vapor boilup, which decouples the temperature loops from the liquid-level loops. Column pressure here is controlled by flooding both condenser and accumulator; however, there is no LC on the accumulator, so this arrangement will not function with an overloaded condenser. Temperatures are used as indications of composition in this column because of the substantial difference in boiling points between propane and butanes. However, off-key components such as ethane do effect the accuracy of the relationship so that an analyzer controller is used to set the top temperature controller (TC) in cascade.

If the products from a column are especially pure, even this configuration may produce excessive interaction between the composition loops. Then the composition of the less pure product should be controlled by manipulating its own flow; the composition of the remaining product should be controlled by manipulating reflux ratio if it is the distillate or boilup ratio if it is the bottom product.

CHEMICAL REACTORS

Composition Control The first requirement for successful control of a chemical reactor is to establish the proper stoichiometry, that is, to control the flow rates of the reactants in the proportions needed

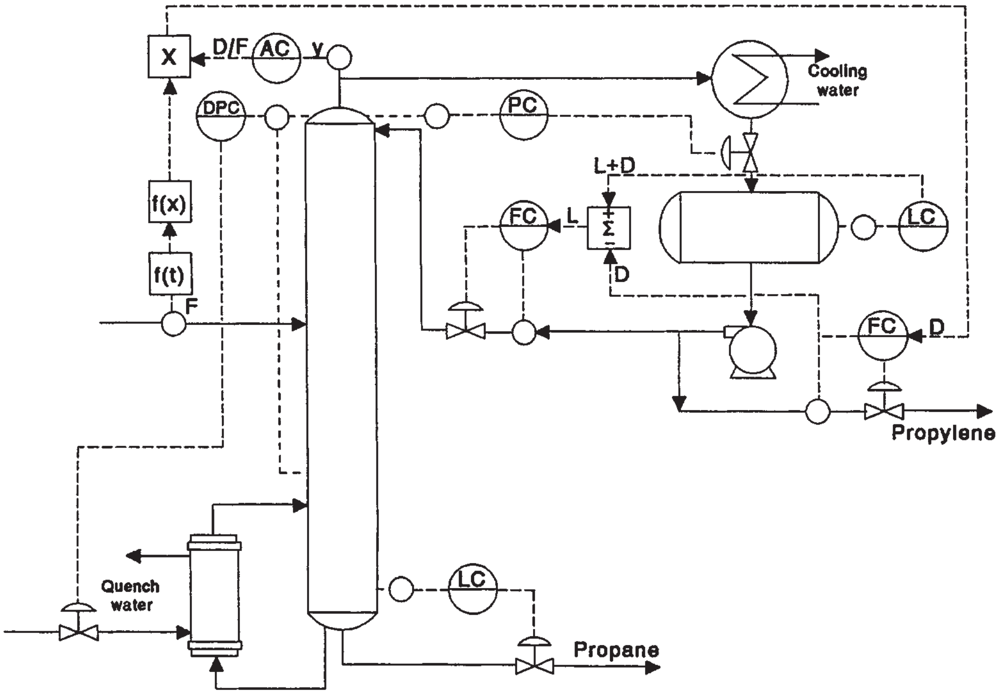


FIG. 8-51 The quality of high-purity propylene should be controlled by manipulating the material balance.

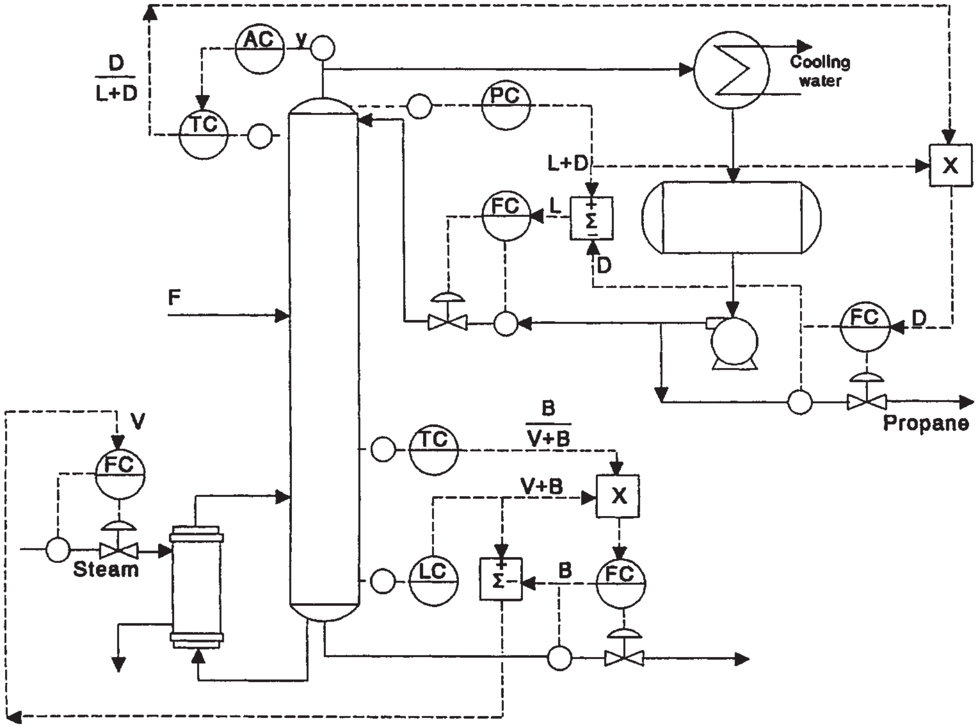


FIG. 8-52 Depropanizers require control of both products, here using reflux-ratio and boilup-ratio manipulation.

to satisfy the reaction chemistry. In a continuous reactor, this begins by setting ingredient flow rates in ratio to one another. However, because of variations in the purity of the feed streams and inaccuracy in flow metering, some indication of excess reactant such as pH or a composition measurement should be used to trim the ratios. Many reactions are incomplete, leaving one or more reactants unconverted. They are separated from the products of the reaction and recycled to the reactor, usually contaminated with inert components. While reactants can be recycled to complete conversion (extinction), inerts can accumulate to the point of impeding the reaction and must be purged from the system. Inerts include noncondensible gases that must be vented and nonvolatiles from which volatile products must be stripped.

If one of the reactants differs in phase from the others and the products, it may be manipulated to close the material balance on that phase. For example, a gas reacting with liquids to produce a liquid product may be added, as it is consumed to control reactor pressure; a gaseous purge would be necessary. Similarly, a liquid reacting with a gas to produce a gaseous product could be added, as it is consumed to control liquid level in the reactor; a liquid purge would be required. Where a large excess of one reactant *A* is used to minimize side reactions, the unreacted excess is typically sent to a storage tank for recycling. Its flow from the recycle storage tank is set in the desired ratio to the flow of reactant *B*, with the flow of fresh *A* manipulated to control recycle tank level if the feed is a liquid or tank pressure if it is a gas. Some catalysts travel with the reactants and must be recycled in the same way.

With batch reactors, it may be possible to add all reactants in their proper quantities initially if the reaction rate can be controlled by injection of initiator or adjustment of temperature. In semibatch operation, one key ingredient is flow-controlled into the batch at a rate that sets the production. This ingredient should not be manipulated for temperature control of an exothermic reactor, as the loop includes two dominant lags—concentration of the reactant and heat capacity of the reaction mass—and can easily go unstable.

Temperature Control Reactor temperature should always be controlled by heat transfer. Endothermic reactions require heat and therefore are eminently self-regulating. Exothermic reactions produce heat, which tends to raise reaction temperature, thereby increasing reaction rate and producing more heat. This positive feedback is countered by negative feedback in the cooling system, which removes more heat as reactor temperature rises. Most continuous reactors have enough heat-transfer surface relative to reaction mass so that negative feedback dominates and they are self-regulating. But most batch reactors do not and are therefore steady-state unstable. Unstable reactors are controllable, but the temperature controller requires a high gain, and the cooling system must have enough margin to accommodate the largest expected disturbance in heat load.

Figure 8-53 shows the recommended system for controlling the temperature of an exothermic reactor, either continuous or batch. The circulating pump on the coolant loop is absolutely essential to effective temperature control in keeping dead time minimum and constant—without it, dead time varies inversely with cooling load, causing limit cycling at low loads. Heating is usually required to raise the temperature to reaction conditions, although it is often locked out in a batch reactor once initiator is introduced. The valves are operated in split range, the heating valve opening from 50–100 percent of controller output, and the cooling valve opening from 0–50 percent. The cascade system linearizes the reactor temperature loop, speeds its response, and protects it from disturbances in the cooling system. The flow of heat removed per unit of coolant flow is directly proportional to the temperature rise of the coolant, which varies with both the temperature of the reactor and the rate of heat transfer from it. Using an equal-percentage cooling valve helps compensate for this nonlinearity, although it is incomplete—a preferred arrangement would be to manipulate coolant flow using a heat-flow controller as described in Fig. 8-50.

The flow of heat across the heat-transfer surface is linear with both temperatures, leaving the primary loop with a constant gain. Using the coolant exit rather than inlet temperature as the secondary controlled variable moves the jacket dynamics from the primary to the secondary loop, reducing the period of the primary loop. Performance and robustness are both improved by using the secondary temperature measurement as the feedback signal to the integral mode of the primary controller. (This feature may only be available with controllers that integrate by positive feedback.) This places the entire secondary loop in the integral path of the primary controller, effectively pacing its integral time to the rate at which the secondary temperature is able to respond. The primary controller may also be left in the automatic mode at all times without integral windup.

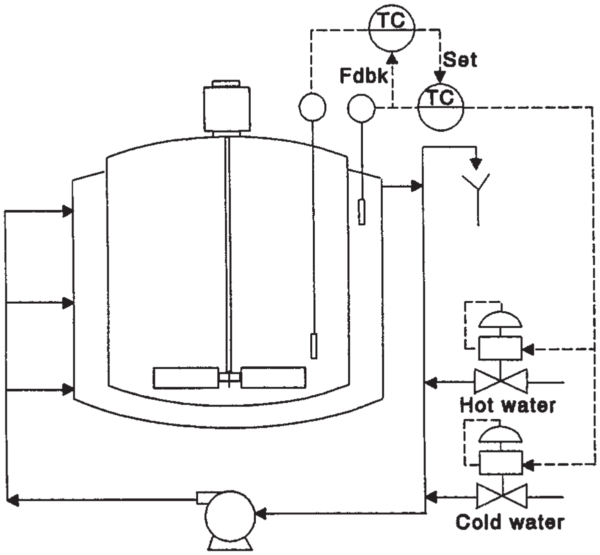


FIG. 8-53 The reactor temperature controller sets coolant outlet temperature in cascade, with primary integral feedback taken from the secondary temperature measurement.

The primary time constant of the reactor is

$$\tau_1 = \frac{M_r C_r}{UA} \tag{8-79}$$

where M_r and C_r are the mass and heat capacity of the reactants, and *U* and *A* are the overall heat-transfer coefficient and area respectively. This system was tested on a pilot reactor where the heat-transfer area and mass could both be changed by a factor of two, changing τ_1 by a factor of four as confirmed by observations of rates of temperature rise. Yet the controllers configured as described in Fig. 8-53 did not require retuning as τ_1 varied. The primary controller should be PID, and the secondary controller at least PI in this system; if the secondary controller has no integral mode, the primary controller will control with offset. Set point overshoot in batch reactor control can be avoided by setting derivative time of the primary controller higher than its integral time, but this is only effective with interacting PID controllers.

CONTROLLING EVAPORATORS

The most important consideration in controlling the quality of concentrate from an evaporator is forcing the vapor rate to match the flow of excess solvent entering in the feed. The mass flow of solid material entering and leaving are equal in the steady state:

$$M_0 x_0 = M_n x_n \tag{8-80}$$

where M_0 and x_0 are the mass flow and solid fraction of the feed, and M_n and x_n are their values in the product after *n* effects of evaporation. The total solvent evaporated from all the effects must then be

$$\sum W = M_0 - M_n = M_0\left(1 - \frac{x_0}{x_n}\right) \tag{8-81}$$

For a steam-heated evaporator, each unit of steam W_0 applied produces a known amount of evaporation based on the number of effects and their fractional economy E:

$$\sum W = nEW_0 \tag{8-82}$$

(A comparable statement can be made with regard to the power applied to a mechanical recompression evaporator.) In summary, the steam flow required to increase the solid content of the feed from x_0 to x_n is

$$W_0 = \frac{M_0(1 - x_0/x_n)}{nE} \tag{8-83}$$

The usual measuring device for feed flow is a magnetic flowmeter, which is a volumetric device whose output F must be multiplied by density ρ to produce mass flow M_0. For most aqueous solutions which are fed to evaporators, the product of density and the function of solid content appearing above is linear with density:

$$F\rho\left(1 - \frac{x_0}{x_n}\right) \approx F[1 - m(\rho - 1)] \tag{8-84}$$

where slope m is determined by the desired product concentration, and density is in g/ml. The required steam flow in lb/h for feed measured in gal/min is then

$$W_0 = \frac{500F[1 - m(\rho - 1)]}{nE} \tag{8-85}$$

where the factor of 500 converts gal/min of water to lb/h. The factor nE is about 1.74 for a double-effect evaporator and 2.74 for a triple-effect. Using a thermocompressor (ejector) driven with 150-lb/in^2 steam on a single-effect evaporator gives an nE of 2.05; it essentially adds the equivalent of one effect to the evaporator train.

A cocurrent evaporator train with its controls is illustrated in Fig. 8-54. The control system applies equally well to countercurrent or mixed-feed evaporators, the principal difference being the tuning of the dynamic compensator $f(t)$, which must be done in the field to minimize the short-term effects of changes in feed flow on product quality. Solid concentration in the product is usually measured as density; feedback trim is applied by the AC adjusting slope m of the density function, which is the only term related to x_n. This recalibrates the system whenever x_n must move to a new set point.

The accuracy of the system depends on controlling heat flow; therefore, if steam pressure varies, compensation must be applied to correct for both steam density and enthalpy as a function of pressure. Some evaporators must use unreliable sources of low-pressure steam. In this case, the measurement of pressure-compensated steam flow can be used to set feed flow by solving the last equation for F using W_0 as a variable. The steam-flow controller would be set for a given production rate, but the dynamically compensated steam-flow measurement would be the input signal to calculate the feed-flow set point. Both of these configurations are widely used in controlling corn-syrup concentrators.

DRYING OPERATIONS

Controlling dryers is much different than controlling evaporators because on-line measurements of feed rate and composition and product composition are rarely available. Most dryers transfer moisture from wet feed into hot dry air in a single pass. The process is generally very self-regulating, in that moisture becomes progressively harder to remove from the product as it dries: This is known as falling-rate drying. Controlling the temperature of the air leaving a cocurrent dryer tends to regulate the moisture in the product, as long as feed rate and the moisture in the feed and air are reasonably constant. At constant outlet air temperature, product moisture tends to rise with all three of these variables.

In the absence of moisture analyzers, regulation of product quality can be improved by raising the temperature of the exhaust air in proportion to the evaporative load. The evaporative load can be estimated by the loss in temperature of the air passing through the dryer in the steady state. Changes in load are first observed in upsets in exhaust temperature at a given inlet temperature; the controller then responds by returning the exhaust air to its original temperature by changing that of the inlet air.

Figure 8-55 illustrates the simplest application of this principal as the linear relationship

$$T_0 = T_b + K\Delta T \tag{8-86}$$

where T_0 is the set point for exhaust temperature elevated above a base temperature T_b corresponding to zero-load operation, and ΔT is the drop in air temperature from inlet to outlet. Coefficient K must be set to regulate product moisture over the expected range of evaporative load. If set too low, product moisture will increase with increasing load; if set too high, it will decrease with increasing load. While K can be estimated from the model of a dryer, it does depend on the rate-of-

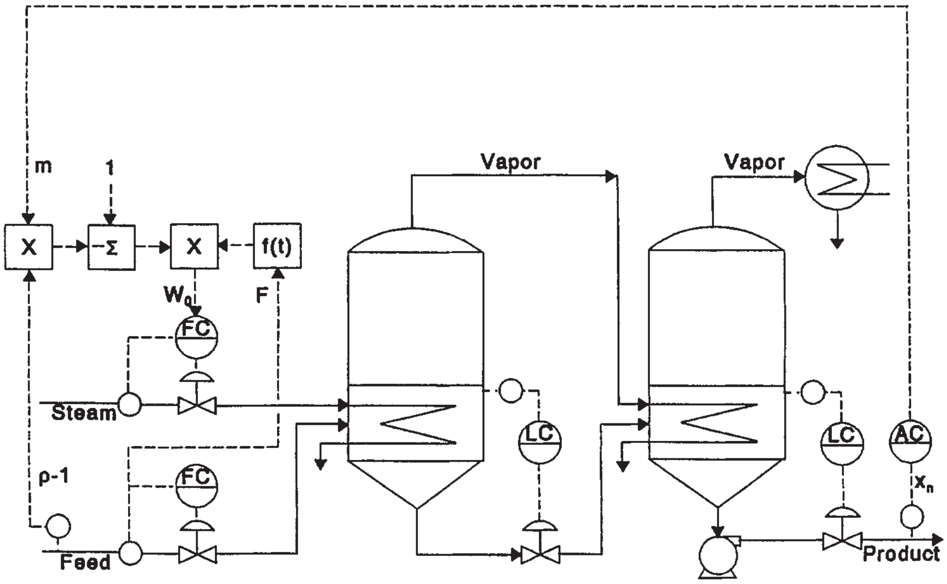


FIG. 8-54 Controlling evaporators requires matching steam flow and evaporative load, here using feedforward control.

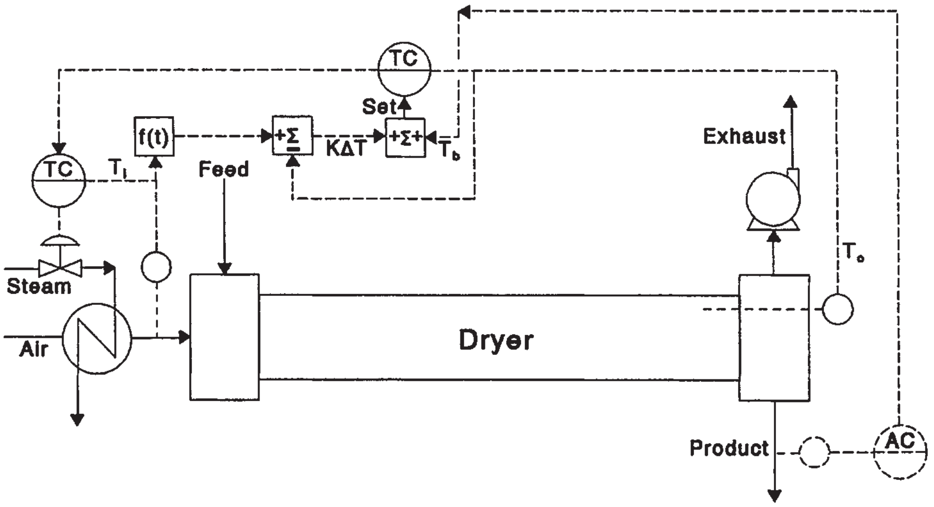


FIG. 8-55 Product moisture from a cocurrent dryer can be regulated through temperature control indexed to heat load.

drying curve for the product, its particle size, and whether the load variations are due primarily to changes in feed rate or feed moisture.

It is important to have the most accurate measurement of exhaust temperature attainable. Note that Fig. 8-55 shows the sensor inserted into the dryer upstream of the rotating seal, because leakage there could cause the temperature in the exhaust duct to read low—even lower than the wet-bulb temperature, an impossibility without leakage of either heat or outside air.

The calculation of exhaust-temperature set point forms a positive-feedback loop capable of destabilizing the dryer. For example, an increase in load causes the controller to raise inlet temperature, which will in turn raise the calculated set point calling for a further increase in inlet temperature. The gain in the set point loop, K, typically is well below the gain of the exhaust temperature measurement responding to the same change in inlet temperature. Negative feedback then dominates in the steady state, but the response of the exhaust temperature measurement is delayed by the dryer. A similar lag $f(t)$ is shown inserted in the set point loop to prevent positive feedback from dominating in the short term, which could cause cycling.

If product moisture is measured off-line, analytical results can be used to adjust K and T_b manually. If an on-line analyzer is used, the analyzer controller would be most effective in adjusting the bias T_b, as shown in the figure.

While the rotary dryer shown is commonly used for grains and minerals, this system has been successfully applied to fluid-bed drying of plastic pellets, air-lift drying of wood fibers, and spray drying of milk solids. The air may be steam-heated as shown or heated by direct combustion of fuel, provided that a representative measurement of inlet air temperature can be made. If it cannot, then evaporative load can be inferred from a measurement of fuel flow, replacing ΔT in the set point calculation.

If the feed flows countercurrent to the air, as is the case when drying granulated sugar, exhaust temperature does not respond to variations in product moisture. For these dryers, product moisture can better be regulated by controlling its temperature at the point of discharge. Conveyor-type dryers are usually divided into a number of zones, each separately heated with recirculation of air which raises its wet-bulb temperature. Only the last two zones may require indexing of exhaust-air temperature as a function of ΔT.

Batch drying, used on small lots like pharmaceuticals, begins operation by blowing air at constant inlet temperature through saturated product in constant-rate drying, where ΔT is constant at its maximum value ΔT_c. When product moisture reaches the point where falling-rate drying begins, the exhaust temperature begins to rise. The desired product moisture will be reached at a corresponding exhaust temperature T_f, which is related to the temperature T_c observed during constant-rate drying, as well as ΔT_c:

$$T_f = T_c + K\Delta T_c \tag{8-87}$$

The control system requires the values of T_c and ΔT_c observed during the first minutes of operation to be stored as the basis for the above calculation of end point. When the exhaust temperature then reaches the value calculated, drying is terminated. Coefficient K can be estimated from models but requires adjustment on-line to reach product specifications repeatedly. Products having different moisture specifications or particle size will require different settings of K, but the system does compensate for variations in feed moisture, batch size, air moisture, and inlet temperature. Some exhaust air may be recirculated to control the dewpoint of the inlet air, thereby conserving energy toward the end of the batch and when the ambient air is especially dry.

BATCH PROCESS CONTROL

BATCH VERSUS CONTINUOUS PROCESSES

General References: Fisher, *Batch Control Systems: Design, Application, and Implementation,* ISA, Research Triangle Park, North Carolina, 1990; Rosenof and Ghosh, *Batch Process Automation,* Van Nostrand Reinhold, New York, 1987.

When categorizing process plants, the following two extremes can be identified:

1. *Commodity plants.* These plants are custom-designed to produce large amounts of a single product (or a primary product plus one or more secondary products). An example is a chlorine plant, where the primary product is chlorine and the secondary products are hydrogen and sodium hydroxide. Usually the margins (product value less manufacturing costs) for the products from commodity plants are small, so the plants must be designed and operated for best possible efficiencies. Although a few are batch, most commodity plants are

continuous. Factors such as energy costs are life-and-death issues for such plants.

2. *Specialty plants.* These plants are capable of producing small amounts of a variety of products. Such plants are common in fine chemicals, pharmaceuticals, foods, and so on. In specialty plants, the margins are usually high, so factors such as energy costs are important but not life-and-death issues. As the production amounts are relatively small, it is not economically feasible to dedicate processing equipment to the manufacture of only one product. Instead, batch processing is utilized so that several products (perhaps hundreds) can be manufactured with the same process equipment. The key issue in such plants is to manufacture consistently each product in accordance with its specifications.

The above two categories represent the extremes in process configurations. The term *semibatch* designates plants in which some processing is continuous but other processing is batch. Even processes that are considered to be continuous can have a modest amount of batch processing. For example, the reformer unit within a refinery is thought of as a continuous process, but the catalyst regeneration is normally a batch process.

In a continuous process, the conditions within the process are largely the same from one day to the next. Variations in feed composition, plant utilities (e.g., cooling water temperature), catalyst activities, and other variables occur, but normally these changes are either about an average (e.g., feed compositions) or exhibit a gradual change over an extended period of time (e.g., catalyst activities). Summary data such as hourly averages, daily averages, and the like are meaningful in a continuous process.

In a batch process, the conditions within the process are continually changing. The technology for making a given product is contained in the product recipe that is specific to that product. Such recipes normally state the following:

1. *Raw material amounts.* This is the stuff needed to make the product.
2. *Processing instructions.* This is what must be done with the stuff in order to make the desired product.

This concept of a recipe is quite consistent with the recipes found in cookbooks.

Sometimes the term *recipe* is used to designate only the raw material amounts and other parameters to be used in manufacturing a batch. Although appropriate for some batch processes, this concept is far too restrictive for others. For some products, the differences from one product to the next are largely physical as opposed to chemical. For such products, the processing instructions are especially important. The term *formula* is more appropriate for the raw material amounts and other parameters, with *recipe* designating the formula and the processing instructions.

The above concept of a recipe permits the following three different categories of batch processes to be identified:

1. *Cyclical batch.* Both the formula and the processing instructions are the same from batch to batch. Batch operations within processes that are primarily continuous often fall into this category. The catalyst regenerator within a reformer unit is a cyclical batch process.
2. *Multigrade.* The processing instructions are the same from batch to batch, but the formula can be changed to produce modest variations in the product. In a batch PVC plant, the different grades of PVC are manufactured by changing the formula. In a batch pulp digester, the processing of each batch or *cook* is the same, but at the start of each cook, the process operator is permitted to change the formula values for chemical-to-wood ratios, cook time, cook temperature, and so on.
3. *Flexible batch.* Both the formula and the processing instructions can change from batch to batch. Emulsion polymerization reactors are a good example of a flexible batch facility. The recipe for each product must detail both the raw materials required and how conditions within the reactor must be sequenced in order to make the desired product.

Of these, the flexible batch is by far the most difficult to automate and requires a far more sophisticated control system than either the cyclical batch or the multigrade batch facility.

Batches and Recipes Each batch of product is manufactured in accordance with a product recipe, which contains all information (formula and processing instructions) required to make a batch of the product (see Fig. 8-56). For each batch of product, there will be one and only one product recipe. However, a given product recipe is nor-

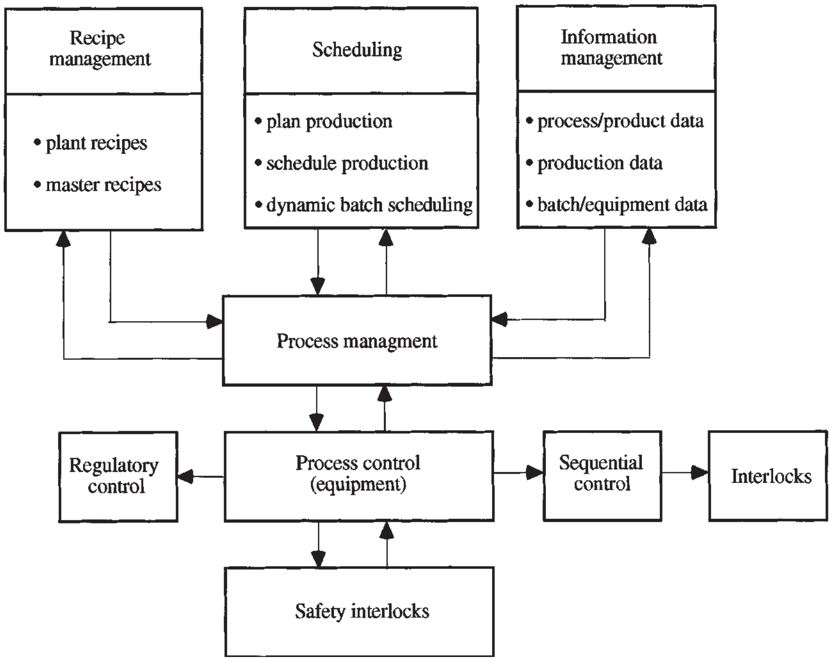


FIG. 8-56 Batch control overview.

mally used to make several batches of product. To uniquely identify a batch of product, each batch is assigned a unique identifier called the batch ID. Most companies adopt a convention for generating the batch ID, but this convention varies from one company to the next.

In most batch facilities, more than one batch of product will be in some stage of production at any given time. The batches in progress may or may not be using the same recipe. The maximum number of batches that can be in progress at any given time is a function of the equipment configuration for the plant.

The existence of multiple batches in progress at a given time presents numerous opportunities for the process operator to make errors, such as charging a material to the wrong batch. Charging a material to the wrong batch is almost always detrimental to the batch to which the material is incorrectly charged. Unless this error is recognized quickly so that the proper charge can be made, the error is also detrimental to the batch to which the charge was supposed to be made. Such errors usually lead to an off-specification batch, but the consequences could be more serious and result in a hazardous condition.

Recipe management refers to the assumption of such duties by the control system. Each batch of product is tracked throughout its production, which may involve multiple processing operations on various pieces of processing equipment. Recipe management assures that all actions specified in the product recipe are performed on each batch of product made in accordance with that recipe. As the batch proceeds from one piece of processing equipment to the next, recipe management is also responsible for assuring that the proper type of process equipment is used and that this processing equipment is not currently in use by another batch.

By assuming such responsibilities, the control system greatly reduces the incidences where operator error results in off-specification batches. Such a reduction in error is essential to implement just-in-time production practices, where each batch of product is manufactured at the last possible moment. When a batch (or batches) are made today for shipment by overnight truck, there is insufficient time for producing another batch to make up for an off-specification batch.

Routing and Production Monitoring In some facilities, batches are individually scheduled. However, in most facilities, production is scheduled by product runs, where a run is the production of a stated quantity of a given product. From the stated quantity and the standard yield of each batch, the number of batches can be determined. As this is normally more than one batch of product, a production run is normally a sequence of some number of batches of the same product.

In executing a production run, the following issues must be addressed (see Fig. 8-56):

1. *Processing equipment must be dedicated to making the run.* More than one run is normally in progress at a given time. The maximum number of runs simultaneously in progress depends on the equipment configuration of the plant. Routing involves determining which processing equipment will be used for each production run.
2. *Raw material must be utilized.* When a production run is scheduled, the necessary raw materials must be allocated to the production run. As the individual batches proceed, the consumption of raw materials must be monitored for consistency with the allocation of raw materials to the production run.
3. *The production quantity for the run must be achieved by executing the appropriate number of batches.* The number of batches is determined from a standard yield for each batch. However, some batches may achieve yields higher than the standard yield, but other batches may achieve yields lower than the standard yield. The actual yields from each batch must be monitored and significant deviations from the expected yields must be communicated to those responsible for scheduling production.

The last two activities are key components of production monitoring, although production monitoring may also involve other activities such as tracking equipment utilization.

Production Scheduling In this regard, it is important to distinguish between scheduling runs (sometimes called long-term scheduling) and assigning equipment to runs (sometimes called routing or short-term scheduling). As used herein, production scheduling refers to scheduling runs and is usually a corporate-level as opposed to a plant-level function. Short-term scheduling or routing was previously discussed and is implemented at the plant level.

The long-term scheduling is basically a material resources planning (MRP) activity involving the following:

1. *Forecasting.* Orders for long-delivery raw materials are issued at the corporate level based on the forecast for the demand for products. The current inventory of such raw materials is also maintained at the corporate level. This constitutes the resources from which products can be manufactured.
2. *The orders for products.* Orders are normally received at the corporate level and then assigned to individual plants for production and shipment. Although the scheduling of some products is based on required product inventory levels, scheduling based on orders and shipping directly to the customer (usually referred to as just-in-time) avoids the costs associated with maintaining product inventories.
3. *Plant locations and capacities.* While producing a product at the nearest plant usually lowers transportation costs, plant capacity limitations sometimes dictate otherwise.

Any company competing in the world economy needs the flexibility to accept orders on a worldwide basis and then assign them to individual plants to be filled. Such a function is logically implemented within the corporate-level information technology framework.

BATCH AUTOMATION FUNCTIONS

Automating a batch facility requires a spectrum of functions.

Interlocks Some of these are provided for safety and are properly called safety interlocks. However, others are provided to avoid mistakes in processing the batch and are properly called process interlocks.

Discrete Device States Discrete devices such as two-position valves can be driven to either of two possible states. Such devices can be optionally outfitted with limit switches that indicate the state of the device. For two-position valves, the following combinations are possible:

1. No limit switches
2. One limit switch on the closed position
3. One limit switch on the open position
4. Two limit switches

In process-control terminology, the discrete device driver is the software routine that generates the output to a discrete device such as a valve and also monitors the state feedback information to ascertain that the discrete device actually attains the desired state. Given the variety of discrete devices used in batch facilities, this logic must include a variety of capabilities. For example, valves do not instantly change states, but instead each valve exhibits a travel time for the change from one state to another. To accommodate this characteristic of the field device, the processing logic within the discrete device driver must provide for a user-specified transition time for each field device. When equipped with limit switches, the potential states for a valve are as follows:

1. *Open.* The valve has been commanded to open, and the limit switch inputs are consistent with the open state.
2. *Closed.* The valve has been commanded to close, and the limit switch inputs are consistent with the closed state.
3. *Transition.* This is a temporary state that is only possible after the valve has been commanded to change state. The limit switch inputs are not consistent with the commanded state, but the transition time has not expired.
4. *Invalid.* The transition time has expired, and the limit switch inputs are not consistent with the commanded state for the valve.

The invalid state is an abnormal condition that is generally handled in a manner similar to process alarms. The transition state is not considered to be an abnormal state but may be implemented in either of the following ways:

1. *Drive and wait.* Further actions are delayed until the device attains its commanded state.
2. *Drive and proceed.* Further actions are initiated while the device is in the transition state.

The latter is generally necessary for devices with long travel times, such as flush-fitting reactor discharge valves that are motor-driven.

Closing such valves is normally via drive and wait; however, drive and proceed is usually appropriate when opening the valve.

Although two-state devices are most common, the need occasionally arises for devices with three or more states. For example, an agitator may be on high speed, on slow speed, or off.

Process States Batch processing usually involves imposing the proper sequence of states on the process. For example, a simple blending sequence might be as follows:

1. Transfer specified amount of material from tank *A* to tank *R.* The process state is "Transfer from *A.*"
2. Transfer specified amount of material from tank *B* to tank *R.* The process state is "Transfer from *B.*"
3. Agitate for specified period of time. The process state is "Agitate without cooling."
4. Cool (with agitation) to specified target temperature. The process state is "Agitate with cooling."

For each process state, the various discrete devices are expected to be in a specified device state. For process state "Transfer from *A,*" the device states might be as follows:

1. Tank *A* discharge valve: open.
2. Tank *R* inlet valve: open.
3. Tank *A* transfer pump: running.
4. Tank *R* agitator: off.
5. Tank *R* cooling valve: closed.

For many batch processes, process state representations are a very convenient mechanism for representing the batch logic. A grid or table can be constructed, with the process states as rows and the discrete device states as columns (or vice versa). For each process state, the state of every discrete device is specified to be one of the following:

1. Device state 0, which may be valve closed, agitator off, and so on
2. Device state 1, which may be valve open, agitator on, and so on
3. No change or don't care

This representation is easily understandable by those knowledgeable about the process technology and is a convenient mechanism for conveying the process requirements to the control engineers responsible for implementing the batch logic.

Many batch software packages also recognize process states. A configuration tool is provided to define a process state. With such a mechanism, the batch logic does not need to drive individual devices but can simply command that the desired process state be achieved. The system software then drives the discrete devices to the device states required for the target process state. This normally includes the following:

1. Generating the necessary commands to drive each device to its proper state
2. Monitoring the transition status of each device to determine when all devices have attained their proper states
3. Continuing to monitor the state of each device to assure that the devices remain in their proper states

Should any discrete device not remain in its target state, failure logic must be initiated.

We will use the control of a simple mixing process (Fig. 8-57) to demonstrate various batch control strategies found in commercial systems. To start the operation sequence, a solenoid valve (VN7) is opened to introduce liquid *A.* When the liquid level in the tank reaches an intermediate level (LH2), flow *B* is started to turn on the mixer. When the liquid level is high (LXH2), flow *B* is stopped and the discharge valve is opened (VN9). The discharge valve is closed and the motor stopped when the tank level reaches the low limit (LL2). The operator may start another mixing cycle by depressing the start button again. It should be noted that this simplified control strategy does not deal with emergency process conditions. Timing of equipment sequencing, such as making sure valve 8 is closed before opening the discharge valve, is not considered. However, this example fully demonstrates the device interlocking and signal latching often encountered in sequential process control.

This process is event triggered and can be easily programmed using sequential logic [Figure 8-58*a*]. Many PLC implementations start the programming phase with sequential logic design. Gate 1 ensures that the process will not start, when requested, if the tank level is not low. Gate 3 opens valve 7 for flow *A* only if valve 8 is not opened. Gate 2 latches the operator request once valve 7 is opened such that the operator may release the push button. Gate 4 starts flow *B* and the mixer motor when the intermediate level is reached. The start signal is fed into gate 3 to terminate flow *A.* At the high tank level, gate 6 opens the discharge valve. This signal is fed into gate 4 to stop flow *B* and the mixer motor. Gate 5 latches in the discharge signal until the tank is drained. Note that for a DCS, this sequential logic can be entered entirely as Boolean functional blocks.

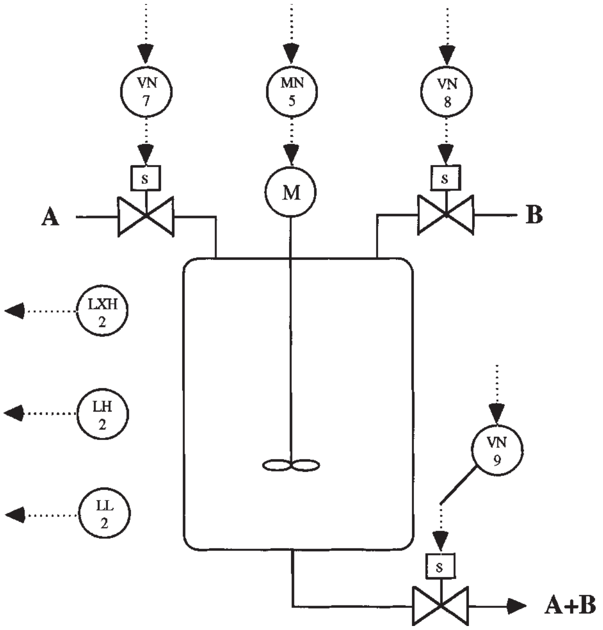


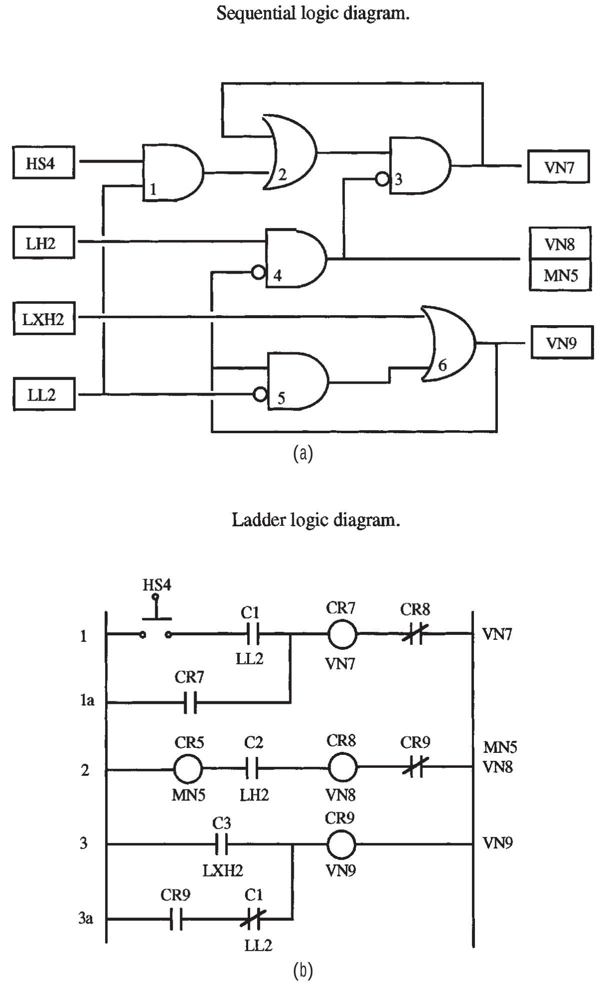
FIG. 8-57 Process schematics of a mixing tank.

Figure 8-58*b* is the ladder logic diagram for the same mixing process. It involves rungs of parallel circuits containing relays (the circles) and contacts. Parallel bars on the rungs represent contacts. A slashed pair of bars depict a normally closed contact. A normally open momentary contact is shown on rung 1 in Fig. 8-58*b.* The ladder logic and diagram builder in PLCs can be programmed easily because there are only a limited number of symbols required in ladder logic diagrams.

The translation from sequential logic to ladder logic is straightforward. In general, two or more contacts on the same rung forms an AND gate. Contacts on branches of a rung form an OR gate. For example, contact C1 on rung 1 is normally open, unless the tank level is low. Contact CR8 is normally closed unless relay CR8 on rung 2 is energized. An operator-actuated push button, HS4, and the contact C1 forms an AND gate equivalent to gate 1 in Fig. 8-58*a.* Therefore, when the operator depresses the push button when the tank level is low, relay CR7 is energized, which closes contact CR7 on branch rung 1*A.* Once contact CR7 is latched in, the operator may release the button. The junction connecting rungs 1 and 1*A* is equivalent to the output of the OR gate 2 in Fig. 8-58*a.*

Regulatory Control For most batch processes, the discrete logic requirements overshadow the continuous control requirements. For many batch processes, the continuous control can be provided by simple loops for flow, pressure, level, and temperature. However, very sophisticated advanced control techniques are occasionally applied. As temperature control is especially critical in reactors, the simple feedback approach is replaced by model-based strategies that rival if not exceed the sophistication of advanced control loops in continuous plants.

In some installations, alternative approaches for regulatory control



Sequential logic diagram.

Ladder logic diagram.

FIG. 8-58 Logic diagrams for the control of the mixing tank.

may be required. Where a variety of products are manufactured, the reactor may be equipped with alternative heat-removal capabilities, including the following:

1. Jacket filled with cooling water. Most such jackets are once-through, but some are recirculating.
2. Heat exchanger in a pump-around loop.
3. Reflux condenser.

The heat removal capability to be used usually depends on the product being manufactured. Therefore, regulatory loops must be configured for each possible option, and sometimes for certain combinations of the possible options. These loops are enabled and disabled depending on the product being manufactured.

The interface between continuous controls and discrete controls is also important. For example, a feed might be metered into a reactor at a variable rate, depending on another feed or possibly on reactor temperature. However, the product recipe calls for a specified quantity of this feed. The flow must be totalized (i.e., integrated), and when the flow total attains a specified value, the feed must be terminated.

The discrete logic must have access to operational parameters such as controller modes. That is, the discrete logic must be able to switch a controller to manual, auto, or cascade. Furthermore, the discrete logic must be able to force the controller output to a specified value.

Sequence Logic Sequence logic must not be confused with discrete logic. Discrete logic is especially suitable for interlocks or permissives, e.g., the reactor discharge valve must be closed in order for the feed valve to be opened. Sequence logic is used to force the process to attain the proper sequence of states. For example, a feed preparation might be to first charge *A,* then charge *B,* then mix, and finally cool. Although discrete logic can be used to implement sequence logic, other alternatives are often more attractive.

Sequence logic is often, but not necessarily, coupled with the concept of a process state. Basically, the sequence logic determines when the process should proceed from the current state to the next, and sometimes what the next state should be.

Sequence logic must encompass both normal and abnormal process operations. Thus, sequence logic is often viewed as consisting of two distinct but related parts:

1. *Normal logic.* This sequence logic provides for the normal or expected progression from one process state to another.
2. *Failure logic.* This logic provides for responding to abnormal conditions, such as equipment failures.

Of these, the failure logic can easily be the most demanding. The simplest approach is to stop or hold on any abnormal condition, and let the process operator sort things out. However, this is not always acceptable. Some failures lead to hazardous conditions that require immediate action; waiting for the operator to decide what to do is not acceptable. The appropriate response to such situations is best determined in conjunction with the process hazards analysis.

No single approach has evolved as the preferred way to implement sequence logic. The approaches utilized include the following:

1. *Discrete logic.* Sequence logic can be implemented via ladder logic, and this approach is common when sequence logic is implemented in programmable logic controllers (PLCs).
2. *Programming languages.* Traditional procedural languages do not provide the necessary constructs for implementing sequence logic. This necessitates one of the following:

a. Special languages. The necessary extensions for sequence logic are provided by extending the syntax of the programming language. This is the most common approach within distributed control systems (DCSs). The early implementations used BASIC as the starting point for the extensions; the later implementations used C as the starting point. A major problem with this approach is portability, especially from one manufacturer to the next but sometimes from one product version to the next within the same manufacturer's product line.

b. Subroutine or function libraries. The facilities for sequence logic are provided via subroutines or functions that can be referenced from programs written in FORTRAN or C. This requires a general-purpose program development environment and excellent facilities to trap the inevitable errors in such programs. Operating systems with such capabilities have long been available on the larger computers, but not for the microprocessors utilized within DCS systems. However, such operating systems are becoming more common within DCS systems.

3. *State machines.* This technology is commonly applied within the discrete manufacturing industries. However, its migration to process batch applications has been limited.
4. *Graphical implementations.* For sequence logic, the flowchart traditionally used to represent the logic of computer programs must be extended to provide parallel execution paths. Such extensions have been implemented in a graphical representation called Grafcet. As process engineers have demonstrated a strong dislike for ladder logic, PLC manufacturers are considering providing Grafcet either in addition to or as an alternative to ladder logic.

As none of the above have been able to dominate the industry, it is quite possible that future developments will provide a superior approach for implementing sequence logic.

BATCH PRODUCTION FACILITIES

Especially for flexible batch applications, the batch logic must be properly structured in order to be implemented and maintained in a reasonable manner. An underlying requirement is that the batch process equipment be properly structured. The following structure is appropriate for most batch production facilities.

Plant A plant is the collection of production facilities at a geographical site. The production facilities at a site normally share warehousing, utilities, and the like.

Equipment Suite An equipment suite is the collection of equipment available for producing a group of products. Normally, this group of products is similar in certain respects. For example, they might all be manufactured from the same major raw materials. Within the equipment suite, material transfer and metering capabilities are available for these raw materials. The equipment suite contains all of the necessary types of processing equipment (reactors, separators, and so on) required to convert the raw materials into salable products. A plant may consist of only one suite of equipment, but large plants usually contain multiple equipment suites.

Process Unit or Batch Unit A process unit is a collection of processing equipment that can, at least at certain times, be operated in a manner completely independent from the remainder of the plant. A process unit normally provides a specific function in the production of a batch of product. For example, a process unit might be a reactor complete with all associated equipment (jacket, recirculation pump, reflux condenser, and so on). However, each feed preparation tank is usually a separate process unit. With this separation, preparation of the feed for the next batch can be started as soon as the feed tank is emptied for the current batch.

All but the very simplest equipment suites contain multiple process units. The minimum number of process units is one for each type of processing equipment required to make a batch of product. However, many equipment suites contain multiple process units of each type. In such equipment suites, multiple batches and multiple production runs can be in progress at a given time.

Item of Equipment An item of equipment is a hardware item that performs a specific purpose. Examples are pumps, heat exchangers, agitators, and the like. A process unit could consist of a single item of equipment, but most process units consist of several items of equipment that must be operated in harmony in order to achieve the function expected of the process unit.

Device A device is the smallest element of interest to batch logic. Examples of devices include measurement devices and actuators.

STRUCTURED BATCH LOGIC

Flexible batch applications must be pursued using a structured approach to batch logic. In such applications, the same processing equipment is used to make a variety of products. In most facilities, little or no proprietary technology is associated with the equipment itself; the proprietary technology is how this equipment is used to produce each of the products.

The primary objective of the structured approach is to separate cleanly the following two aspects of the batch logic:

Product Technology Basically, this encompasses the product technology, such as how to mix certain molecules to make other molecules. This technology ultimately determines the chemical and physical properties of the final product. The product recipe is the principal source for the product technology.

Process Technology The process equipment permits certain processing operations (e.g., heat to a specified temperature) to be undertaken. Each processing operation will involve certain actions (e.g., opening appropriate valves).

The need to keep these two aspects separated is best illustrated by a situation where the same product is to be made at different plants. While it is possible that the processing equipment at the two plants is identical, this is rarely the case. Suppose one plant uses steam for heating its vessels, but the other uses a hot oil system as the source of heat. When a product recipe requires that material is to be heated to a specified temperature, each plant can accomplish this objective, but will go about it in quite different ways.

The ideal case for a product recipe is as follows:

1. Contains all of the product technology required to make a product
2. Contains no equipment-dependent information, that is, no process technology

In the previous example, such a recipe would simply state that the product must be heated to a specified temperature. Whether heating is undertaken with steam or hot oil is irrelevant to the product technology. By restricting the product recipe to a given product technology, the same product recipe can be used to make products at different sites. Timing diagrams (such as Fig. 8-59) are one way to represent a recipe.

At a given site, the specific approach to be used to heat a vessel is important. The traditional approach is for an engineer at each site to expand the product recipe into a document that explains in detail how the product is to be made at the site. This document goes by various names, although standard operating procedure or SOP is a common one. Depending on the level of detail to which it is written, the SOP could specify exactly which valves must be opened in order to heat the contents of a vessel. Thus, the SOP is site-dependent, and contains both product technology and process technology.

In structuring the logic for a flexible batch application, the following organization permits product technology to be cleanly separated from process technology:

- A recipe consists of a formula and one or more processing operations. Ideally, only product technology is contained in a recipe.
- A processing operation consists of one or more phases. Ideally, only product technology is contained in a processing operation.
- A phase consists of one or more actions. Ideally, only process technology is contained in a phase.

In this structure, the recipe and processing operations would be the same at each site that manufactures the product. However, the logic that comprises each phase would be specific to a given site. Using the heating example from above, each site would require a phase to heat the contents of the vessel. However, the logic within the phase at one site would accomplish the heating by opening the appropriate steam valves, while the logic at the other site would accomplish the heating by opening the appropriate hot oil valves.

Usually the critical part of structuring batch logic is the definition of the phases. There are two ways to approach this:

1. Examine the recipes for the current products for commonality, and structure the phases to reflect this commonality.
2. Examine the processing equipment to determine what processing capabilities are possible, and write phases to accomplish each possible processing capability.

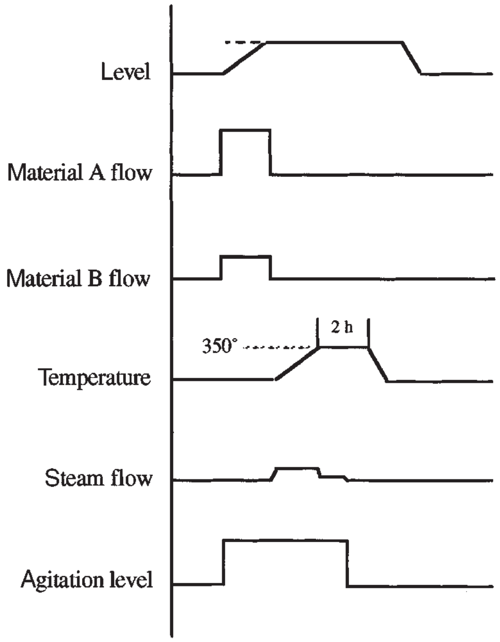


FIG. 8-59 Sample process timing diagram.

There is the additional philosophical issue of whether to have a large number of simple phases with few options each, or a small number of complex phases with numerous options. The issues are a little different from structuring a complex computer program into subprograms. Each possible alternative will have advantages and disadvantages.

As the phase contains no product technology, the implementation of a phase must be undertaken by those familiar with the process equipment. Furthermore, they should undertake this on the basis that the result will be used to make a variety of products, not just those that are initially contemplated. The development of the phase logic must also encompass all equipment-related safety issues. The phase should accomplish a clearly defined objective, so the implementers should be able to thoroughly consider all relevant issues in accomplishing this objective. The phase logic is defined in detail, implemented in the control system, and then thoroughly tested. Except when the processing equipment is modified, future modifications to the phase should be infrequent. The result should be a very dependable module that can serve as a building block for batch logic.

Even for flexible batch applications, a comprehensive menu of phases should permit most new products to be implemented using currently existing phases. By reusing exising phases, numerous advantages accrue:

1. The engineering effort to introduce a new recipe at a site is reduced.
2. The product is more likely to be on-spec the first time, thus avoiding the need to dispose of off-spec product.
3. The new product can be supplied to customers sooner, hopefully before competitors can supply the product.

There is also a distinct advantage in maintenance. When a problem with a phase is discovered and the phase logic is corrected, the correction is effectively implemented in all recipes that use the phase. If a change is implemented in the processing equipment, the affected phases must be modified accordingly and then thoroughly tested. These modifications are also effectively implemented in all recipes that use these phases.

PROCESS MEASUREMENTS

General References: Benedict, *Fundamentals of Temperature, Pressure, and Flow Measurements,* Wiley, New York, 1969. Considine, *Process Instruments and Control Handbook,* McGraw-Hill, New York, 1993. Considine and Ross, *Handbook of Applied Instrumentation,* McGraw-Hill, New York, 1964. Doebelin, *Measurement Systems: Application and Design,* 4th ed. McGraw-Hill, New York, 1990. Ginesi and Annarummo, "User Tips for Mass, Volume Flowmeters," *Tech,* 41, April, 1994. *ISA Transducer Compendium,* 2d ed., Plenum, New York, 1969. Liptak, *Instrument Engineers Handbook,* Chilton, Philadelphia, 1995. Michalski, Eckersdorf, and McGhee, *Temperature Measurement,* Wiley, Chichester, 1991. Nichols, G. D., *On-Line Process Analyzers,* Wiley, New York, 1988.

GENERAL CONSIDERATIONS

Process measurements encompass the application of the principles of metrology to the process in question. The objective is to obtain values for the current conditions within the process and make this information available in a form usable by either the control system, process operators, or any other entity that needs to know. The term "measured variable" or "process variable" designates the process condition that is being determined.

Process measurements fall into two categories:

1. *Continuous measurements.* An example of a continuous measurement is a level measurement device that determines the liquid level in a tank (in meters).
2. *Discrete measurements.* An example of a discrete measurement is a level switch that indicates the presence or absence of liquid at the location at which the level switch is installed.

In continuous processes, most process control applications rely on continuous measurements. In batch processes, many of the process control applications will utilize discrete as well as continuous measurements. In both types of processes, the safety interlocks and process interlocks rely largely on discrete measurements.

Continuous Measurements In most applications, continuous measurements are considerably more ambitious than discrete measurements. Basically, discrete measurements involve a yes/no decision, whereas continuous measurements may entail considerable signal processing.

The components of a typical continuous measurement device are as follows:

- *Sensor.* This component produces a signal that is related in a known manner to the process variable of interest. The sensors in use today are primarily of the electrical analog variety, and the signal is in the form of a voltage, a resistance, a capacitance, or some other directly measurable electrical quantity. Prior to the mid 1970s, instruments tended to use sensors whose signal was mechanical in nature, and thus compatible with pneumatic technology. Since that time the fraction of sensors that are digital in nature has grown considerably, often eliminating the need for analog-to-digital conversion.
- *Signal processing.* The signal from most sensors is related in a nonlinear fashion to the process variable of interest. In order for the output of the measurement device to be linear with respect to the process variable of interest, linearization is required. Furthermore, the signal from the sensor might be affected by variables other than the process variable. In this case, additional variables must be sensed and the signal from the sensor compensated to account for the other variables. For example, reference junction compensation is required for thermocouples (except when used for differential temperature measurements).
- *Transmitter.* The measurement device output must be a signal that can be transmitted over some distance. Where electronic analog transmission is used, the low range on the transmitter output is 4 milliamps, and the upper range is 20 milliamps. Microprocessor-based transmitters (often referred to as smart transmitters) are usually capable of transmitting the measured variable digitally in engineering units.

Accuracy and Repeatability Definitions of terminology pertaining to process measurements can be obtained from standard S51.1 from the International Society of Measurment and Control (ISA) and standard RC20-11 from the Scientific Apparatus Manufacturers Association (SAMA), both of which are updated periodically. An appreciation of accuracy and repeatability is especially important. Some applications depend on the accuracy of the instrument, but other applications depend on repeatability. Excellent accuracy implies excellent repeatability; however, an instrument can have poor accuracy but excellent repeatability. In some applications, this is acceptable, as discussed below.

Range and Span A continuous measurement device is expected to provide credible values of the measured value between a lower range and an upper range. The difference between the upper range and the lower range is the span of the measurement device. The maximum value for the upper range and the minimum value for the lower range depend on the principles on which the measurement device is based and on the design chosen by the manufacturer of the measurement device. If the measured variable is greater than the upper range or less than the lower range, the measured variable is said to be out-of-range or the measurement device is said to be overranged.

Accuracy Accuracy refers to the difference between the measured value and the true value of the measured variable. Unfortunately, the true value is never known, so in practice accuracy refers to the difference between the measured value and an accepted standard value for the measured variable.

Accuracy can be expressed in a number of ways:

1. As an absolute difference in the units of the measured variable
2. As a percent of the current reading
3. As a percent of the span of the measured variable
4. As a percent of the upper range of the span

For process measurements, accuracy as a percent of span is the most common.

Manufacturers of measurement devices always state the accuracy of the instrument. However, these statements always specify specific or reference conditions at which the measurement device will perform with the stated accuracy, with temperature and pressure most often appearing in the reference conditions. When the measurement device is applied at other conditions, the accuracy is affected. Manufacturers usually also provide some statements on how accuracy is affected when the conditions of use deviate from the referenced conditions in the statement of accuracy. Although appropriate calibration procedures can minimize some of these effects, rarely can they be totally eliminated. It is easily possible for such effects to cause a measurement device with a stated accuracy of 0.25 percent of span at reference conditions to ultimately provide measured values with accuracies of 1 percent or less. Microprocessor-based measurement devices usually provide better accuracy than the traditional electronic measurement devices.

In practice, most attention is given to accuracy when the measured variable is the basis for billing, such as in custody transfer applications. However, whenever a measurement device provides data to any type of optimization strategy, accuracy is very important.

Repeatability Repeatability refers to the difference between the measurements when the process conditions are the same. This can also be viewed from the opposite perspective. If the measured values are the same, repeatability refers to the difference between the process conditions.

For regulatory control, repeatability is of major interest. The basic objective of regulatory control is to maintain uniform process operation. Suppose that on two different occasions, it is desired that the temperature in a vessel be 80°C. The regulatory control system takes appropriate actions to bring the measured variable to 80°C. The difference between the process conditions at these two times is determined by the repeatability of the measurement device.

In the use of temperature measurement for control of the separation in a distillation column, repeatability is crucial but accuracy is not. Composition control for the overhead product would be based on a measurement of the temperature on one of the trays in the rectifying section. A target would be provided for this temperature. However, at periodic intervals, a sample of the overhead product is analyzed in the laboratory and the information provided to the process operator. Should this analysis be outside acceptable limits, the operator would adjust the set point for the temperature. This procedure effectively compensates for an inaccurate temperature measurement; however, the success of this approach requires good repeatability from the temperature measurement.

Dynamics of Process Measurements Especially where the measurement device is incorporated into a closed loop control configuration, dynamics are important. The dynamic characteristics depend on the nature of the measurement device, and also on the nature of components associated with the measurement device (for example, thermowells and sample conditioning equipment). The term *measurement system* designates the measurement device and its associated components.

The following dynamics are commonly exhibited by measurement systems:

- *Time constants.* Where there is a capacity and a throughput, the measurement device will exhibit a time constant. For example, any temperature measurement device has a thermal capacity (mass times heat capacity) and a heat flow term (heat transfer coefficient and area). Both the temperature measurement device and its associated thermowell will exhibit behavior typical of time constants.
- *Dead time.* Probably the best example of a measurement device that exhibits pure dead time is the chromatograph, because the analysis is not available for some time after a sample is injected. Additional dead time results from the transportation lag within the sample system. Even continuous analyzer installations are plagued by dead time from the sample system.
- *Underdamped behavior.* Measurement devices with mechanical components often have a natural harmonic and can exhibit underdamped behavior. The displacer type of level measurement device is capable of such behavior.

While the manufacturers of measurement devices can supply some information on the dynamic characteristics of their devices, interpretation is often difficult. Measurement device dynamics are quoted on varying bases, such as rise time, time to 63 percent response, settling time, and so on. Even where the time to 63 percent response is quoted, it might not be safe to assume that the measurement device exhibits first-order behavior.

Where the manufacturer of the measurement device does not supply the associated equipment (thermowells, sample conditioning equipment, and the like), the user must incorporate the characteristics of these components to obtain the dynamics of the measurement system.

An additional complication is that most dynamic data are stated for configurations involving reference materials such as water, air, and so on. The nature of the process material will affect the dynamic characteristics. For example, a thermowell will exhibit different characteristics when immersed in a viscous organic emulsion than when immersed in water. It is often difficult to extrapolate the available data to process conditions of interest.

Similarly, it is often impossible, or at least very difficult, to experimentally determine the characteristics of a measurement system under the conditions where it is used. It is certainly possible to fill an emulsion polymerization reactor with water and determine the dynamic characteristics of the temperature measurement system. However, it is not possible to determine these characteristics when the reactor is filled with the emulsion under polymerization conditions.

The primary impact of unfavorable measurement dynamics is on the performance of closed loop control systems. This explains why most control engineers are very concerned with measurement dynamics. The goal to improve the dynamic characteristics of measurement devices is made difficult because the discussion regarding measurement dynamics is often subjective.

Selection Criteria The selection of a measurement device entails a number of considerations given below, some of which are almost entirely subjective.

1. *Measurement span.* The measurement span required for the measured variable must lie entirely within the instrument's envelope of performance.
2. *Performance.* Depending on the application, accuracy, repeatability, or perhaps some other measure of performance is appropriate. Where closed loop control is contemplated, speed of response must be included.
3. *Reliability.* Data available from the manufacturers can be expressed in various ways and at various reference conditions. Often, previous experience with the measurement device within the purchaser's organization is weighted most heavily.
4. *Materials of construction.* The instrument must withstand the process conditions to which it is exposed. This encompasses considerations such as operating temperatures, operating pressures, corrosion, and abrasion. For some applications, seals or purges may be necessary.
5. *Prior use.* For the first installation of a specific measurement device at a site, training of maintenance personnel and purchases of spare parts might be necessary.
6. *Potential for releasing process materials to the environment.* Fugitive emissions are receiving ever increasing attention. Exposure considerations, both immediate and long term, for maintenance personnel are especially important when the process fluid is either corrosive or toxic.
7. *Electrical classification.* Article 500 of the National Electric Code provides for the classification of the hazardous nature of the process area in which the measurement device will be installed. If the measurement device is not inherently compatible with this classification, suitable enclosures must be purchased and included in the installation costs.

8. *Physical access.* Subsequent to installation, maintenance personnel must have physical access to the measurement device for maintenance and calibration. If additional structural facilities are required, they must be included in the installation costs.

9. *Cost.* There are two aspects of the cost:

a. Initial purchase and installation (capital costs).

b. Recurring costs (operational expense). This encompasses instrument maintenance, instrument calibration, consumables (for example, titrating solutions must be purchased for automatic titrators), and any other costs entailed in keeping the measurement device in service.

Calibration Calibration entails the adjustment of a measurement device so that the value from the measurement device agrees with the value from a standard. The International Standards Organization (ISO) has developed a number of standards specifically directed to calibration of measurement devices. Furthermore, compliance with the ISO 9000 standards requires that the working standard used to calibrate a measurement device must be traceable to an internationally recognized standard such as those maintained by the National Institute of Standards and Technology (NIST).

Within most companies, the responsibility for calibrating measurement devices is delegated to a specific department. Often, this department may also be responsible for maintaining the measurement device. The specific calibration procedures depend on the type of measurement device. The frequency of calibration is normally predetermined, but earlier action may be dictated if the values from the measurement device become suspect.

Calibration of some measurement devices involves comparing the measured value with the value from the working standard. Pressure and differential pressure transmitters are calibrated in this manner. Calibration of analyzers normally involves using the measurement device to analyze a specially prepared sample whose composition is known. These and similar approaches can be applied to most measurement devices.

Flow is an important measurement whose calibration presents some challenges. When a flow measurement device is used in applications such as custody transfer, provision is made to pass a known flow through the meter. However, such a provision is costly and is not available for most in-process flowmeters. Without such a provision, a true calibration of the flow element itself is not possible. For orifice meters, calibration of the flowmeter normally involves calibration of the differential pressure transmitter, and the orifice plate is usually only inspected for deformation, abrasion, and so on. Similarly, calibration of a magnetic flowmeter normally involves calibration of the voltage measurement circuitry, which is analogous to calibration of the differential pressure transmitter for an orifice meter.

TEMPERATURE MEASUREMENTS

Measurement of the hotness or coldness of a body or fluid is commonplace in the process industries. Temperature-measuring devices utilize systems with properties that vary with temperature in a simple, reproducible manner and thus can be calibrated against known references (sometimes called *secondary thermometers*). The three dominant measurement devices used in automatic control are thermocouples, resistance thermometers, and pyrometers and are applicable over different temperature regimes.

Thermocouples Temperature measurements using thermocouples are based on the discovery by Seebeck in 1821 that an electric current flows in a continuous circuit of two different metallic wires if the two junctions are at different temperatures. The thermocouple may be represented diagrammatically as shown in Fig. 8-60. *A* and *B* are the two metals, and T_1 and T_2 are the temperatures of the junctions. Let T_1 and T_2 be the reference junction (cold junction) and the measuring junction, respectively. If the thermoelectric current *i* flows in the direction indicated in Fig. 8-60, metal *A* is customarily referred to as thermoelectrically positive to metal *B.* Metal pairs used for thermocouples include platinum-rhodium (the most popular and accurate), chromel-alumel, copper-constantan, and iron-constantan. The thermal emf is a measure of the difference in temperature between T_2 and T_1. In control systems the reference junction is usually located at the emf-measuring device. The reference junction may be held at constant temperature such as in an ice bath or a thermostated oven, or it may be at ambient temperature but electrically compensated (cold-junction-compensated circuit) so that it appears to be held at a constant temperature.

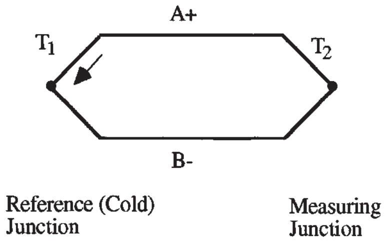


FIG. 8-60 Basic circuit of Seebeck effect.

Resistance Thermometers The resistance thermometer depends upon the inherent characteristics of materials to change in electrical resistance when they undergo a change in temperature. Industrial resistance thermometers are usually constructed of platinum, copper, or nickel, and more recently semiconducting materials such as thermistors are being used. Basically, a resistance thermometer is an instrument for measuring electrical resistance that is calibrated in units of temperature instead of in units of resistance (typically ohms). Several common forms of bridge circuits are employed in industrial resistance thermometry, the most common being the Wheatstone bridge. A resistance thermometer detector (RTD) consists of a resistance conductor (metal), which generally shows an increase in resistance with temperature. The following equation represents the variation of resistance with temperature (°C):

$$R_T = R_0(1 + a_1T + a_2T^2 + \cdots + a_nT^n)$$

$$R_0 = \text{resistance at } 0°\text{C} \qquad (8\text{-}88)$$

The temperature coefficient of resistance α_T is expressed as:

$$\alpha_T = \frac{1}{R_T}\frac{dR_T}{dT} \qquad (8\text{-}89)$$

For most metals, α_T is positive. For many pure metals, the coefficient is essentially constant and stable over large portions of their useful range. Typical resistance versus temperature curves for platinum, copper, and nickel are given in Fig. 8-61, with platinum usually the metal of choice. Platinum has a useful range of −200°C to 800°C, while Nickel (−80°C to 320°C) and copper (−100°C to 100°C) are more limited. Detailed resistance versus temperature tables are available from the National Bureau of Standards and suppliers of resistance thermometers. Table 8-7 gives recommended temperature measurement ranges for thermocouples and RTDs. Resistance thermometers are receiving increased usage because they are about ten times more accurate than thermocouples.

Thermistors Thermistors are nonlinear temperature-dependent resistors, and normally only the materials with negative temperature

TABLE 8-7 Recommended Temperature Measurement Ranges for RTDs and Thermocouples

Resistance thermometer detectors (RTDs)	
100V Pt	−200°C–+850°C
120V Ni	−80°C–+320°C
Thermocouples	
Type B	700°C–+1820°C
Type E	−175°C–+1000°C
Type J	−185°C–+1200°C
Type K	−175°C–+1372°C
Type N	0°C–+1300°C
Type R	125°C–+1768°C
Type S	150°C–+1768°C
Type T	−170°C–+400°C

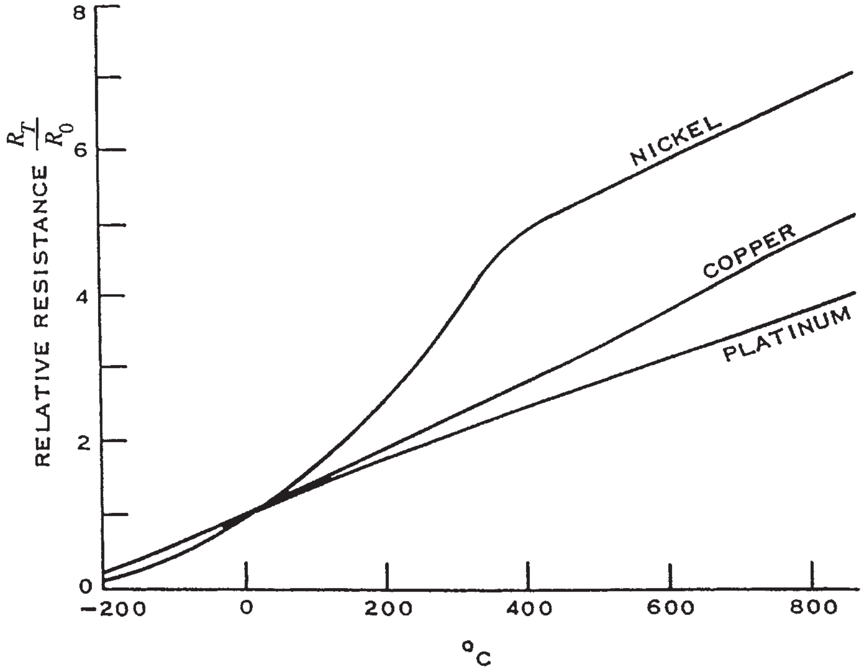


FIG. 8-61 Typical resistance-thermometer curves for platinum, copper, and nickel wire, where R_T = resistance at temperature T and R_0 = resistance at 0°C.

coefficient of resistance (NTC type) are used. The resistance is related to temperature as:

$$R_T = R_{T_r} \exp\left[\beta\left(\frac{1}{T} - \frac{1}{T_r}\right)\right] \tag{8-90}$$

where T_r is a reference temperature, which is generally 298 K. Thus

$$\alpha_T = \frac{1}{R_T}\frac{dR_T}{dT} \tag{8-91}$$

The value of β is of the order of 4000, so at room temperature (298 K), $\alpha_T = -0.045$ for thermistor and 0.0035 for 100 Ω Platinum RTD. Compared with RTDs, NTC type thermistors are advantageous in that the detector dimension can be made small, the resistance value is higher (less affected by the resistances of the connecting leads), the temperature sensitivity is higher, and the thermal inertia of the sensor is low. Disadvantages of thermistors to RTDs include nonlinear characteristics and low measuring temperature range.

Filled-System Thermometers The filled-system thermometer is designed to provide an indication of temperature some distance removed from the point of measurement. The measuring element (bulb) contains a gas or liquid that changes in volume, pressure, or vapor pressure with temperature. This change is communicated through a capillary tube to a Bourdon tube or other pressure- or volume-sensitive device. The Bourdon tube responds so as to provide a motion related to the bulb temperature. Those systems that respond to volume changes are completely filled with a liquid. Systems that respond to pressure changes either are filled with a gas or are partially filled with a volatile liquid. Changes in gas or vapor pressure with changes in bulb temperatures are carried through the capillary to the Bourdon. The latter bulbs are sometimes constructed so that the capillary is filled with a nonvolatile liquid.

Fluid-filled bulbs deliver enough power to drive controller mechanisms and even directly actuate control valves. These devices are characterized by large thermal capacity, which sometimes leads to slow response, particularly when they are enclosed in a thermal well for process measurements. Filled-system thermometers are used extensively in industrial processes for a number of reasons. The simplicity of these devices allows rugged construction, minimizing the possibility of failure with a low level of maintenance, and inexpensive overall design of control equipment. In case of system failure, the entire unit must be replaced or repaired.

As normally used in the process industries, the sensitivity and percentage of span accuracy of these thermometers are generally the equal of those of other temperature-measuring instruments. Sensitivity and absolute accuracy are not the equal of those of short-span electrical instruments used in connection with resistance-thermometer bulbs. Also, the maximum temperature is somewhat limited.

Bimetal Thermometers Thermostatic bimetal can be defined as a composite material made up of strips of two or more metals fastened together. This composite, because of the different expansion rates of its components, tends to change curvature when subjected to a change in temperature. With one end of a straight strip fixed, the other end deflects in proportion to the temperature change, the square of the length, and inversely as the thickness, throughout the linear portion of the deflection characteristic curve. If a bimetallic strip is wound into a helix or a spiral and one end is fixed, the other end will rotate when heat is applied. For a thermometer with uniform scale divisions, a bimetal must be designed to have linear deflection over the desired temperature range. Bimetal thermometers are used at temperatures ranging from 580°C down to −180°C and lower. However, at the low temperatures the rate of deflection drops off quite rapidly. Bimetal thermometers do not have long-time stability at temperatures above 430°C.

Pyrometers Planck's distribution law gives the radiated energy flux $q_b(\lambda, T)d\lambda$ in the wavelength range λ to $\lambda + d\lambda$ from a black surface:

$$q_b(\lambda, T) = \frac{C_1}{\lambda^5}\frac{1}{e^{C_2/\lambda T} - 1} \tag{8-92}$$

where $C_1 = 3.7418 \times 10^{10}$ μW μm^4 cm^{-2}, and $C_2 = 14{,}388$ μm K.

If the target object is a black body and if the pyrometer has a detector that measures the specific wavelength signal from the object, the temperature of the object can be exactly estimated from Eq. (8-92). While it is possible to construct a physical body that closely approximates black body behavior, most real-world objects are not black bodies. The deviation from a black body can be described by the spectral emissivity

$$\varepsilon_T = \frac{q(T)}{q_b(T)} \tag{8-93}$$

where $q(\lambda, T)$ is the radiated energy flux from a real body in the wavelength range λ to $\lambda + d\lambda$ and $0 < \varepsilon_{\lambda,T} < 1$. Integrating Eq. (8-92) over all wavelengths gives the Stefan-Boltzmann equation

$$\begin{aligned} q_b(T) &= \int_0^{\infty} q_b(\lambda, T)\, d\lambda \\ &= \sigma T^4 \end{aligned} \tag{8-94}$$

where σ is the Stefan-Boltzmann constant. Similar to Eq. (8-93), the emissivity ε_T for the total radiation is

$$\varepsilon_T = \frac{q(T)}{q_b(T)} \tag{8-95}$$

where $q(T)$ is the radiated energy flux from a real body with emissivity ε_T.

Total Radiation Pyrometers In total radiation pyrometers, the thermal radiation is detected over a large range of wavelengths from the object at high temperature. The detector is normally a thermopile, which is built by connecting several thermocouples in series to increase the temperature measurement range. The pyrometer is calibrated for black bodies, so the indicated temperature T_p should be converted for non-black body temperature.

Photoelectric Pyrometers Photoelectric pyrometers belong to the class of band radiation pyrometers. The thermal inertia of thermal radiation detectors does not permit the measurement of rapidly changing temperatures. For example, the smallest time constant of a thermal detector is about 1 msec, while the smallest time constant of a photoelectric detector can be about 1 or 2 sec. Photoelectric pyrometers may use photoconductors, photodiodes, photovoltaic cells, or vacuum photocells. Photoconductors are built from glass plates with thin film coatings of 1 μm thickness, using PbS, CdS, PbSe or PbTe. When the incident radiation has the same wavelength as the materials are able to absorb, the captured incident photons free photoelectrons, which form an electric current. Photodiodes in germanium or silicon are operated with a reverse bias voltage applied. Under the influence of the incident radiation their conductivity as well as their reverse saturation current is proportional to the intensity of the radiation within the spectral response band from 0.4 to 1.7 μm for Ge and 0.6 to 1.1 μm for Si. Because of the above characteristics, the operating range of a photoelectric pyrometer can be either spectral or in a specific band. Photoelectric pyrometers can be applied for a specific choice of the wavelength.

Disappearing Filament Pyrometers Disappearing filament pyrometers can be classified as spectral pyrometers. The brightness of a lamp filament is changed by adjusting the lamp current until the filament disappears against the background of the target, at which point the temperature is measured. Since the detector is the human eye, it is difficult to calibrate for on-line measurements.

Ratio Pyrometers The ratio pyrometer is also called the two-color pyrometer. Two different wavelengths are utilized for detecting the radiated signal. If one uses Wien's law for small values of λT, the detected signals from spectral radiant energy flux emitted at the wavelengths λ_1 and λ_2 with emissivities ε_{λ_1} and ε_{λ_2} are

$$S_{\lambda_1} = KC_1\varepsilon_{\lambda_1}\lambda_1^{-5} \exp^{-C_2/\lambda_1 T} \tag{8-96}$$

$$S_{\lambda_2} = KC_1\varepsilon_{\lambda_2}\lambda_2^{-5} \exp^{-C_2/\lambda_2 T} \tag{8-97}$$

The ratio of the signals S_{λ_1} and S_{λ_2} is

$$\frac{S_{\lambda_1}}{S_{\lambda_2}} = \frac{\varepsilon_{\lambda_1}}{\varepsilon_{\lambda_2}} \left(\frac{\lambda_2}{\lambda_1}\right)^5 \exp\left[\frac{C_2}{T}\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right)\right] \tag{8-98}$$

Nonblack or nongrey bodies are characterized by wavelength dependence of their spectral emissivity. Let T_c be defined as the temperature of the body corresponding to the temperature of a black body. If the ratio of its radiant intensities at the wavelengths λ_1, and λ_2 equals the ratio of the radiant intensities of the nonblack body, whose temperature is to be measured at the same wavelength, then Wien's law gives

$$\frac{\varepsilon_{\lambda_1} \exp^{-C_2/\lambda_1 T}}{\varepsilon_{\lambda_2} \exp^{-C_2/\lambda_2 T}} = \frac{\exp^{-C_2/\lambda_1 T_c}}{\exp^{-C_2/\lambda_2 T_c}} \tag{8-99}$$

where T is the true temperature of the body. Rearranging Eq. (8-99) gives

$$T = \left[\frac{\ln \varepsilon_{\lambda_1}/\varepsilon_{\lambda_2}}{C_2\left(\dfrac{1}{\lambda_1} - \dfrac{1}{\lambda_2}\right)} + \frac{1}{T_c}\right]^{-1} \tag{8-100}$$

For black or grey bodies, Eq. (8-98) reduces to

$$\frac{S_{\lambda_1}}{S_{\lambda_2}} = \left(\frac{\lambda_2}{\lambda_1}\right)^5 \exp\left[\frac{C_2}{T}\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right)\right] \tag{8-101}$$

Thus, by measuring S_{λ_1} and S_{λ_2}, the temperature T can be estimated.

Accuracy of Pyrometers Most of the temperature estimation methods for pyrometers assume that the object is either a grey body or has known emissivity values. The emissivity of the nonblack body depends on the internal state or the surface geometry of the objects. Also, the medium through which the thermal radiation passes is not always transparent. These inherent uncertainties of the emissivity values make the accurate estimation of the temperature of the target objects difficult. Proper selection of the pyrometer and accurate emissivity values can provide a high level of accuracy.

PRESSURE MEASUREMENTS

Pressure defined as force per unit area is usually expressed in terms of familiar units of weight-force and area or the height of a column of liquid that produces a like pressure at its base. Process pressure-measuring devices may be divided into three groups: (1) those that are based on the measurement of the height of a liquid column, (2) those that are based on the measurement of the distortion of an elastic pressure chamber, and (3) electrical sensing devices.

Liquid-Column Methods Liquid-column pressure-measuring devices are those in which the pressure being measured is balanced against the pressure exerted by a column of liquid. If the density of the liquid is known, the height of the liquid column is a measure of the pressure. Most forms of liquid-column pressure-measuring devices are commonly called manometers. When the height of the liquid is observed visually, the liquid columns are contained in glass or other transparent tubes. The height of the liquid column may be measured in length units or be calibrated in pressure units. Depending on the pressure range, water and mercury are the liquids most frequently used. Since the density of the liquid used varies with temperature, the temperature must be taken into account for accurate pressure measurements.

Elastic-Element Methods Elastic-element pressure-measuring devices are those in which the measured pressure deforms some elastic material (usually metallic) within its elastic limit, the magnitude of the deformation being approximately proportional to the applied pressure. These devices may be loosely classified into three types: Bourdon tube, bellows, and diaphragm.

Bourdon-Tube Elements Probably the most frequently used process pressure-indicating device is the C-spring Bourdon-tube pressure gauge. Gauges of this general type are available in a wide variety of pressure ranges and materials of construction. Materials are selected on the basis of pressure range, resistance to corrosion by the process materials, and effect of temperature on calibration. Gauges calibrated with pressure, vacuum, compound (combination pressure and vacuum), and suppressed-zero ranges are available.

Bellows Element The bellows element is an axially elastic cylinder with deep folds or convolutions. The bellows may be used unopposed, or it may be restrained by an opposing spring. The pressure to be measured may be applied either to the inside or to the space outside the bellows, with the other side exposed to atmospheric pressure. For measurement of absolute pressure either the inside or the space outside of the bellows can be evacuated and sealed. Differential pressures may be measured by applying the pressures to opposite sides of a single bellows or to two opposing bellows.

Diaphragm Elements Diaphragm elements may be classified into two principal types: those that utilize the elastic characteristics of the diaphragm and those that are opposed by a spring or other separate elastic element. The first type usually consists of one or more capsules, each composed of two diaphragms bonded together by soldering, brazing, or welding. The diaphragms are flat or corrugated circular metallic disks. Metals commonly used in diaphragm elements include brass, phosphor bronze, beryllium copper, and stainless steel. Ranges are available from fractions of an inch of water to about 206.8 kPa gauge. The second type of diaphragm is used for containing the pressure and exerting a force on the opposing elastic element. The diaphragm is a flexible or slack diaphragm of rubber, leather, impregnated fabric, or plastic. Movement of the diaphragm is opposed by a spring that determines the deflection for a given pressure. This type of diaphragm is used for the measurement of extremely low pressure, vacuum, or differential pressure.

Electrical Methods

Strain Gauges When a wire or other electrical conductor is stretched elastically, its length is increased and its diameter is decreased. Both of these dimensional changes result in an increase in the electrical resistance of the conductor. Devices utilizing resistance-wire grids for measuring small distortions in elastically stressed materials are commonly called strain gauges. Pressure-measuring elements utilizing strain gauges are available in a wide variety of forms. They usually consist of one of the elastic elements described earlier to which one or more strain gauges have been attached to measure the deformation. There are two basic strain-gauge forms: bonded and unbonded. Bonded strain gauges are those which are bonded directly to the surface of the elastic element whose strain is to be measured. The unbonded-strain-gauge transducer consists of a fixed frame and an armature which moves with respect to the frame in response to the measured pressure. The strain-gauge wire filaments are stretched between the armature and frame. The strain gauges are usually connected electrically in a Wheatstone-bridge configuration.

Strain-gauge pressure transducers are manufactured in many forms for measuring gauge, absolute, and differential pressures and vacuum. Full-scale ranges from 25.4 mm of water to 10,134 MPa are available. Strain gauges bonded directly to a diaphragm pressure-sensitive element usually have an extremely fast response time and are suitable for high-frequency dynamic-pressure measurements.

Piezoresistive Transducers A variation of the conventional strain-gauge pressure transducer uses bonded single-crystal semiconductor wafers, usually silicon, whose resistance varies with strain or distortion. Transducer construction and electrical configurations are similar to those using conventional strain gauges. A permanent magnetic field is applied perpendicular to the resonating sensor. An AC current causes the resonator to vibrate, and the resonant frequency is a function of the pressure (tension) of the resonator. The principal advantages of piezoresistive transducers are a much higher bridge voltage output and smaller size. Full-scale output voltages of 50 to 100 mV/V of excitation are typical. Some newer devices provide digital rather than analog output.

Piezoelectric Transducers Certain crystals produce a potential difference between their surfaces when stressed in appropriate directions. Piezoelectric pressure transducers generate a potential difference proportional to a pressure-generated stress. Because of the extremely high electrical impedance of piezoelectric crystals at low frequency, these transducers are usually not suitable for measurement of static process pressures.

FLOW MEASUREMENTS

Flow, defined as volume per unit of time at specified temperature and pressure conditions, is generally measured by positive-displacement or rate meters. The term "positive-displacement meter" applies to a device in which the flow is divided into isolated measured volumes when the number of fillings of these volumes is counted in some manner. The term "rate meter" applies to all types of flowmeters through which the material passes without being divided into isolated quantities. Movement of the material is usually sensed by a primary measuring element that activates a secondary device. The flow rate is then inferred from the response of the secondary device by means of known physical laws or from empirical relationships.

The principal classes of flow-measuring instruments used in the process industries are variable-head, variable-area, positive-displacement, and turbine instruments, mass flowmeters, vortex-shedding and ultrasonic flowmeters, magnetic flowmeters, and more recently, Coriolis mass flowmeters. Head meters are covered in more detail in Sec. 5.

Orifice Meter The most widely used flowmeter involves placing a fixed-area flow restriction (an orifice) in the pipe carrying the fluid. This flow restriction causes a pressure drop that can be related to flow rate. The sharp-edge orifice is popular because of its simplicity, low cost, and the large amount of research data on its behavior. For the orifice meter, the flow rate Q_a for a liquid is given by

$$Q_a = \frac{C_d A_2}{\sqrt{1-(A_2/A_1)^2}} \cdot \sqrt{\frac{2(p_1 - p_2)}{\rho}} \qquad (8\text{-}102)$$

where $p_1 - p_2$ is the pressure drop, ρ is the density, A_1 is the pipe cross-sectional area, A_2 is the orifice cross-sectional area, and C_d is the discharge coefficient. The discharge coefficient C_d varies with the Reynolds number at the orifice and can be calibrated with a single fluid, such as water (typically $C_d \approx 0.6$). If the orifice and pressure taps are constructed according to certain standard dimensions, quite accurate (about 0.4 to 0.8 percent error) values of C_d may be obtained. It should also be noted that the standard calibration data assume no significant flow disturbances such as elbows, valves, and so on, for a certain minimum distance upstream of the orifice. The presence of such disturbances close to the orifice can cause errors of as much as 15 percent. Accuracy in measurements limits the meter to a range of 3:1. The orifice has a relatively large permanent pressure loss that must be made up by the pumping machinery.

Venturi Meter The venturi tube operates on exactly the same principle as the orifice [see Eq. (8-102)]. Discharge coefficients of venturis are larger than those for orifices and vary from about 0.94 to 0.99. A venturi gives a definite improvement in power losses over an orifice and is often indicated for measuring very large flow rates, where power losses can become economically significant. The initial higher cost of a venturi over an orifice may thus be offset by reduced operating costs.

Rotameter A rotameter consists of a vertical tube with a tapered bore in which a float changes position with the flow rate through the tube. For a given flow rate the float remains stationary since the vertical forces of differential pressure, gravity, viscosity, and buoyancy are balanced. The float position is the output of the meter and can be made essentially linear with flow rate by making the tube area vary linearly with the vertical distance.

Turbine Meter If a turbine wheel is placed in a pipe containing a flowing fluid, its rotary speed depends on the flow rate of the fluid. A turbine can be designed whose speed varies linearly with flow rate. The speed can be measured accurately by counting the rate at which turbine blades pass a given point, using magnetic pickup to produce voltage pulses. By feeding these pulses to an electronic pulse-rate meter, one can measure flow rate by summing the pulses during a timed interval. Turbine meters are available with full-scale flow rates ranging from about 0.1 to 30,000 gpm for liquids and 0.1 to 15,000 ft^3/min for air. Nonlinearity can be less than 0.05 percent in the larger sizes. Pressure drop across the meter varies with the square of flow rate and is about 3 to 10 psi at full flow. Turbine meters can follow flow transients quite accurately since their fluid/mechanical time constant is of the order of 2 to 10 msec.

Vortex-Shedding Flowmeters These flowmeters take advantage of vortex shedding, which occurs when a fluid flows past a non-streamlined object (a blunt body). The flow cannot follow the shape of the object and separates from it, forming turbulent vortices or eddies at the object's side surfaces. As the vortices move downstream, they grow in size and are eventually shed or detached from the object.

Shedding takes place alternately at either side of the object, and the rate of vortex formation and shedding is directly proportional to the volumetric flow rate. The vortices are counted and used to develop a signal linearly proportional to the flow rate. The digital signals can easily be totaled over an interval of time to yield the flow rate. Accuracy can be maintained regardless of density, viscosity, temperature, or pressure when the Reynolds number is greater than 10,000. There is usually a low flow cutoff point below which the meter output is clamped at zero. This flowmeter is recommended for use with relatively clean, low viscosity liquids, gases, and vapors, and rangeability of 10:1 to 20:1 is typical. A sufficient length of straight-run pipe is necessary to prevent distortion in the fluid velocity profile.

Ultrasonic Flowmeters All ultrasonic flowmeters are based upon the variable time delays of received sound waves that arise when a flowing liquid's rate of flow is varied. Two fundamental measurement techniques, depending upon liquid cleanliness, are generally used. In the first technique, two opposing transducers are inserted in a pipe so that one transducer is downstream from the other. These transducers are then used to measure the difference between the velocity at which the sound travels with the direction of flow and the velocity at which it travels against the direction of flow. The differential velocity is measured either by (1) direct time delays using sound wave burst or (2) frequency shifts derived from beat-together, continuous signals. The frequency-measurement technique is usually preferred because of its simplicity and independence of the liquid static velocity. A relatively clean liquid is required to preserve the uniqueness of the measurement path.

In the second technique, the flowing liquid must contain scatters in the form of particles or bubbles that will reflect the sound waves. These scatters should be traveling at the velocity of the liquid. A Doppler method is applied by transmitting sound waves along the flow path and measuring the frequency shift in the returned signal from the scatters in the process fluid. This frequency shift is proportional to liquid velocity.

Magnetic Flowmeters The principle behind these flowmeters is Faraday's law of electromagnetic inductance. The magnitude of the voltage induced in a conductive medium moving at right angles through a magnetic field is directly proportional to the product of the magnetic flux density, the velocity of the medium, and the path length between the probes. A minimum value of fluid conductivity is required to make this approach viable. The pressure of multiple phases or undissolved solids can affect the accuracy of the measurement if the velocities of the phases are different than that for straight-run pipe. Magmeters are very accurate over wide flow ranges and are especially accurate at low flow rates. Typical applications include metering viscous fluids, slurries, or highly corrosive chemicals. Because magmeters should be filled with fluid, the preferred installation is in vertical lines with flow going upwards. However, magmeters can be used in tight piping schemes where it is impractical to have long pipe runs, typically requiring lengths equivalent to five or more pipe diameters.

Coriolis Mass Flowmeters Coriolis mass flowmeters utilize a vibrating tube in which Coriolis acceleration of a fluid in a flow loop can be created and measured. They can be used with virtually any liquid and are extremely insensitive to operating conditions, with high pressure over ranges of 100:1. These meters are more expensive than volumetric meters and range in size from g to 6 inches. Due to the circuitous path of flow through the meter, Coriolis flowmeters exhibit higher than average pressure changes. The meter should be installed so that it will remain full of fluid, with the best installation in a vertical pipe with flow going upward. There is no Reynolds number limitation with this meter, and it is quite insensitive to velocity profile distortions and swirl, hence there is no requirement for straight piping upstream.

LEVEL MEASUREMENTS

The measurement of level can be defined as the determination of the location of the interface between two fluids, separable by gravity, with respect to a fixed datum plane. The most common level measurement is that of the interface between a liquid and a gas. Other level measurements frequently encountered are the interface between two liquids, between a granular or fluidized solid and a gas, and between a liquid and its vapor.

A commonly used basis for classification of level devices is as follows: float-actuated, displacer, and head devices, and a miscellaneous group that depends mainly on fluid characteristics.

Float-Actuated Devices Float-actuated devices are characterized by a buoyant member that floats at the interface between two fluids. Since a significant force is usually required to move the indicating mechanism, float-actuated devices are generally limited to liquid-gas interfaces. By properly weighting the float, they can be used to measure liquid-liquid interfaces. Float-actuated devices may be classified on the basis of the method used to couple the float motion to the indicating system as discussed below.

Chain or Tape Float Gauge In these types of gauges, the float is connected to the indicating mechanism by means of a flexible chain or tape. These gauges are commonly used in large atmospheric storage tanks. The gauge-board type is provided with a counterweight to keep the tape or chain taut. The tape is stored in the gauge head on a spring-loaded reel. The float is usually a pancake-shaped hollow metal float with guide wires from top to bottom of the tank to constrain it.

Lever and Shaft Mechanisms In pressurized vessels, float-actuated lever and shaft mechanisms are frequently used for level measurement. This type of mechanism consists of a hollow metal float and lever attached to a rotary shaft, which transmits the float motion to the outside of the vessel through a rotary seal.

Magnetically Coupled Devices A variety of float-actuated level devices that transmit the float motion by means of magnetic coupling have been developed. Typical of this class of devices are magnetically operated level switches and magnetic-bond float gauges. A typical magnetic-bond float gauge consists of a hollow magnet-carrying float that rides along a vertical nonmagnetic guide tube. The follower magnet is connected and drives an indicating dial similar to that on a conventional tape float gauge. The float and guide tube are in contact with the measured fluid and come in a variety of materials for resistance to corrosion and to withstand high pressures or vacuum. Weighted floats for liquid-liquid interfaces are available.

Head Devices A variety of devices utilize hydrostatic head as a measure of level. As in the case of displacer devices, accurate level measurement by hydrostatic head requires an accurate knowledge of the densities of both heavier-phase and lighter-phase fluids. The majority of this class of systems utilize standard-pressure and differential-pressure measuring devices.

Bubble-Tube Systems The commonly used bubble-tube system sharply reduces restrictions on the location of the measuring element. In order to eliminate or reduce variations in pressure drop due to the gas flow rate, a constant differential regulator is commonly employed to maintain a constant gas flow rate. Since the flow of gas through the bubble tube prevents entry of the process liquid into the measuring system, this technique is particularly useful with corrosive or viscous liquids, liquids subject to freezing, and liquids containing entrained solids.

Electrical Methods Two electrical characteristics of fluids—conductivity and dielectric constant—are frequently used to distinguish between two phases for level-measurement purposes. An application of electrical conductivity is the fixed-point level detection of a conductive liquid such as high and low water levels. A voltage is applied between two electrodes inserted into the vessel at different levels. When both electrodes are immersed in the liquid, a current flows. Capacitance-type level measurements are based on the fact that the electrical capacitance between two electrodes varies with the dielectric constant of the material between them. A typical continuous level-measurement system consists of a rod electrode positioned vertically in a vessel, the other electrode usually being the metallic vessel wall. The electrical capacitance between the electrodes is a measure of the height of the interface along the rod electrode. The rod is usually conductively insulated from process fluids by a coating of plastic. The dielectric constant of most liquids and solids is markedly higher than that of gases and vapors. The dielectric constant of water and other polar liquids is also higher than that of hydrocarbons and other nonpolar liquids.

Thermal Methods Level-measuring systems may be based on the difference in thermal characteristics between the fluids, such as temperature or thermal conductivity. A fixed-point level sensor based on the difference in thermal conductivity between two fluids consists of an electrically heated thermistor inserted into the vessel. The temperature of the thermistor and consequently its electrical resistance increase as the thermal conductivity of the fluid in which it is immersed decreases. Since the thermal conductivity of liquids is markedly higher than that of vapors, such a device can be used as a point level detector for liquid-vapor interface.

Sonic Methods A fixed-point level detector based on sonic-propagation characteristics is available for detection of a liquid-vapor interface. This device uses a piezoelectric transmitter and receiver, separated by a short gap. When the gap is filled with liquid, ultrasonic energy is transmitted across the gap, and the receiver actuates a relay. With a vapor filling the gap, the transmission of ultrasonic energy is insufficient to actuate the receiver.

PHYSICAL PROPERTY MEASUREMENTS

Physical-property measurements are sometimes equivalent to composition analyzers, because the composition can frequently be inferred from the measurement of a selected physical property.

Density and Specific Gravity For binary or pseudobinary mixtures of liquids or gases or a solution of a solid or gas in a solvent, the density is a function of the composition at a given temperature and pressure. Specific gravity is the ratio of the density of a noncompressible substance to the density of water at the same physical conditions. For nonideal solutions, empirical calibration will give the relationship between density and composition. Several types of measuring devices are described below.

Liquid Column Density may be determined by measuring the gauge pressure at the base of a fixed-height liquid column open to the atmosphere. If the process system is closed, then a differential pressure measurement is made between the bottom of the fixed height liquid column and the vapor over the column. If vapor space is not always present, the differential-pressure measurement is made between the bottom and top of a fixed-height column with the top measurement being made at a point below the liquid surface.

Displacement There are a variety of density-measurement devices based on displacement techniques. A hydrometer is a constant-weight, variable-immersion device. The degree of immersion, when the weight of the hydrometer equals the weight of the displaced liquid, is a measure of the density. The hydrometer is adaptable to manual or automatic usage. Another modification includes a magnetic float suspended below a solenoid, the varying magnetic field maintaining the float at a constant distance from the solenoid. Change in position of the float, resulting from a density change, excites an electrical system which increases or decreases the current through the solenoid.

Direct Mass Measurement One type of densitometer measures the natural vibration frequency and relates the amplitude to changes in density. The density sensor is a U-shaped tube held stationary at its node points and allowed to vibrate at its natural frequency. At the curved end of the U is an electrochemical device that periodically strikes the tube. At the other end of the U, the fluid is continuously passed through the tube. Between strikes, the tube vibrates at its natural frequency. The frequency changes directly in proportion to changes in density. A pickup device at the curved end of the U measures the frequency and electronically determines the fluid density. This technique is useful because it is not affected by the optical properties of the fluid. However, particulate matter in the process fluid can affect the accuracy.

Radiation-Density Gauges Gamma radiation may be used to measure the density of material inside a pipe or process vessel. The equipment is basically the same as for level measurement, except that here the pipe or vessel must be filled over the effective, irradiated sample volume. The source is mounted on one side of the pipe or vessel and the detector on the other side with appropriate safety radiation shielding surrounding the installation. Cesium 137 is used as the radiation source for path lengths under 610 mm (24 in) and cobalt 60 above 610 mm. The detector is usually an ionization gauge. The absorption of the gamma radiation is a function of density. Since the absorption path includes the pipe or vessel walls, an empirical calibration is used. Appropriate corrections must be made for the source intensity decay with time.

Viscosity Continuous viscometers generally measure either the resistance to flow or the drag or torque produced by movement of an element (moving surface) through the fluid. Each installation is normally applied over a narrow range of viscosities. Empirical calibration over this range allows use on both newtonian and nonnewtonian fluids. One such device uses a piston inside a cylinder. The hydrodynamic pressure of the process fluid raises the piston to a preset height. Then the inlet valve closes and the piston is allowed to free-fall, and the time of travel (typically a few seconds) is a measure of viscosity. Other geometries include the rotation of a spindle inside a sample chamber and a vibrating probe immersed in the fluid. Because viscosity depends on temperature, the viscosity measurement must be thermostated with a heater or cooler.

Refractive-Index When light travels from one medium (e.g., air or glass) into another (e.g., a liquid), it undergoes a change of velocity and, if the angle of incidence is not 90°, a change of direction. For a given interface, angle, temperature, and wavelength of light the amount of deviation or refraction will depend on the composition of the liquid. If the sample is transparent, the normal method is to measure the refraction of light transmitted through the glass-sample interface. If the sample is opaque, the reflectance near the critical angle at a glass-sample interface is measured. In an on-line refractometer, the process fluid is separated from the optics by a prism material. A beam of light is focused on a point in the fluid which creates a conic section of light at the prism, striking the fluid at different angles (greater than or less than the critical angle). The critical angle depends on the species concentrations; as the critical angle changes, the proportions of reflected and refracted light change. A photodetector produces a voltage signal proportional to the light refracted, when compared to a reference signal. Refractometers can be used with opaque fluids and in streams that contain particulates.

Dielectric Constant The dielectric constant of material represents its ability to reduce the electric force between two charges separated in space. This property is useful in process control for polymers, ceramic materials, and semiconductors. Dielectric constants are measured with respect to vacuum (1.0); typical values range from 2 (benzene) to 33 (methanol) to 80 (water). The value for water is higher than for most plastics. A measuring cell is made of glass or some other insulating material and is usually doughnut-shaped, with the cylinders coated with metal, which constitute the plates of the capacitor.

Thermal Conductivity All gases and vapor have the ability to conduct heat from a heat source. At a given temperature and physical environment, radiation, and convection heat losses will be stabilized and the temperature of the heat source will be mainly dependent on the thermal conductivity and thus the composition of the surrounding gases. Thermal-conductivity analyzers normally consist of a sample cell and a reference cell, each containing a combined heat source and detector. These cells are normally contained in a metal block with two small cavities in which the detectors are mounted. The sample flows through the sample-cell cavity past the detector. The reference cell is an identical cavity with a detector through which a known gas flows. The combined heat source and detectors are normally either wire filaments or thermistors heated by a constant current. Since their resistance is a function of temperature, the sample-detector resistance will vary with sample composition while the reference-detector resistance will remain constant. The output from the detector bridge will be a function of sample composition.

CHEMICAL COMPOSITION ANALYZERS

A number of composition analyzers used for process monitoring and control require chemical conversion of one or more sample components preceding quantitative measurement. These reactions include formation of suspended solids for turbidimetric measurement, formation of colored materials for colorimetric detection, selective oxidation or reduction for electrochemical measurement, and formation of electrolytes for measurement by electrical conductance. Some nonvolatile materials may be separated and measured by gas chromatography after conversion into volatile derivatives.

Chromatographic Analyzers Chromatographic analyzers are widely used for the separation and measurement of volatile compounds and of compounds that can be quantitatively converted into volatile derivatives. These materials are separated by placing a portion of the sample in a chromatographic column and carrying the compounds through the column with a gas stream. As a result of the different affinities of the sample components for the column packing, the compounds emerge successively as binary mixtures with the carrier gas. A detector at the column outlet measures some physical property which can be related to the concentrations of the compounds in the carrier gas. Both the concentration peak height and the peak height-time integral, (i.e., peak area) can be related to the concentration of the compound in the original sample. The two detectors most commonly used for process chromatographs are the thermal-conductivity detector and the hydrogen-flame ionization detector. Thermal-conductivity detectors, discussed earlier, require calibration for the thermal response of each compound. Hydrogen-flame ionization detectors are more complicated than thermal-conductivity detectors but are capable of 100 to 10,000 times greater sensitivity for hydrocarbons and organic compounds. For ultrasensitive detection of trace impurities, carrier gases must be specially purified.

Infrared Analyzers Many gaseous and liquid compounds absorb infrared radiation to some degree. The degree of absorption at specific wavelengths depends on molecular structure and concentration. There are two common detector types for nondispersive infrared analyzers. These analyzers normally have two beams of radiation, an analyzing and a reference beam. One type of detector consists of two gas-filled cells separated by a diaphragm. As the amount of infrared energy absorbed by the detector gas in one cell changes, the cell pressure changes. This causes movement in the diaphragm, which in turn causes a change in capacitance between the diaphragm and a reference electrode. This change in electrical capacitance is measured as the output. The second type of detector consists of two thermopiles or two bolometers, one in each of the two radiation beams. The infrared radiation absorbed by the detector is measured by a differential thermocouple output or a resistance-thermometer (bolometer) bridge circuit.

With gas-filled detectors, a chopped light system is normally used in which one side of the detector sees the source through the analyzing beam and the other side the reference beam, alternating at a frequency of a few hertz.

Ultraviolet and Visible-Radiation Analyzers Many gas and liquid compounds absorb radiation in the near-ultraviolet or visible region. For example, organic compounds containing aromatic and carbonyl structural groups are good absorbers in the ultraviolet region. Also many inorganic salts and gases absorb in the ultraviolet or visible region. In contrast, straight-chain and saturated hydrocarbons, inert gases, air, and water vapor are essentially transparent. Process analyzers are designed to measure the absorbance in a particular wavelength band. The desired band is normally isolated by means of optical filters. When the absorbance is in the visible region, the term "colorimetry" is used. A phototube is the normal detector. Appropriate optical filters are used to limit the energy reaching the detector to the desired level and the desired wavelength region. Since absorption by the sample is logarithmic if a sufficiently narrow wavelength region is used, an exponential amplifier is sometimes used to compensate and produce a linear output.

Paramagnetism A few gases including O_2, NO, and NO_2 exhibit paramagnetic properties as a result of unpaired electrons. In a nonuniform magnetic field, paramagnetic gases, because of their magnetic susceptibility, tend to move toward the strongest part of the field, thus displacing diamagnetic gases. Paramagnetic susceptibility of these gases decreases with temperature. These effects permit measurement of the concentration of the strongest paramagnetic gas, oxygen. This analyzer used a dumbbell suspended in the magnetic field which is repelled or attracted toward the magnetic field depending on the magnetic susceptibility of the gas.

ELECTROANALYTICAL INSTRUMENTS

Conductometric Analysis Solutions of electrolytes in ionizing solvents (e.g., water) conduct current when an electrical potential is applied across electrodes immersed in the solution. Conductance is a function of ion concentration, ionic charge, and ion mobility. Conductance measurements are ideally suited for measurement of the concentration of a single strong electrolyte in dilute solutions. At higher concentrations, conductance becomes a complex, nonlinear function of concentration requiring suitable calibration for quantitative measurements.

Measurement of pH The primary detecting element in pH measurement is the glass electrode. A potential is developed at the pH-sensitive glass membrane as a result of differences in hydrogen ion activity in the sample and a standard solution contained within the electrode. This potential measured relative to the potential of the reference electrode gives a voltage that is expressed as pH. Instrumentation for pH measurement is among the most widely used process-measurement devices. Rugged electrode systems and highly reliable electronic circuits have been developed for this use.

After installation, the majority of pH measurement problems are sensor-related, mostly on the reference side, including junction plugging, poisoning, and depletion of electrolyte. For the glass (measuring electrode), common difficulties are broken or cracked glass, coating, and etching or abrasion. Symptoms such as drift, sluggish response, unstable readings, and inability to calibrate are indications of measurement problems. On-line diagnostics such as impedance measurements, wiring checks, and electrode temperature are now available in most instruments. Other characteristics that can be measured off-line include efficiency or slope and asymmetry potential (offset), which indicate whether the unit should be cleaned or changed [Nichols, *Chem. Engr. Prog.,* **90**(12), 64, 1994; McMillan, *Chem. Engr. Prog.,* **87**(12), 30, 1991].

Specific-Ion Electrodes In addition to the pH glass electrode specific for hydrogen ions, a number of electrodes that are selective for the measurement of other ions have been developed. This selectivity is obtained through the composition of the electrode membrane (glass, polymer, or liquid-liquid) and the composition of the electrode. These electrodes are subject to interference from other ions, and the response is a function of the total ionic strength of the solution. However, electrodes have been designed to be highly selective for specific ions, and when properly used, these provide valuable process measurements.

MOISTURE MEASUREMENT

Moisture measurements are important in the process industries because moisture can foul products, poison reactions, damage equipment, or cause explosions. Moisture measurements include both absolute-moisture methods and relative-humidity methods. The absolute methods are those that provide a primary output that can be directly calibrated in terms of dew-point temperature, molar concentration, or weight concentration. Loss of weight on heating is the most familiar of these methods. The relative-humidity methods are those that provide a primary output that can be more directly calibrated in terms of percentage of saturation of moisture.

Dew-Point Method For many applications, the dew point is the desired moisture measurement. When concentration is desired, the relation between water content and dew point is well-known and available. The dew-point method requires an inert surface whose temperature can be adjusted and measured, a sample gas stream flowing past the surface, a manipulated variable for adjusting the surface temperature to the dew point, and a means of detecting the onset of condensation.

Although the presence of condensate can be detected electrically, the original and most often used method is the optical detection of change in light reflection from an inert metallic-surface mirror. Some instruments measure the attenuation of reflected light at the onset of condensation. Others measure the increase of light dispersed and scattered by the condensate instead of, or in addition to, the reflected-light measurement. Surface cooling is obtained with an expendable refrigerant liquid, conventional mechanical refrigeration, or thermoelectric cooling. Surface-temperature measurement is usually made with a thermocouple or a thermistor.

Piezoelectric Method A piezoelectric crystal in a suitable oscillator circuit will oscillate at a frequency dependent on its mass. If the crystal has a stable hygroscopic film on its surface, the equivalent mass of the crystal varies with the mass of water sorbed in the film. Thus, the frequency of oscillation depends on the water in the film. The analyzer contains two such crystals in matched oscillator circuits. Typically, valves alternately direct the sample to one crystal and a dry gas to the other on a 30-s cycle. The oscillator frequencies of the two circuits are compared electronically, and the output is the difference between the two frequencies. This output is then representative of the moisture content of the sample. The output frequency is usually converted to a variable DC voltage for meter readout and recording. Multiple ranges are provided for measurement from about 1 ppm to near saturation. The dry reference gas is preferably the same as the sample except for the moisture content of the sample. Other reference gases which are adsorbed in a manner similar to the dried sample gas may be used. The dry gas is usually supplied by an automatic dryer. The method requires a vapor sample to the detector. Mist striking the detector destroys the accuracy of measurement until it vaporizes or is washed off the crystals. Water droplets or mist may destroy the hygroscopic film, thus requiring crystal replacement. Vaporization or gas-liquid strippers may sometimes be used for the analysis of moisture in liquids.

Capacitance Method Several analyzers utilize the high dielectric constant of water for its detection in solutions. The alternating electric current through a capacitor containing all or part of the sample between the capacitor plates is measured. Selectivity and sensitivity are enhanced by increasing the concentration of moisture in the cell by filling the capacitor sample cell with a moisture-specific sorbent as part of the dielectric. This both increases the moisture content and reduces the amount of other interfering sample components. Granulated alumina is the most frequently used sorbent. These detectors may be cleaned and recharged easily and with satisfactory reproducibility if the sorbent itself is uniform.

Oxide Sensors Aluminum oxide can be used as a sensor for moisture analysis. A conductivity cell has one electrode node of aluminum, which is anodized to form a thin film of aluminum oxide, followed by coating with a thin layer of gold (the opposite electrode). Moisture is selectively adsorbed through the gold layer and into the hygroscopic aluminum oxide layer, which in turn determines the electrical conductivity between gold and aluminum oxide. This value can be related to ppm water in the sample. This sensor can operate between near vacuum to several hundred atmospheres, and it is independent of flow rate (including static conditions). Temperature, however, must be carefully monitored. A similar device is based on phosphorous pentoxide. Moisture content influences the electrical current between two inert metal electrodes, which are fabricated as a helix on the inner wall of a tubular nonconductive sample cell. For a constant DC voltage applied to the electrodes, a current flows which is proportional to moisture. The moisture is absorbed into the hygroscopic phosphorous pentoxide, where the current electrolyzes the water molecules into hydrogen and oxygen. This sensor will handle moisture up to 1000 ppm and 6 atm pressure. Similar to the aluminum oxide ion, temperature control is very important.

Photometric Moisture Analysis This analyzer requires a light source, a filter wheel rotated by a synchronous motor, a sample cell, a detector to measure the light transmitted, and associated electronics. Water has two absorption bands in the near infrared region at 1400 and 1900 nm. This analyzer can measure moisture in liquid or gaseous samples at levels from 5 ppm up to 100 percent, depending on other chemical species in the sample. Response time is less than 1 s, and samples can be run up to 300°C and 400 psig.

OTHER TRANSDUCERS

Gear Train Rotary motion and angular position are easily transduced by various types of gear arrangements. A gear train in conjunction with a mechanical counter is a direct and effective way to obtain a digital readout of shaft rotations. The numbers on the counter can mean anything desired, depending on the gear ratio and the actuating device used to turn the shaft. A pointer attached to a gear train can be used to indicate a number of revolutions or a small fraction of a revolution for any specified pointer rotation.

Differential Transformer These devices produce an AC electrical output from linear movement of an armature. They are very versatile in that they can be designed for a full range of output with any range of armature travel up to several inches. The transformers have one or two primaries and two secondaries connected to oppose each other. With an AC voltage applied to the primary, the output voltage depends on the position of the armature and the coupling. Such devices produce accuracies of 0.5 to 1.0 percent of full scale and are used to transmit forces, pressures, differential pressures, or weights up to 1500 m. They can also be designed to transmit rotary motion.

Hall-Effect Sensors Some semiconductor materials exhibit a phenomenon in the presence of a magnetic field which is adaptable to sensing devices. When a current is passed through one pair of wires attached to a semiconductor, such as germanium, another pair of wires properly attached and oriented with respect to the semiconductor will develop a voltage proportional to the magnetic field present and the current in the other pair of wires. Holding the exciting current constant and moving a permanent magnet near the semiconductor produce a voltage output proportional to the movement of the magnet. The magnet may be attached to a process-variable measurement device which moves the magnet as the variable changes. Hall-effect devices provide high speed of response, excellent temperature stability, and no physical contact.

SAMPLING SYSTEMS FOR PROCESS ANALYZERS

The sampling system consists of all the equipment required to present a process analyzer with a clean representative sample of a process stream and to dispose of that sample. When the analyzer is part of an automatic control loop, the reliability of the sampling system is as important as the reliability of the analyzer or the control equipment. Sampling systems have several functions. The sample must be withdrawn from the process, transported, conditioned, introduced into the analyzer, and disposed. Probably the most common problem in sample-system design is the lack of realistic information concerning the properties of the process material at the sampling point. Another common problem is the lack of information regarding the conditioning required so that the analyzer may utilize the sample without malfunction for long periods of time. Some samples require enough conditioning and treating that the sampling systems become equivalent to miniature online processing plants. These systems possess many of the same fabrication, reliability, and operating problems as small-scale pilot plants except that the sampling system must generally operate reliably for much longer periods of time.

Selecting the Sampling Point The selection of the sampling point is based primarily on supplying the analyzer with a sample whose composition or physical properties are pertinent to the control function to be performed. Other considerations include selecting locations that provide representative homogeneous samples with minimum transport delay, locations that collect a minimum of contaminating material, and locations that are accessible for test and maintenance procedures.

Sample Withdrawal from Process A number of considerations are involved in the design of sample-withdrawal devices that will provide representative samples. For example, in a horizontal pipe that conveys process fluid, a sample point on the bottom of the pipe will collect a maximum amount of rust, scale, or other solid materials being carried along by the process fluid. In a gas stream, such a location will also collect a maximum amount of liquid contaminants. A sample point on the top side of a pipe will, for liquid streams, collect a maximum amount of vapor contaminants being carried along. Bends in the piping that produce swirls or cause centrifugal concentration of the denser phase may cause maximum contamination to be at unexpected locations. Two-phase process materials are difficult to sample for a total-composition representative sample.

A typical method for obtaining a sample of process fluid well away from vessel or pipe walls is an eduction tube inserted through a packing gland. This sampling method withdraws liquid sample and vaporizes it for transporting to the analyzer location. The transport lag time from the end of the probe to the vaporizer is minimized by using tubing having a small internal volume compared with pipe and valve volumes.

This sample probe may be removed for maintenance and reinstalled without shutting down the process. The eduction tube is made of material that will not corrode so that it will slide through the packing gland even after long periods of service. There may be a small amount of process-fluid leakage until the tubing is withdrawn sufficiently to close the gate valve. A swaged ferrule on the end of the tube prevents accidental ejection of the eduction tube prior to removal of the packing gland. The section of pipe surrounding the eduction tube and extending into the process vessel provides mechanical protection for the eduction tube.

Sample Transport Transport time, the time elapsed between sample withdrawal from the process and its introduction into the analyzer, should be minimized, particularly if the analyzer is an automatic analyzer-controller. Any sample-transport time in the analyzer-controller loop must be treated as equivalent to process dead time in determining conventional feedback controller settings or in evaluating controller performance. Reduction in transport time usually means transporting the sample in the vapor state.

Design considerations for sample-lines are as follows:

1. The structural strength or protection must be compatible with the area through which the sample line runs.
2. Line size and length must be small enough to meet transport-time requirements without excessive pressure drop or excessive bypass of sample at the analyzer input.
3. Line size and internal-surface quality must be adequate to prevent clogging by the contaminants in the sample.
4. The prevention of a change of state of the sample may require insulation, refrigeration, or heating of the sample line.
5. Sample-line material must be such as to minimize corrosion due to sample or environment.

Sample Conditioning Sample conditioning usually involves the removal of contaminants or some deleterious component from the sample mixture and/or the adjustment of temperature, pressure, and flow rate of the sample to values acceptable to the analyzer. Some of the more common contaminants that must be removed are rust, scale, corrosion products, deposits due to chemical reactions, and tar. In sampling some process streams, the material to be removed may include the primary-process product such as polymer or the main constituent of the stream such as oil. In other cases, the material to be removed is present in trace quantities. For example, water in an online chromatograph sample can damage the chromatographic column packing. When contaminants or other materials that will hinder analysis represent a large percentage of the stream composition, their removal may significantly alter the integrity of the sample. In some cases, removal must be done as part of the analysis function so that removed material can be accounted for. In other cases, proper calibration of the analyzer output will suffice.

TELEMETERING AND TRANSMISSION

ANALOG SIGNAL TRANSMISSION

Modern control systems permit the measurement device, the control unit, and the final actuator to be physically separated by several hundred meters, if necessary. This requires the transmission of the measured variable from the measurement device to the control unit, and the transmission of the controller output from the control unit to the final actuator.

In each case, transmission of a single value in only one direction is required. Such requirements can be met by analog signal transmission. A span is defined for the value to be transmitted, and the value is basically transmitted as a percent of this span. For the measured variable, the logical span is the measurement span. For the controller output, the logical span is the range of the final actuator (e.g., valve fully closed to valve fully open).

For pneumatic transmission systems, the signal range used for the transmission is 3 to 15 psig. In each pneumatic transmission system, there can be only one transmitter, but there can be any number of receivers. When most measurement devices were pneumatic, pneumatic transmission was the logical choice. However, with the displacement of pneumatic measurement devices by electronic devices, pneumatic transmission is becoming less common but is unlikely to totally disappear.

In order for electronic transmission systems to be less susceptible to interference from magnetic fields, current is used for the transmission signal instead of voltage. The signal range is 4 to 20 milliamps. In each circuit or “current loop,” there can be only one transmitter. There can be more than one receiver, but not an unlimited number. For each receiver, a 250 ohm “range resistor” is inserted into the current loop, which provides a 1- to 5-volt input to the receiving device. The number of receivers is limited by the power available from the transmitter.

Both pneumatic and electronic transmission use a “live zero.” This enables the receiver to distinguish a transmitted value of zero percent of span from a transmitter or transmission system failure. Transmission of zero percent of span provides a signal of 4 milliamps in electronic transmission. Should the transmitter or the transmission system fail (i.e., an open circuit in a current loop), the signal level would be zero milliamps.

For most measurement variable transmissions, the lower range of the measurement span corresponds to 4 milliamps and the upper range of the measurement span corresponds to 20 milliamps. On an open circuit, the measured variable would fail to its lower range. In some applications, this is undesirable. For example, in a fired heater that is heating material to a target temperature, failure of the temperature measurement to its lower span value would drive the output of the combustion control logic to the maximum possible firing rate. In such applications, the analog transmission signal is normally inverted, with the upper range of the measurement span corresponding to 4 milliamps and the lower range of the measurement span corresponding to 20 milliamps. On an open circuit, the measured variable would fail to its upper range. For the fired heater, failure of the measured variable to its upper span would drive the output of the combustion control logic to the minimum firing rate.

DIGITAL SYSTEMS

With the advent of the microprocessor, digital technology began to be used for data collection, feedback control, and all other information processing requirements in production facilities. Such systems must acquire data from a variety of measurement devices, and control systems must drive final actuators.

Analog Input and Outputs Analog inputs are generally divided into two categories:

1. *High level.* Where the source is a process transmitter, the range resistor in the current loop converts the 4–20 milliamp signal into a 1–5 volt signal. The conversion equipment can be unipolar (i.e., capable of processing only positive voltages).

2. *Low level.* The most common low level signals are inputs from thermocouples. These inputs rarely exceed 30 millivolts, and could be zero or even negative. The conversion equipment must be bipolar (i.e., capable of processing positive and negative voltages).

Ultimately, such signals are converted to digital values via an analog-to-digital (A/D) converter. However, the A/D converter is normally preceded by two other components:

1. *Multiplexer.* This permits one A/D converter to service multiple analog inputs. The number of inputs to a multiplexer is usually between 8 and 256.

2. *Amplifier.* As A/D converters require high level signals, a high gain amplifier is required to convert low-level signals into high-level signals.

One of the important parameters for the A/D converter is its resolution. The resolution is stated in terms of the number of significant binary digits (bits) in the digital value. As the repeatability of most process transmitters is around 0.1 percent, the minimum acceptable resolution for a bipolar A/D converter is 12 bits, which translates to 11 data bits plus one bit for the sign. With this resolution, the analog input values can be represented to 1 part in 2^{11}, or one part in 2048. Normally, a 5-volt input is converted to a digital value of 2000, which effectively gives a resolution of 1 part in 2000 or 0.05 percent. Very few process control systems utilize resolutions higher than 14 bits, which translates to a resolution of 1 part in 8000 or 0.0125 percent.

For 4–20 milliamp inputs, the resolution is not quite as good as stated above. For a 12-bit bipolar A/D converter, 1-volt converts to a digital value of 400. Thus, the range for the digital value is 400 to 2000, making the effective input resolution 1 part in 1600, or 0.0625 percent.

On the output side, dedicated digital-to-analog converters are provided for each analog output. Outputs are normally unipolar, and require a lower resolution than inputs. A 10-bit resolution is normally sufficient, giving a resolution of 1 part in 1000 or 0.1%.

Pulse Inputs Where the sensor within the measurement device is digital in nature, analog-to-digital conversion can be avoided. For rotational devices, the rotational element can be outfitted with a shaft encoder that generates a known number of pulses per revolution. The digital system can process such inputs in either of the following ways:

1. Count the number of pulses over a fixed interval of time.
2. Determine the time for a specified number of pulses.
3. Determine the duration of time between the leading (or trailing) edges of successive pulses.

Of these, the first option is the most commonly used in process applications.

Turbine flowmeters are probably the most common example where pulse inputs are used. Another example is a watt-hour meter. Basically any measurement device that involves a rotational element can be interfaced via pulses.

Occasionally, a nonrotational measurement device can generate pulse outputs. One example is the vortex shedding meter, where a pulse can be generated when each vortex passes over the detector.

Serial Interfaces Some very important measurement devices cannot be reasonably interfaced via either analog or pulse inputs. Two examples are the following:

1. Chromatographs can perform a total composition analysis for a sample. It is possible but inconvenient to provide an analog input for each component. Furthermore, it is often desirable to capture other information, such as the time that the analysis was made (normally the time the sample was injected).

2. Load cells are capable of resolutions of 1 part in 100,000. A/D converters for analog inputs cannot even approach such resolutions.

One approach to interfacing with such devices is serial interfaces. This involves two aspects:

1. *Hardware interface.* The RS-232 interface standard is the basis for most serial interfaces.

2. *Protocol.* This is interpreting the sequence of characters transmitted by the measurement device. There are no standards for protocols, which means that custom software is required.

One advantage of serial interfaces is that two-way communication is possible. For example, a "tare" command can be issued to a load cell.

Microprocessor-Based Transmitters The cost of microprocessor technology has declined to the point where it is economically feasible to incorporate a microprocessor into each transmitter. Such microprocessor-based transmitters are often referred to as "smart" transmitters. As opposed to conventional or "dumb" transmitters, the smart transmitters offer the following capabilities:

1. Checks on the internal electronics, such as verifying that the voltage levels of internal power supplies are within specifications.
2. Checks on environmental conditions within the instruments, such as verifying that the case temperature is within specifications.
3. Compensation of the measured value for conditions within the instrument, such as compensating the output of a pressure transmitter for the temperature within the transmitter. Smart transmitters are much less affected by temperature and pressure variations than conventional transmitters.
4. Compensation of the measured value for other process conditions, such as compensating the output of a capacitance level transmitter for variations in process temperature.
5. Linearizing the output of the transmitter. Functions such as square root extraction of the differential pressure for a head-type flowmeter can be done within the instrument instead of within the control system.
6. Configuring the transmitter from a remote location, such as changing the span of the transmitter output.
7. Automatic recalibration of the transmitter. Although this is highly desired by users, the capabilities, if any, in this respect depend on the type of measurement.

Due to these capabilities, smart transmitters offer improved performance over conventional transmitters.

Transmitter/Actuator Networks With the advent of smart transmitters and smart actuators, the limitations of the 4–20 milliamp analog signal transmission retard the full utilization of the capabilities of the smart devices. For smart transmitters, the following capabilities are required:

1. *Transmission of more than one value from a transmitter.* Information beyond the measured variable is available from the smart transmitter. For example, a smart pressure transmitter can also report the temperature within its housing. Knowing that this temperature is above normal values permits corrective action to be taken before the device fails. Such information is especially important during the initial commissioning of a plant.

2. *Bidirectional transmission.* Configuration parameters such as span, engineering units, resolution, and so on, must be communicated to the smart transmitter.

Similar capabilities are required for smart actuators.

In order to meet their initial requirements, several manufacturers have developed digital communications capabilities for communicating with smart transmitters. These can be used either in addition to or in lieu of the 4–20 milliamp signal. Although most manufacturers release enough information on their communications features to permit another manufacturer to provide compatible instruments (and in some cases provide an open communication standard), the communications capability provided by a manufacturer may be proprietary.

Users purchase their transmitters from a variety of manufacturers, so this situation limits the full utilization of the capabilities of smart transmitters and valves. Efforts to develop a standard for a communications network have not proceeded smoothly. The International Society for Measurement and Control (ISA) has attempted to develop a standard generally referred to as fieldbus. The standards effort attempted to develop a world standard, encompassing European, Japanese, and American products. This effort focused on developing a single standard with which all manufacturers would comply. Currently, efforts are mostly being directed to providing the capability for interoperability between the products of the manufacturers with competing communications networks. Meanwhile, users are reluctant to make major commitments, and are continuing to rely primarily on the traditional 4–20 milliamp transmission.

FILTERING AND SMOOTHING

A signal received from a process transmitter generally contains the following features:

1. Low-frequency process disturbances. The control system is expected to react to these disturbances.
2. High-frequency process disturbances. The frequency of these disturbances is beyond the capability of the control system to effectively react.
3. Measurement noise.
4. Stray electrical pickup, primarily 50- or 60-cycle AC. Frequencies are measured in Hertz (Hz), with 60-cycle AC being a 60-Hz frequency.

The objective of filtering and smoothing is to remove the last three components, leaving only the low frequency process disturbances.

Normally this has to be accomplished using the proper combination of analog and digital filters. Sampling a continuous signal results in a phenomenon often referred to as aliasing or foldover. In order to represent a sinusoidal signal, a minimum of four samples are required during each cycle. That is, the sampling interval must be at least 1/4th the period of the sinusoid. Consequently, when a signal is sampled at a frequency ω_s, all frequencies higher than $(\pi/2)\omega_s$ cannot be represented at their original frequency. Instead, they are present in the sampled signal with their original amplitude but at a lower frequency harmonic.

Because of the aliasing or foldover issues, a combination of analog and digital filtering is usually required. The sampler (i.e., the A/D converter) must be preceded by an analog filter that rejects those high-frequency components such as stray electrical pickup that would result in foldover when sampled. In commercial products, analog filters are normally incorporated into the input processing hardware by the manufacturer. The software then permits the user to specify digital filtering to remove any undesirable low-frequency components.

On the analog side, the filter is often the conventional resistor-capacitor or RC filter. However, other possibilities exist. For example, one type of A/D converter is called an "integrating A/D" because the converter basically integrates the input signal over a fixed interval of time. By making the interval 1/60th second, this approach provides excellent rejection of any 60-Hz electrical noise.

On the digital side, the input processing software generally provides for smoothing via the exponentially weighted moving average, which is the digital counterpart to the RC network analog filter. The smoothing equation is as follows:

$$y_i = \alpha x_i + (1 - \alpha) y_{i-1} \tag{8-103}$$

where x_i = current value of input
y_i = current output from filter
y_{i-1} = previous output from filter
α = filter coefficient

The degree of smoothing is determined by the filter coefficient α, with $\alpha = 1$ being no smoothing and $\alpha = 0$ being infinite smoothing (no effect of new measurements). The filter coefficient α is related to the filter time constant τ_F and the sampling interval Δt by the following equation:

$$\alpha = 1 - \exp\left(\frac{-\Delta t}{\tau_F}\right) \tag{8-104}$$

or by the approximation

$$\alpha = \frac{\Delta t}{\Delta t + \tau_F} \tag{8-105}$$

Another approach to smoothing is to use the arithmetic moving average, which is represented by the following equation:

$$y_i = \frac{\left[\sum_{j=1}^{n} x_{i+1-j}\right]}{n} \tag{8-106}$$

The term "moving" is applied because the filter software maintains a storage array with the previous n values of the input. When a new value is received, the oldest value in the storage array is replaced with the new value, and the arithmetic average recomputed. This permits the filtered value to be updated each time a new input value is received.

In process applications, determining τ_F (or α) for the exponential filter and n for the moving average filter is often done merely by observing the behavior of the filtered value. If the filtered value is "bouncing," the degree of smoothing (that is, τ_F or n) is increased. This can easily lead to an excessive degree of filtering, which will limit the performance of any control system that uses the filtered value. The degree of filtering is best determined from the frequency spectrum of the measured input, but such information is rarely available for process measurements.

ALARMS

The purpose of an alarm is to alert the process operator to a process condition that requires immediate attention. An alarm is said to occur whenever the abnormal condition is detected and the alert is issued. An alarm is said to return to normal when the abnormal condition no longer exists.

Analog alarms can be defined on measured variables, calculated variables, controller outputs, and the like. For analog alarms, the following possibilities exist:

1. *High/low alarms.* A high alarm is generated when the value is greater than or equal to the value specified for the high-alarm limit. A low alarm is generated when the value is less than or equal to the value specified for the low-alarm limit.
2. *Deviation alarms.* An alarm limit and a target are specified. A high deviation alarm is generated when the value is greater than or equal to the target plus the deviation alarm limit. A low deviation alarm is generated when the value is less than or equal to the target minus the deviation alarm limit.
3. *Trend or rate-of-change alarms.* A limit is specified for the maximum rate of change, usually specified as a change in the measured value per minute. A high trend alarm is generated when the rate of change of the variable is greater than or equal to the value specified for the trend alarm limit. A low trend alarm is generated when the rate of change of the variable is less than or equal to the negative of the value specified for the trend alarm limit.

Most systems permit multiple alarms of a given type to be configured for a given value. For example, configuring three high alarms provides a high alarm, a high-high alarm, and a high-high-high alarm.

One operational problem with analog alarms is that noise in the variable can cause multiple alarms whenever its value approaches a limit. This can be avoided by defining a deadband on the alarm. For example, a high alarm would be processed as follows:

1. *Occurrence.* The high alarm is generated when the value is greater than or equal to the value specified for the high-alarm limit.
2. *Return to normal.* The high-alarm return to normal is generated when the value is less than or equal to the high alarm limit less the deadband.

As the degree of noise varies from one input to the next, the deadband must be individually configurable for each alarm.

Discrete alarms can be defined on discrete inputs, limit switch inputs from on/off actuators and so on. For discrete alarms, the following possibilities exist:

1. *Status alarms.* An expected or normal state is specified for the discrete value. A status alarm is generated when the discrete value is other than its expected or normal state.
2. *Change-of-state alarm.* A change-of-state alarm is generated on any change of the discrete value.

The expected sequence of events on an alarm is basically as follows:

1. The alarm occurs. This usually activates an audible annunciator.
2. The alarm occurrence is acknowledged by the process operator. When all alarms have been acknowledged, the audible annunciator is silenced.
3. Corrective action is initiated by the process operator.
4. The alarm condition returns to normal.

However, additional requirements are imposed at some plants. Sometimes the process operator must acknowledge the alarm's return to normal. Some plants require that the alarm occurrence be reissued if the alarm remains in the occurred state longer than a specified period of time. Consequently, some "personalization" of the alarming facilities is done.

When alarms were largely hardware-based (i.e., the panel alarm systems), the purchase and installation of the alarm hardware imposed a certain discipline on the configuration of alarms. With digital systems, the suppliers have made it extremely easy to configure alarms. In fact, it is sometimes easier to configure alarms on a measured value than not to configure the alarms. Furthermore, the engineer assigned the responsibility for defining alarms should ensure that an abnormal process condition will not go undetected because an alarm has not been configured. When alarms are defined on every measured and calculated variable, the result is an excessive number of alarms, most of which are duplicative and unnecessary.

The accident at the Three Mile Island nuclear plant clearly demonstrated that an alarm system can be counterproductive. An excessive number of alarms can distract the operator's attention from the real problem that needs to be addressed. Alarms that merely tell the operator something that is already known do the same. In fact, a very good definition of a nuisance alarm is one that informs the operator of a situation of which the operator is already aware. The only problem with applying this definition is determining what the operator already knows.

Unless some discipline is imposed, engineering personnel, especially where contractors are involved, will define far more alarms than plant operations require. This situation may be addressed by simply setting the alarm limits to values such that the alarms never occur. However, changes in alarms and alarm limits are changes from the perspective of the Process Safety Management regulations. It is prudent to impose the necessary discipline to avoid an excessive number of alarms. Potential guidelines are as follows:

1. For each alarm, a specific action is expected from the process operator. Operator actions such as "call maintenance" are inappropriate with modern systems. If maintenance needs to know, modern systems can inform maintenance directly.
2. Alarms should be restricted to abnormal situations for which the process operator is responsible. A high alarm on the temperature in one of the control system cabinets should not be issued to the process operator. Correcting this situation is the responsibility of maintenance, not the process operator.
3. Process operators are expected to be exercising normal surveillance of the process. Therefore, alarms are not appropriate for situations known to the operator either through previous alarms or through normal process surveillance. The "sleeping operator" problem can be addressed by far more effective means than the alarm system.
4. When the process is operating normally, no alarms should be triggered. Within the electric utility industry, this design objective is known as "darkboard." Application of darkboard is especially important in batch plants, where much of the process equipment is operated intermittently.

Ultimately, guidelines such as those above will be taken seriously only if production management carefully configures the alarms. The consequences of excessive and redundant alarms will be felt primarily by those responsible for production operations. Therefore, production management must make adequate resources available for reviewing and analyzing the proposed alarm configurations.

Another serious distraction to a process operator is the multiple alarm event, where a single event within the process results in multiple alarms. When the operator must individually acknowledge each alarm, considerable time can be lost in silencing the obnoxious annunciator before the real problem is addressed. Air-handling systems are especially vulnerable to this, where any fluctuation in pressure (for example, resulting from a blower trip) can cause a number of pressure alarms to occur. Point alarms (high alarms, low alarms, status alarms, etc.) are especially vulnerable to the multiple alarm event. This can be addressed in one of two ways:

1. *Ganging alarms.* Instead of individually issuing the point alarms, all alarms associated with a certain aspect of the process are simply wired to give a single trouble alarm. The responsibility rests entirely with the operator to determine the nature of the problem.
2. *Intelligent alarms.* Logic is incorporated into the alarm system to determine the nature of the problem and then issue a single alarm to the process operator. Sometimes this is called an expert system.

While the intelligent alarm approach is clearly preferable, substantial process analysis is required to support intelligent alarming. Meeting the following two objectives is quite challenging:

1. The alarm logic must consistently detect abnormal conditions within the process.
2. The alarm logic must not issue an alert to an abnormal condition when in fact none exists.

Often the latter case is more challenging than the former.

Logically, the intelligent alarm effort must be linked to the process hazards analysis. Developing an effective intelligent alarming system requires substantial commitments of effort, involving both process engineers, control systems engineers, and production personnel. Methodologies such as expert systems can facilitate the implementation of an intelligent alarming system, but they must still be based on a sound analysis of the potential process hazards.

DIGITAL TECHNOLOGY FOR PROCESS CONTROL

GENERAL REFERENCES: Fortier, *Design and Analysis of Distributed Real-Time Systems,* McGraw-Hill, New York, 1985; Hawryszkiewycs, *Database Analysis and Design,* Science Research Associates Inc., Chicago, 1984; Khambata, *Microprocessors/Microcomputers: Architecture, Software, and Systems,* 2d ed., Wiley, New York, 1987; Liptak, *Instrument Engineers Handbook,* Chilton Book Company, Philadelphia, 1995; Mellichamp (ed.), *Real-Time Computing with Applications to Data Acquisition and Control,* Van Nostrand Reinhold, New York, 1983.

Since the 1970s, process controls have evolved from pneumatic analog technology to electronic analog technology to microprocessor-based controls. Electronic analog technology has virtually disappeared from process controls. Pneumatic controls continue to be manufactured, but they are relegated to special situations where pneumatics can offer a unique advantage. Process controls are dominated by programmable electronic systems (PES), most of which are based on microprocessor technology.

HIERARCHY OF INFORMATION SYSTEMS

Coupling digital controls with networking technology permits information to be passed from level-to-level within a corporation at high rates of speed. This technology is capable of presenting the measured variable from a flow transmitter installed in a plant in a remote location anywhere in the world to the company headquarters in less than a second.

A hierarchical representation of the information flow within a company leads to a better understanding of how information is passed from one layer to the next. Such representations can be developed in varying degrees of detail, and most companies have developed one that describes their specific practices. The following hierarchy consists of five levels.

Measurement Devices and Actuators Often referred to as level 0, this layer couples the control and information systems to the process. The measurement devices provide information on the current conditions within the process. The actuators permit control decisions to be imposed on the process. Although traditionally analog, smart transmitters and smart valves based on microprocessor technology will eventually dominate this layer.

Regulatory Controls The objective of this layer is to operate the process at or near the targets supplied by others, be it the process operator or a higher layer in the hierarchy. In order to achieve consistent process operations, a high degree of automatic control is required from the regulatory layer. The direct result is a reduction in variance in the key process variables. More uniform product quality is an obvious benefit. However, consistent process operation is a prerequisite for optimizing the process operations. To ensure success for the upper level functions, the first objective of any automation effort must be to achieve a high degree of regulatory control.

Supervisory Controls The regulatory layer blindly attempts to operate the process at the specified targets, regardless of the appropriateness of these targets. Determining the most appropriate targets is the responsibility of the supervisory layer. Given the current production targets for a unit, supervisory control determines how the process can be best operated to meet the production targets. Usually this optimization has a limited scope, being confined to a single production unit or possibly even a single unit operation within a production unit. Supervisory control translates changes in factors such as current process efficiencies, current energy costs, cooling medium temperatures, and so on, to changes in process operating targets so as to optimize process operations.

Production Controls The nature of the production control logic differs greatly between continuous and batch plants. A good example of production control in a continuous process is refinery optimization. From the assay of the incoming crude oil, the values of the various possible refined products, the contractual commitments to deliver certain products, the performance measures of the various units within a refinery, and the like, it is possible to determine the mix of products that optimizes the economic return from processing this crude. The solution of this problem involves many relationships and constraints and is solved with techniques such as linear programming.

In a batch plant, production control often takes the form of routing or short-term scheduling. For a multiproduct batch plant, determining the long term schedule is basically a manufacturing resource planning (MRP) problem, where the specific products to be manufactured and the amounts to be manufactured are determined from the outstanding orders, the raw materials available for production, the production capacities of the process equipment, and other factors. The goal of the MRP effort is the long-term schedule, which is a list of the products to be manufactured over a specified period of time (often one week). For each product on the list, a target amount is also specified. To manufacture this amount usually involves several batches. The term "production run" often refers to the sequence of batches required to make the target amount of product, so in effect the long term schedule is a list of production runs.

Most multiproduct batch plants have more than one piece of equipment of each type. Routing refers to determining the specific pieces of equipment that will be used to manufacture each run on the long term production schedule. For example, the plant might have five reactors, eight neutralization tanks, three grinders, and four packing machines. For a given run, a rather large number of possible routes are possible. Furthermore, rarely is only one run in progress at a given time. The objective of routing is to determine the specific pieces of production equipment to be used for each run on the long-term production schedule. Given the dynamic nature of the production process (equipment failures, insertion/deletion of runs into the long-term schedule, etc.), the solution of the routing problem continues to be quite challenging.

Corporate Information Systems Terms such as *management information systems* (MIS) and *information technology* (IT) are frequently used to designate the upper levels of computer systems within a corporation. From a control perspective, the functions performed at this level are normally long-term and/or strategic. For example, in a processing plant, long-term contracts are required with the providers of the feedstocks. A forecast must be developed for the demand for possible products from the plant. This demand must be translated into needed raw materials, and then contracts executed with the suppliers to deliver these materials on a relatively uniform schedule.

While most companies within the process industries recognize the importance of information technology in managing their businesses, this technology has been a source of considerable frustration and disappointment. Schedule delays, cost overruns, and failure of the final product to perform as expected have often eroded the credibility of information technology. However, immense potential remains for the technology, and process companies have no choice but to seek continuous improvement.

DISTRIBUTED CONTROL SYSTEMS

Although digital control technology was first applied to process control in 1959, the total dependence of the early centralized architectures on a single computer for all control and operator interface functions resulted in complex systems with dubious reliability. Adding a second processor increased both the complexity and the cost. Consequently, many installations provided analog backup systems to protect against a computer malfunction.

Microprocessor technology permitted these technical issues to be addressed in a cost-effective manner. In the mid-1970s, a process control architecture referred to as a distributed control system (DCS) was introduced and almost instantly became a commercial success. A DCS consists of some number of microprocessor-based nodes that are interconnected by a digital communications network, often called a data highway. The key features of this architecture are as follows:

1. The process control functions and the operator interface, also referred to as man-machine interface (MMI) or human-machine interface (HMI), is provided by separate nodes. This approach is referred to as *split-architecture,* and it permits considerable flexibility in choosing a configuration that most appropriately meets the needs of the application.

2. The process control functions can be distributed functionally and/or geographically. Functional distribution permits related control functions to be grouped and implemented in a single node. Geographical distribution permits the process control nodes to be physically located near the equipment being controlled. As the digital communications network is based on local area network (LAN) technology, the nodes within the DCS can be physically separated by thousands of meters.

3. Redundancy can be provided where appropriate, the following being typical:

a. Multiple operator interface nodes can be provided to reduce the impact of an operator interface node failure.

b. The digital communications network is normally redundant to the extent that at least two independent paths are available between any two nodes of the DCS.

c. Consisting basically of processor and memory, the process control nodes are highly reliable, with mean-times-between-failures approaching 100 years. Redundant configurations are available for especially critical applications.

4. As the data within the DCS are digital in nature, interfaces to upper level computers are technically easier to implement. Unfortunately, the proprietary nature of the communications networks within commercial DCS products complicate the implementation of such interfaces. Truly open DCS architectures, at least as the term "open" is used in the mainstream of computing, are not yet available.

Figure 8-62 depicts a hypothetical distributed control system. A number of different unit configurations are illustrated. This system consists of many commonly used DCS components, including multiplexers (MUXs), single/multiple-loop controllers, programmable logic controllers (PLCs), and smart devices. A typical system includes the following elements as well:

- *Host computers.* These are the most powerful computers in the system, capable of performing functions not normally available in other units. They act as the arbitrator unit to route internodal communications. An operator interface is supported and various peripheral devices are coordinated. Computationally intensive tasks, such as optimization or advanced control strategies, are processed here.
- *Data highway.* This is the communication link between com-

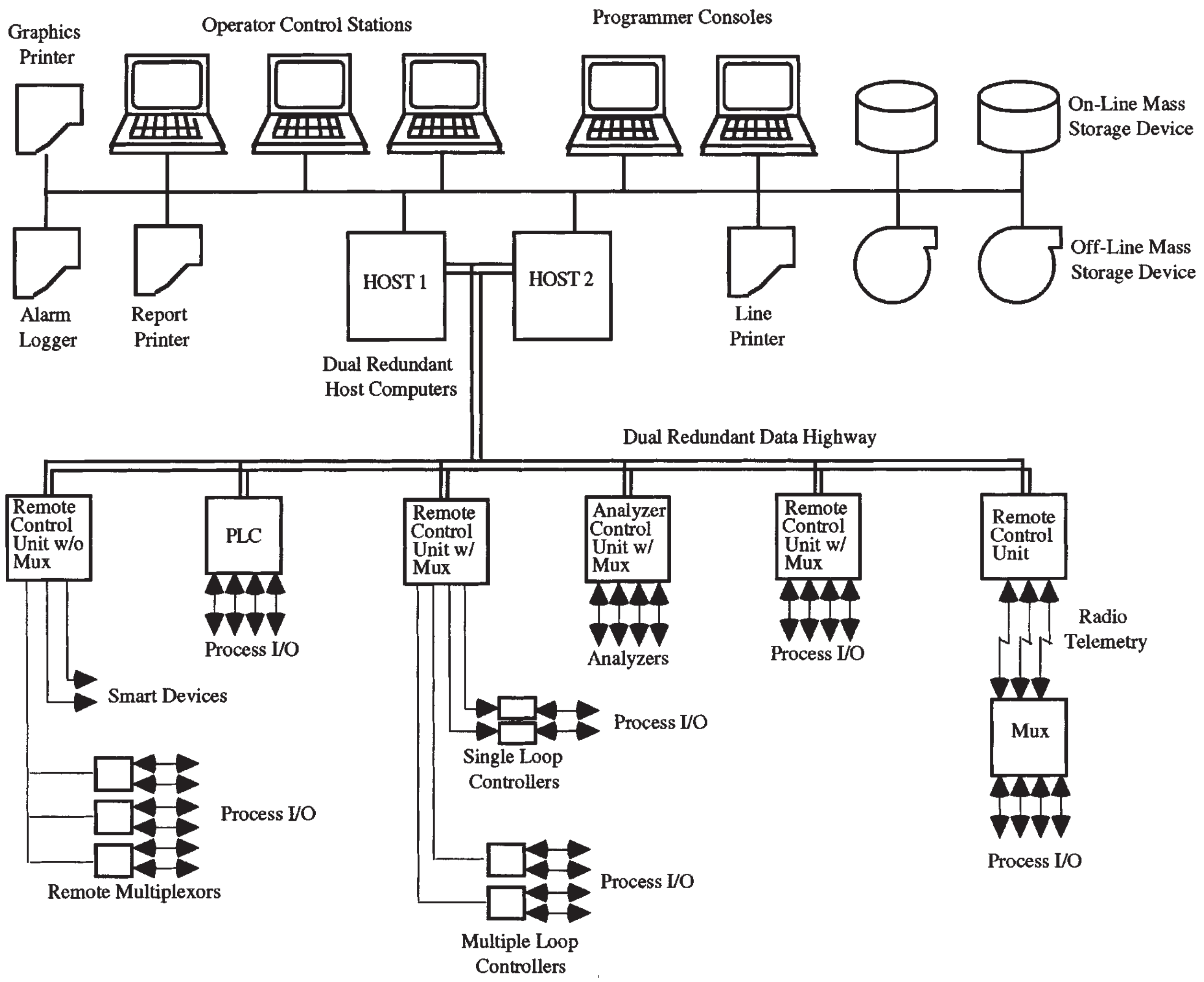

Graphics Printer
Operator Control Stations
Programmer Consoles
On-Line Mass Storage Device
Alarm Logger
Report Printer
HOST 1
HOST 2
Line Printer
Off-Line Mass Storage Device
Dual Redundant Host Computers
Dual Redundant Data Highway
Remote Control Unit w/o Mux
PLC
Remote Control Unit w/ Mux
Analyzer Control Unit w/ Mux
Remote Control Unit w/ Mux
Remote Control Unit
Process I/O
Analyzers
Process I/O
Radio Telemetry
Smart Devices
Process I/O
Single Loop Controllers
Mux
Process I/O
Remote Multiplexors
Process I/O
Process I/O
Multiple Loop Controllers

FIG. 8-62 A typical DDC system.

ponents of a network. Coaxial cable is often used. A redundant pair is normally supplied to reduce possibility of link failure.

- *Real-time clocks (RTCs).* Real-time systems are required to respond to events, as they occur, in a timely manner. This is especially crucial in process control systems where control actions applied at the wrong time may amplify process deviations or destabilize the processes. The nodes in the systems are interrupted periodically by the real-time clocks to maintain the actual elapsed times.
- *Operator control stations.* These typically consist of color graphics monitors with special keyboards, in addition to a conventional alphanumeric keyboard, containing keys to perform dedicated functions. Operators may supervise and control processes from these stations. A control station may contain a number of printers for alarm logging, report printing, or hard-copying process graphics.
- *Remote control units.* These units are used to control unit processes. Basic control functions such as the PID algorithm are implemented here. Depending on other hardware components used, data acquisition capability may be required to perform digital control. They may be configured to supply process set points to single-loop controllers. Radio telemetry may be installed to communicate with MUX units located at great distances.
- *Programmer consoles.* These are programming terminals. Developing system software on the host machines is a common practice by many system suppliers. This eliminates compatibility problems between development and target environments. Programming capability is normally retained when the system is delivered such that system users may develop their own application programs.
- *Mass storage device.* Typically, fixed-head hard disk drives are used to store active data, including on-line and historical databases and non-memory-resident programs. Memory-resident programs are stored to allow loading at system startups. The tape drives are used for archives and backups.

DISTRIBUTED DATABASE AND THE DATABASE MANAGER

A database is a centralized location for data storage. The use of databases enhances system performance by maintaining complex relations between data elements while reducing data redundancy. A database may be built based on the relational model, the entity relationship model, or some other model. The database manager is a system utility program or programs acting as the gatekeeper to the databases. All functions retrieving or modifying data must submit a request to the manager. Information required to access the database include the tag name of the database entity, often referred to as a point, the attributes to be accessed, and the values if modifying. The database manager maintains the integrity of the databases by executing a request only when not processing other conflicting requests. Although a number of functions may read the same data item at the same time, writing by a number of functions or simultaneous read and write of the same data item is not permitted.

To allow flexibility, the database manager must also perform point addition or deletion. However, the ability to create a point type or to add or delete attributes of a point type is not normally required because, unlike other data processing systems, a process control system normally involves a fixed number of point types and related attributes. For example, analog and binary input and output types are required for process I/O points. Related attributes for these point types include tag names, values, and hardware addresses. Different system manufacturers may define different point types using different data structures. We will discuss other commonly used point types and attributes as they appear.

Historical Database Subsystem We have discussed the use of on-line databases. An historical database is built similar to an on-line database. Unlike their on-line counterparts, the information stored in a historical database is not normally accessed directly by other subsystems for process control and monitoring. Periodic reports and long-term trends are generated based on the archived data. The reports are often used for long-term planning and system performance evaluations such as statistical process (quality) control. The trends may be used to detect process drifts or to compare process variations at different times.

The historical data is sampled at user-specified intervals. A typical process plant contains a large number of data points, but it is not feasible to store data for all points at all times. The user determines if a data point should be included in the list of archive points. Most systems provide archive-point menu displays. The operators are able to add or delete data points to the archive point lists. The sampling periods are normally some multiples of their base scan frequencies. However, some systems allow historical data sampling of arbitrary intervals. This is necessary when intermediate virtual data points that do not have the scan frequency attribute are involved. The archive point lists are continuously scanned by the historical database software. On-line databases are polled for data. The times of data retrieval are recorded with the data obtained. To conserve storage space, different data compression techniques are employed by various manufacturers.

The historical data may be used for long-term trending. The live trends data are displayed but not stored. Therefore, these trends cannot be recalled once cleared off the screens. The historical trend of any archive point may be displayed at any time because the values used are extracted from the archived data. Zooming, that is, axis scaling, is allowed by most systems. As a result, the displayed data point intervals may not be multiples of stored data intervals. Many systems provide data interpolation and smoothing functions to process the stored data when they are displayed. The live and historical trend displays are superior to strip charts in many ways. In addition to conventional trend recording, the zoom-in capability allows close examination of recorded data, whereas zoom-out compresses long-term data within a screen. Exact data sampled at any time point can be extracted by cursor positioning. Strip-chart recorders have disappeared from many modern plants.

Periodic reports, including shift, daily, weekly, monthly, and quarterly reports, are printed based on archived data. Some reports may contain simply the stored data in certain specific arrangements. More often, quantities such as mean values, standard deviations, or other calculated values are included. Instead of hard-coding reports to user specifications, many system suppliers provide report generation packages in the form of metalanguages. These packages allow users to configure report formats suitable for their particular requirements. The report generator interprets the configuration files prepared by the users to create reports. Due to the infrequent execution, the report generator is normally operated in the batch mode.

DCS manufacturers have devoted considerable efforts to make it easy to implement and enhance process control configurations within their products. Although programming in the traditional sense is possible within most products, the majority of the functions required for a process control application can be implemented by configuring as opposed to programming.

PROCESS CONTROL LANGUAGE

A digital control system involves software development. The introduction of high-level programming languages such as FORTRAN and BASIC in the 1960s was considered a major breakthrough. For process control applications, some companies have incorporated libraries of software routines for these languages, but others have developed speciality pseudolanguages. These implementations are characterized by their statement-oriented language structure. Although substantial savings in time and efforts can be realized, software development costs continue to be significant.

The most successful and user-friendly approach, which is now adopted by virtually all commercial systems, is the fill-in-the-forms or table-driven process control languages (PCLs). The core of these languages is a number of basic functional blocks or software modules. All modules are defined as database points. Using a module is analogous to calling a subroutine in conventional programs.

In general, each module contains some inputs and an output. The programming involves soft-wiring outputs of blocks to inputs of other blocks. Some modules may require additional parameters to direct module execution. The users are required to fill in the sources of input values, the destinations of output values, and the parameters in blanks of forms or tables prepared for the modules. The source and destination blanks may be filled with process I/Os when appropriate. To connect modules, some systems require filling the tag names of modules originating or receiving data. Additional programs are often required to resolve ambiguities when connecting multiple input-output modules. Another method involves the use of intermediate data points. The blanks in a pair of interconnecting modules are filled with the tag name of the same data point. Batch jobs and/or interactive data entry may be performed to fill the databases. A completed control strategy resembles a data flow diagram. The soft-wiring of modules is similar to hard-wiring analog-electronic circuits in analog computers.

Additional database space must be allocated when intermediate data points are used. A system can be designed to use process I/O points as intermediates. However, the data acquisition software must be programmed to bypass these points when scanned. All system builders provide virtual data point types if the intermediate data storage scheme is adopted. These points are not scanned by the data acquisition software. Memory space requirements are reduced by eliminating unnecessary attributes such as hardware addresses and scan frequencies. It should be noted that the fill-in-the-forms technique is applicable to all data point types.

All process control languages contain PID controller blocks. The digital PID controller is normally programmed to execute in velocity form. A pulse duration output may be used to receive the velocity output directly. Where positional signal is expected, an operating mode bit is used to enable an internal integrator. This flexibility is not normally available in analog controllers. Unlike an analog controller, the three modes in a digital PID controller do not interact. This simplifies the tuning effort. In addition to the tuning constants, a typical digital PID controller contains some entries not normally found in an analog controller:

- When a process error is below certain tolerable deadband, the controller ceases modifying output. This is referred to as gap action.
- The magnitude of change in a velocity output is limited by a change clamp.
- A pair of output clamps is used to restrict a positional output value from exceeding specified limits.
- The controller action can be disabled by triggering a binary deactivate input signal, during process startup, shutdown, or when some abnormal conditions exist.

Although modules are supplied and their internal configurations are different from system to system, their basic functionalities are the same.

SINGLE-LOOP CONTROLS

With the exception of pneumatic controllers for special applications, commercial single-loop controllers are almost entirely microprocessor-based. The most basic products provide only the PID control algorithm, but the more powerful versions provide a set of general-purpose algorithms comparable to those in a DCS. For applications such as cascade control and multivariable control, the manufacturers of single-loop controllers provide multiloop versions of their products. These multiloop controllers have much in common with the process control node in a DCS.

Single-loop controllers provide both the process control functions and the operator interface function. This makes them ideally suited to very small applications, where only two or three loops are required. However, it is possible to couple single-loop controllers to a personal computer (PC) to provide the operator interface function. Such installations are extremely cost effective, and with the keen competition in PC-based products, the capabilities are comparable and sometimes even better than that provided by a DCS. However, this approach makes sense only up to about 25 loops.

Initially, the microprocessor-based single-loop controllers made the power of digital control affordable to those with small processes. To compete with these products in small applications, the DCS suppliers have introduced micro-DCS versions of their products. As a PC-based operator interface is usually a component of the micro-DCS, there is sometimes little distinction between a micro-DCS and a system consisting of single-loop controllers coupled to a PC-based operator interface.

PROGRAMMABLE LOGIC CONTROLLERS

The programmable logic controller (PLC) was the first digital technology to successfully compete with conventional technology in industrial control applications. Initially developed in the early 1970s for applications within the manufacturing industries (principally automotive), the PLC proved to be superb for implementing discrete logic. The earliest PLCs were limited to discrete I/O, basic Boolean logic functions (AND, OR, NOT), timers, and counters. However, versions soon appeared with analog I/O, math functions, PID control algorithms, and other functions required for process control applications.

Developed to replace hard-wired relay logic, the early PLCs were "programmed" using the same ladder logic diagrams used to represent logic implemented with hard-wired relays. As the initial target market was electrical, programming in ladder logic was a definite advantage, and some union contracts specifically required that such discrete logic be presented as ladder diagrams. However, ladder logic is not the programming medium preferred by instrument engineers, which hampers the acceptance of PLCs for process control. Alternatives to ladder logic are available for programming PLCs, but established perceptions are slow to change.

Developed specifically for implementing discrete logic, PLCs continue to provide the best route to implementing such logic. The manufacturers of PLCs provide robust, cost-effective discrete I/O modules. Regardless of its acceptability, ladder logic is the most efficient means for implementing discrete logic. Because PLCs scan the discrete logic very rapidly, a 100-millisecond scan rate is considered very slow for a PLC. The process control modules of a DCS often implement discrete logic using function blocks, which is less efficient than ladder logic and normally results in a slower scan rate. A few DCS process control modules have used ladder logic to implement discrete logic, but their discrete I/O capabilities and slow scan rates rarely match that of a PLC. Consequently, for applications heavy with discrete logic, most DCS suppliers will incorporate one or more PLCs into their system.

Being excellent at discrete logic, PLCs are a potential candidate for implementing interlocks. Process interlocks are clearly acceptable for implementation within a PLC. Implementation of safety interlocks in programmable electronic systems (such as a PLC) is not universally accepted. Many organizations continue to require that all safety interlocks be hard-wired, but implementing safety interlocks in a PLC that is dedicated to safety functions is accepted by some as being equivalent to the hard-wired approach.

INTERCOMPUTER COMMUNICATIONS

A group of computers becomes a network when intercomputer communication is established. Prior to the 1980s, all system suppliers used proprietary protocols to network their systems. Ad hoc approaches were sometimes used to connect third-party equipment, which was not cost-effective in system maintenance, upgrade, and expansion. The recent introduction of standardized protocols has led to a decrease in initial capital cost. Most current DCS network protocol designs are based on the ISO-OSI* seven-layer model.

The most notable effort in standardizing plant automation protocols

* Abbreviated from International Standards Organization-Open System Interconnection. They are the physical, data link, network, transports session, presentation, and application layers. Only the physical, data link, and application layers are present in the mini-MAP.

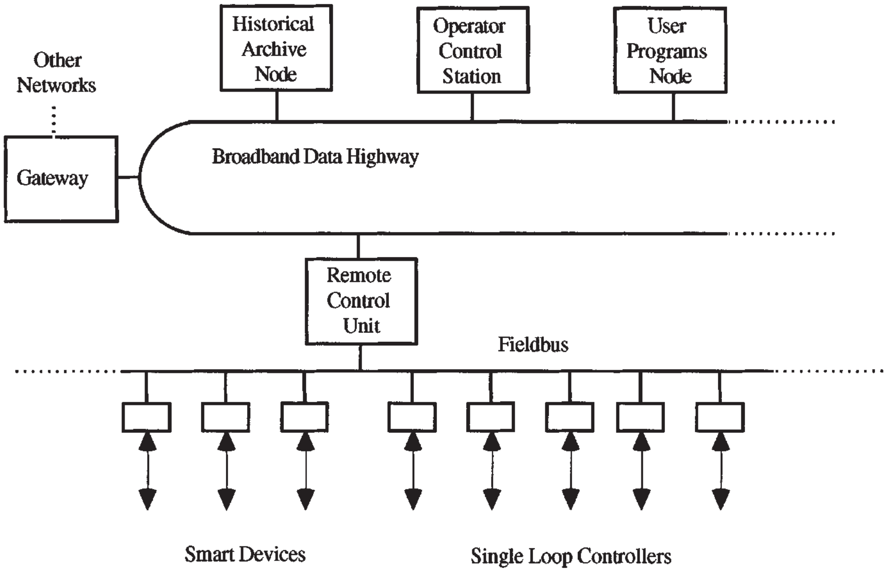


FIG. 8-63 A DCS using broadband data highway and fieldbus.

was initiated by General Motors in the early 1980s. This culminated in the Manufacturing Automation Protocol (MAP), which adopted the ISO-OSI standards as its basis. MAP specifies a broadband backbone local area network (LAN), which incorporates a selection of existing standard protocols suitable for manufacturing automation while defining additional protocols where no standard previously existed. Although intended for discrete component systems, MAP has evolved to address the integration of DCSs used in process control as well. Due to various technical reasons, MAP has gained limited acceptance by the process industries as of 1990. Engineering societies, including ISA and IEEE, and many operating companies are collaborating to refine MAP for wider support.

More microprocessor-based process equipment, such as smart instruments and single-loop controllers, with digital communications capability are now becoming available and are used extensively in process plants. A fieldbus, which is a low-cost protocol, is necessary to perform efficient communication between the DCS and these devices. So-called mini-MAP architecture was developed to satisfy process control and instrumentation requirements while incorporating existing ISA standards. It is intended to improve access time while allowing communications to a large number of microprocessor-based devices. The mini-MAP contains only three of the seven layers specified by the ISO-OSI model; therefore, a mini-MAP device cannot communicate with devices on the MAP bus directly. The development of MAP/EPA (Enhanced Performance Architecture) is parallel to that of the mini-MAP. This scheme adopts the full seven-layer model with a reduced set of MAP protocols. The MAP/EPA is compatible to both the complete MAP and the mini-MAP. Another benefit of standardizing the fieldbus is that it encourages third-party traditional equipment manufacturers to enter the smart equipment market, resulting in increased competition and improved equipment quality. Figure 8-63 illustrates a LAN-based DCS.

Irrespective of the protocol used, communication programs act as servers to the database manager. When some functions request data from a remote node, the database manager will transfer the request to the remote node database manager via the communication programs. The remote node communication programs will relay the request to the resident database manager and return the obtained data. The remote database access and the existence of communications equipment and software are transparent to plant operators.

CONTROLLERS, FINAL CONTROL ELEMENTS, AND REGULATORS

General References: Baumann, *Control Valve Primer,* 2d ed., ISA, 1994; Considine, *Process/Industrial Instruments & Controls Handbook,* 4th ed., McGraw-Hill, 1993; Driskell, *Control-Valve Selection and Sizing,* ISA, 1983; Hammitt, *Cavitation and Multiphase Flow Phenomena,* McGraw-Hill, 1980; Norton, *Fundamentals of Noise and Vibration Analysis for Engineers,* Cambridge University Press, 1989; Ulanski, *Valve and Actuator Technology,* McGraw-Hill, 1991.

External control of the process is achieved by devices that are specially designed, selected and configured for the intended process-control application. The text below covers three very common function classifications of process-control devices: controllers, final control elements, and regulators.

The process controller is the “master” of the process-control system. It accepts a set point and other inputs and generates an output or outputs that it computes from a rule or set of rules that are part of its internal configuration. The controller output serves as an input to another controller or, more often, as an input to a final control element. The final control element is the device that affects the flow in the piping system of the process. The final control element serves as an interface between the process controller and the process. Control valves and adjustable speed pumps are the principal types discussed.

Regulators, though not controllers or final control elements, perform the combined function of these two devices (controller and final control element) along with the measurement function commonly associated with the process variable transmitter. The uniqueness, control performance, and widespread usage of the regulator make it deserving of a functional grouping of its own.

ELECTRONIC AND PNEUMATIC CONTROLLERS

Electronic Controllers Almost all of the electronic process controllers used today are microprocessor-based devices. These processor-based controllers contain, or have access to, input/output (I/O) interface electronics that allow various types of signals to enter and leave the controller’s processor. The controller, depending on its type, uses sufficient read-only-memory (ROM) and read/write-accessible-memory (RAM) to perform the controller function.

The resolution of the analog I/O channels of the controller vary somewhat, with 12-bit and 14-bit conversions quite common. Sample rates for the majority of the constant sample rate controllers range from 1 to 10 samples/second. Hard-wired single-pole, low-pass filters are installed on the analog inputs to the controller to protect the sampler from aliasing errors.

Distributed Control Systems Some knowledge of the distributed control system (DCS) is useful in understanding electronic controllers. A DCS is a process control system with sufficient performance to support large-scale real-time process applications. The DCS has (1) an operations workstation with a cathode ray tube (CRT) for display; (2) a controller subsystem that supports various types of controllers and controller functions; (3) an I/O subsystem for transducing data; (4) a higher-level computing platform for performing process supervision, information processing, and analysis; and (5) communication networks to tie the DCS subsystems, plant areas, and other plant systems together.

The component controllers used in the controller subsystem portion of the DCS can be of various types and include multiloop controllers, programmable logic controllers, personal computer controllers, single-loop controllers, and fieldbus controllers. The type of electronic controller utilized depends on the size and functional characteristic of the process application being controlled. See the earlier section on distributed control systems.

Multiloop Controllers The multiloop controller is a DCS network device that uses a single 32-bit microprocessor to provide control functions to many process loops. The controller operates independent of the other devices on the DCS network and can support from 20 to 500 loops. Data acquisition capability for up to 1000 analog and discrete I/O channels or more can also be provided by this controller.

The multiloop controller contains a variety of function blocks (for example, PID, totalizer, lead/lag compensator, ratio control, alarm, sequencer, and Boolean) that can be “soft-wired” together to form complex control strategies. The multiloop controller, as part of a DCS, communicates with other controllers and man/machine interface (MMI) devices also on the DCS network.

Programmable Logic Controllers The programmable logic controller (PLC) originated as a solid-state replacement for the hard-wired relay control panel and was first used in the automotive industry for discrete manufacturing control. Today, PLCs are used to implement Boolean logic functions, timers, counters, and some math functions and PID control. PLCs are often used with on/off process control valves.

PLCs are classified by the number of the I/O functions supported. There are several sizes available, with the smallest PLCs supporting less than 128 I/O channels and the largest supporting over 1023 I/O channels. I/O modules are available that support high-current motor loads, general-purpose voltage and current loads, discrete inputs, analog I/O and special-purpose I/O for servomotors, stepping motors, high-speed pulse counting, resolvers, decoders, multiplexed displays, and keyboards. PLCs often come with lights or other discrete indicators to determine the status of key I/O channels.

When used as an alternative to a DCS, the PLC is programmed with a handheld or computer-based loader. The PLC is typically programmed with basic ladder logic or a high-level computer language such as BASIC, FORTRAN, or C. Programmable logic controllers use 16- or 32-bit microprocessors and offer some form of point-to-point communications such as RS-232C, RS-422, or RS-485.

Personal Computer Controller Because of its high performance at low cost and its unexcelled ease of use, application of the personal computer (PC) as a platform for process controllers is growing. When configured to perform scan, control, alarm, and data acquisition (SCADA) functions and combined with a spreadsheet or database management application, the PC controller can be a low-cost, basic alternative to the DCS or PLC.

Using the PC for control requires installation of a board into the expansion slot in the computer, or the PC can be connected to an external I/O module using a standard communication port on the PC (RS-232, RS-422, or IEEE-488). The controller card/module supports 16- or 32-bit microprocessors. Standardization and high volume in the PC market has produced a large selection of hardware and software tools for PC controllers.

Single-Loop Controller The single-loop controller (SLC) is a process controller that produces a single output. SLCs can be pneumatic, analog electronic, or microprocessor-based. Pneumatic SLCs are discussed in the pneumatic controller section, and analog electronic SLC is not discussed because it has been virtually replaced by the microprocessor-based design.

The microprocessor-based SLC uses an 8- or 16-bit microprocessor with a small number of digital and analog process input channels with control logic for the I/O incorporated within the controller. Analog inputs and outputs are available in the standard ranges (1–5 volts DC and 4–20 mA DC). Direct process inputs for temperature sensors (thermistor RTD and thermocouple types) are available. Binary outputs are also available. The face of the SLC has some form of visible display and pushbuttons that are used to view or adjust control values and configuration. SLCs are available for mounting in panel openings as small as 48 mm by 48 mm (1.9 by 1.9 inches).

The processor based SLC allows the operator to select a control strategy from a predefined set of control functions. Control functions include PID, on/off, lead/lag, adder/subtractor, multiply/divider, filter functions, signal selector, peak detector, and analog track. SLCs feature auto/manual transfer switching, multi-setpoint, self diagnostics, gain scheduling, and perhaps also time sequencing. Most processor-based SLCs have self-tuning PID control algorithms.

Sample times for the microprocessor-based SLCs vary from 0.1 to 0.4 seconds. Low-pass analog electronic filters are installed on the process inputs to stop aliasing errors caused by fast changes in the process signal. Input filter time constants are typically in the range from 0.1 to 1 s. Microprocessor-based SLCs may be made part of a DCS by using the communication port (RS-488 is common) on the controller or may be operated in a standalone mode independent of the DCS.

Fieldbus Controller The benefits of eliminating all analog communication links to and from the devices in the process loop (including final control elements and measurement transmitters) have stimulated considerable interest in standardizing a suitable digital fieldbus communication network. Although a universal network standard is not currently complete (see "Digital Field Communications" in this section), several manufacturers have made available field devices that feature basic process-controller functionality. These controllers, known as fieldbus controllers, reside in the final control element or measurement transmitter and are considered to be an option available with these control devices. A suitable communications modem is present in the device to interface with a proprietary, PC-based, or hybrid analog/digital bus network.

Presently, fieldbus controllers are single-loop controllers with 8- and 16-bit microprocessors and are options to digital field-control devices. These controllers support the basic PID control algorithm and are projected to increase in functionality as the controller market develops. Parameters relating to intrinsic safety, communication type and bit rate, level of DCS support, and ultimate controller performance differentiate currently available fieldbus controllers.

Controller Reliability and Application Trends Critical process-control applications demand a high level of reliability from the electronic controller. Some methods that improve the reliability of electronic controllers include: (1) focusing on robust circuit design using quality components; (2) using redundant circuits, modules, or subsystems where necessary; (3) using small-sized backup systems when needed; (4) reducing repair time and using more powerful diagnostics; and (5) distributing functionality to more independent modules to limit the impact of a failed module.

Currently, the trend in process control is away from centralized process control and toward an increased number of small distributed-control or PLC systems. This trend will put emphasis on the evolution of the fieldbus controller and continued growth of the PC-based controller. Also, as hardware and software improves, the functionality of the controller will increase, and the supporting hardware will be physically smaller. Hence, the traditional lines between the DCS and the PLC will become less distinct as systems will be capable of supporting either function set.

Pneumatic Controllers The pneumatic controller is an automatic controller that uses pneumatic pressure as a power source and generates a single pneumatic output pressure. The pneumatic controller is used in single-loop control applications and is often installed on the control valve or on an adjacent pipestand or wall in close proximity to the control valve and/or measurement transmitter. Pneumatic controllers are used in areas where it would be hazardous to use electronic equipment, in locations without power, in situations where maintenance personnel are more familiar with pneumatic controllers, or in applications where replacement with modern electronic controls has not been justified.

Process-variable feedback for the controller is achieved by one of two methods. The process variable can (1) be measured and transmitted to the controller by using a separate measurement transmitter with a 0.2–1.0-bar (3–15-psig) pneumatic output, or (2) be sensed directly by the controller, which contains the measurement sensor within its enclosure. Controllers with integral sensing elements are available that sense pressure, differential pressure, temperature, and level. Some controller designs have the set point adjustment knob in the controller, making set point adjustment a local and manual operation. Other types receive a set point from a remotely located pneumatic source, such as a manual air set regulator or another controller, to achieve set point adjustment. There are versions of the pneumatic controller that support the useful one-, two-, and three-mode combinations of proportional, integral, and derivative actions. Other options include auto/manual transfer stations, antireset windup circuitry, on/off control, and process-variable and set point indicators.

Pneumatic controllers are made of Bourdon tubes, bellows, diaphragms, springs, levers, cams, and other fundamental transducers to accomplish the control function. If operated on clean, dry plant air, they offer good performance and are extremely reliable. Pneumatic controllers are available with one or two stages of pneumatic amplification, with the two-stage designs having faster dynamic response characteristics.

An example of a pneumatic PI controller is shown in Fig. 8-64*a*. This controller has two stages of pneumatic amplification and a Bourdon tube input element that measures process pressure. The Bourdon tube element is a flattened tube that has been formed into a curve so that changes in pressure inside the tube cause vertical motions to occur at the ungrounded end. This motion is transferred to the left end of the beam, as shown.

The resulting motion of the beam is detected by the pneumatic nozzle amplifier, which, by proper sizing of the nozzle and fixed orifice diameters, causes the pressure internal to the nozzle to rise and fall with vertical beam motion. The internal nozzle pressure is routed to the pneumatic relay. The relay, which is constructed like the booster relay described in the "Valve Control Devices" subsection, has a direct linear input-to-output pressure characteristic. The output of the relay is the controller's output and is piped away to the final control element.

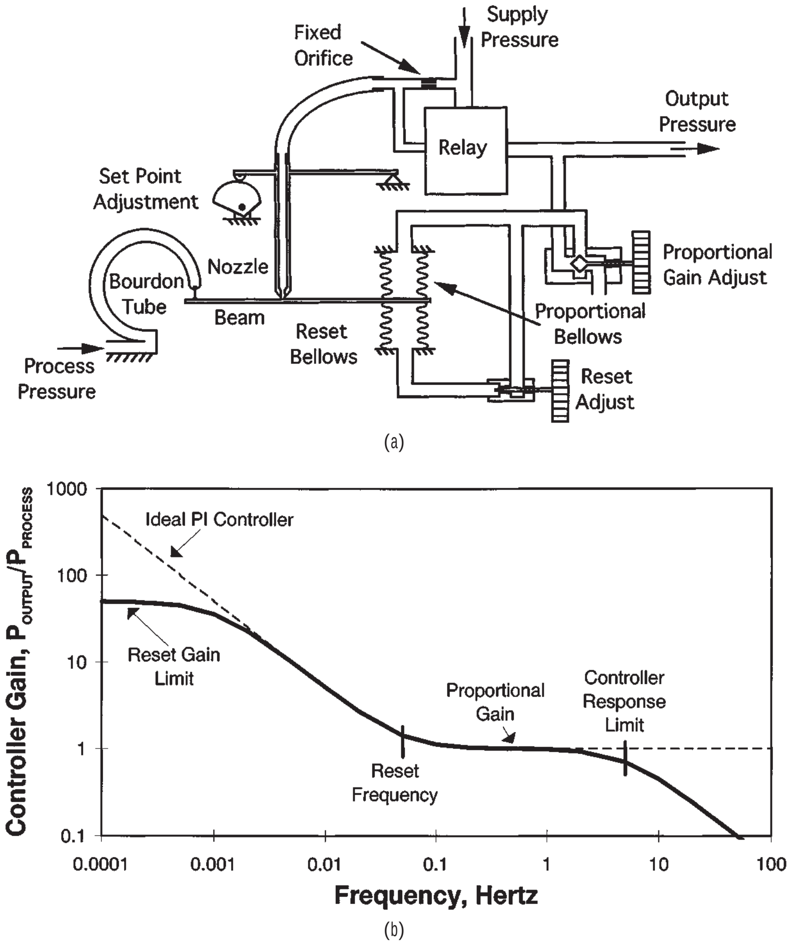


FIG. 8-64 Pneumatic controller: (*a*) example; (*b*) frequency response characteristic.

To generate the *P* and *I* control modes, a feedback circuit consisting of two bellows and two small metering valves has been added to the pneumatic amplifier system described above. The first valve is the proportional gain valve and is adjusted to provide an output pressure that is ratioed to the output pressure from the relay. The ratio used is set by manual adjustment of this valve. The output from the proportional gain valve is connected to a proportional feedback bellows, which provides negative feedback around the pneumatic amplifiers. The output pressure from the proportional gain valve is also connected to another valve, called the reset valve. This valve meters flow to and from a second bellows, known as the reset bellows. The resistance provided by the valve and the capacitance relating to the volume of the bellows forms a time constant that becomes the reset time constant for the controller.

The frequency response characteristics of a pneumatic PI controller and an ideal PI controller are shown in Fig. 8-64*b*. Notice that the gain of the pneumatic controller reaches a limit at the lowest frequencies. This limit is due to a less-than-infinite amount of forward amplifier gain. The manufacturer of the controller designs the forward gain term to be as high as possible to better approximate ideal reset action in the controller but never to reach the ideal. Reset gain for available pneumatic controllers runs between 20 and 100 times the gain implied by a proportional gain of unity. Unity proportional gain implies that a 100 percent change in process input (i.e., a full range change) will generate a 100 percent change in controller output.

The reset time, which is user-adjustable, can range from 0.05 seconds to 80 minutes or more, depending on controller design. The reset time constant, when converted to frequency $1/2(T_R)$ Hz (where T_R is the reset time in seconds), determines the frequency where the reset and proportional response characteristics of the controller merge (see Fig. 8-64*b*). Tuning the reset adjustment on the controller moves the reset frequency to the left or right along the frequency axis and thereby affects the reset action of the controller.

The response limit for the controller is a function of the design of the relay, the size of the load volume to which the controller is attached, the setting of the proportional band valve, and the forward gain designed into the pneumatic amplifiers. The frequency response limit (sometimes called the controller bandwidth) for a pneumatic controller into a small instrument load volume is in the 5- to 8-Hz range for two-stage pneumatic designs like the one shown. The dynamic response of the controller to set point changes is essentially the same as that indicated for process-variable changes. The set point adjustment mechanism affects the vertical motion of the nozzle over the beam and results in actions in the controller similar to those produced by changes in the process pressure.

The main shortcomings of the pneumatic controller is its lack of flexibility when compared to modern electronic controller designs. Increased range of adjustability, choice of alternate control algorithms, the communication link to the control system, and other features and services provided by the electronic controller make it a superior choice in most of today's applications.

CONTROL VALVES

A control valve consists of a valve, an actuator, and possibly one or more valve-control devices. The valves discussed in this section are applicable to throttling control (i.e., where flow through the valve is regulated to any desired amount between maximum and minimum limits). Other valves such as check, isolation, and relief valves are addressed in the next subsection. As defined, control valves are automatic control devices that modify the fluid flow rate as specified by the controller.

Valves Valves are categorized according to their design style. These styles can be grouped into type of stem motion—linear or rotary. The valve stem is the rod, shaft, or spindle that connects the actuator with the closure member (i.e., a movable part of the valve that is positioned in the flow path to modify the rate of flow). Motion of either type is known as travel. The major categories are described briefly below.

Globe and Angle The most common linear stem-motion control valve is the globe valve. The name comes from the globular shaped cavities around the port. In general, a port is any fluid passageway, but often the reference is to the passage that is blocked off by the closure member when the valve is closed. In globe valves, the closure member is called a plug. The plug in the valve shown in Fig. 8-65 is guided by a large-diameter port and moves within the port to provide the flow control orifice of the valve. A very popular alternate construction is a cage-guided plug as illustrated in Fig. 8-66. In many such designs, openings in the cage provide the flow control orifices. The valve seat is the zone of contact between the moving closure member and the stationary valve body, which shuts off the flow when the valve is closed. Often the seat in the body is on a replaceable part known as a seat ring. This stationary seat can also be designed as an integral part of the cage. Plugs may also be port-guided by wings or a skirt that fits snugly into the seat-ring bore.

One distinct advantage of cage guiding is the use of balanced plugs in single-port designs. The unbalanced plug depicted in Fig. 8-65 is subjected to a static pressure force equal to the port area times the valve pressure differential (plus the stem area times the downstream pressure) when the valve is closed. In the balanced design (Fig. 8-66), note that both the top and bottom of the plug are subjected to the same downstream pressure when the valve is closed. Leakage via the plug-to-cage clearance is prevented by a plug seal. Both plug types are subjected to hydrostatic force due to internal pressure acting on the stem area and to dynamic flow forces when the valve is flowing.

The plug, cage, seat ring, and associated seals are known as the trim. A key feature of globe valves is that they allow maintenance of the trim via a removable bonnet without removing the valve body from the line. Bonnets are typically bolted on but may be threaded in smaller sizes.

Angle valves are an alternate form of the globe valve. They often share the same trim options and have the top-entry bonnet style. Angle valves can eliminate the need for an elbow but are especially useful when direct impingement of the process fluid on the body wall is to be avoided. Sometimes it is not practical to package a long trim within a globe body, so an angle body is used. Some angle bodies are self draining, which is an important feature for dangerous fluids.

Butterfly The classic design of butterfly valves is shown in Fig. 8-67. Its chief advantage is high capacity in a small package and a very low initial cost. Much of the size and cost advantage is due to the wafer body design, which is clamped between two pipeline flanges. In the simplest design, there is no seal as such, merely a small clearance gap between the disc OD and the body ID. Often a true seal is provided by a resilient material in the body that is engaged via an interference fit with the disc. In a lined butterfly valve, this material covers the entire body ID and extends around the body ends to eliminate the need for pipeline joint gaskets. In a fully lined valve, the disc is also coated to minimize corrosion or erosion.

A high-performance butterfly valve has a disc that is offset from the shaft center line. This eccentricity causes the seating surface to move away from the seal once the disc is out of the closed position, reducing friction and seal wear. Also known as an eccentric disc valve, its advantage is improved shutoff while maintaining high ultimate capacity at a

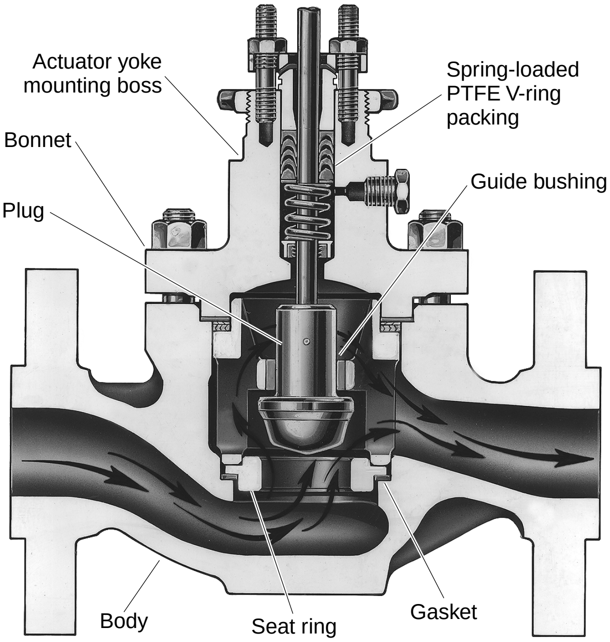

Actuator yoke mounting boss
Bonnet
Plug
Spring-loaded PTFE V-ring packing
Guide bushing
Body
Seat ring
Gasket

FIG. 8-65 Post-guided contour-plug globe valve with metal seat and raised-face flange end connections. (*Courtesy Fisher-Rosemount.*)

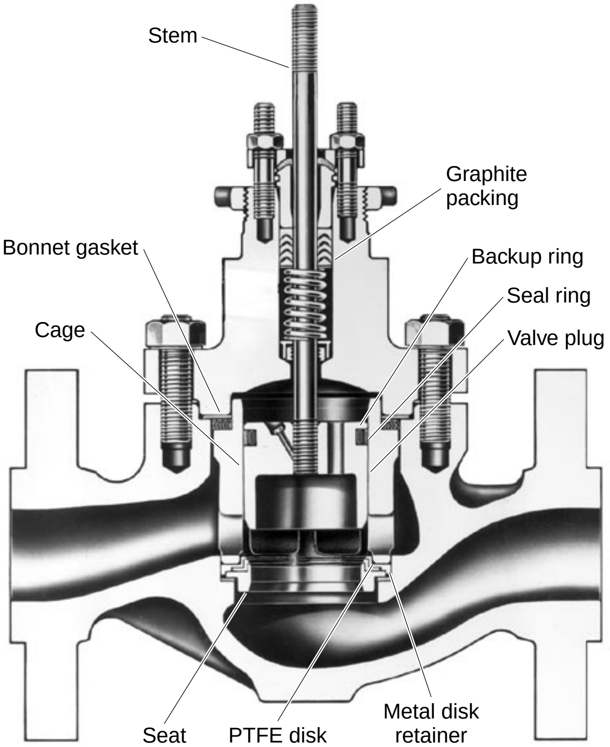


FIG. 8-66 Cage-guided balanced-plug globe valve with polymer seat and plug seal. (*Courtesy Fisher-Rosemount.*)

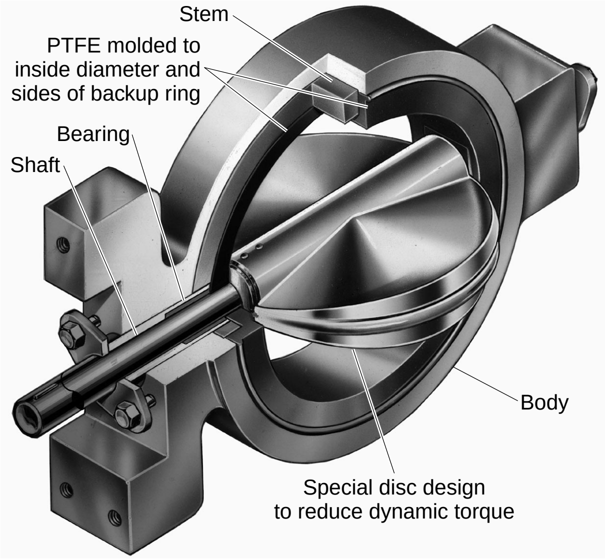

Stem
PTFE molded to inside diameter and sides of backup ring
Bearing
Shaft
Body
Special disc design to reduce dynamic torque

FIG. 8-67 Partial cutaway of wafer-style lined butterfly valve. (*Courtesy Fisher-Rosemount.*)

reasonable cost. This cost advantage relative to other design styles is particularly true in sizes above 6-inch nominal pipe size (NPS). Improved shutoff is due to advances in seal technologies, including polymer, flexing metal, combination metal with polymer inserts, and so on, many utilizing pressure assist.

Ball Ball valves get their name from the shape of the closure member. One version uses a full spherical member with a cylindrical bore through it. The ball is rotated d turn from the full-closed to the full-open position. If the bore is the same diameter as the mating-pipe fitting ID, the valve is referred to as full-bore. If the hole is undersized, the ball valve is considered to be a venturi style. A segmented ball is a portion of a hollow sphere—large enough to block the port when closed. Segmented balls often have a V-shaped contour along one edge, which provides a desirable flow characteristic (see Fig. 8-68). Both full-ball and segmented-ball valves are known for their low resistance to flow when full open. Shutoff leakage is minimized through the use of flexing or spring-loaded elastomeric or metal seals. Bodies are usually in two or three pieces or have a removable retainer to facilitate installing seals. End connections are usually flanged or threaded in small sizes, although segmented-ball valves are offered in wafer style also.

Plug There are two substantially different rotary-valve design categories referred to as plug valves. The first consists of a cylindrical or slightly conical plug with a port through it. The plug rotates to vary the flow much like a ball valve. The body is top-entry but is geometrically simpler than a globe valve and thus can be lined with fluorocarbon polymer to protect against corrosion. These plug valves have excellent shutoff but are generally not for modulating service due to high friction. A variation of the basic design (similar to the eccentric butterfly disc) only makes sealing contact in the closed position and is used for control.

The other rotary plug design is portrayed in Fig. 8-69. The seating surface is substantially offset from the shaft, producing a ball-valve-like motion with the additional cam action of the plug into the seat when closing. In reverse flow, high-velocity fluid motion is directed inward—impinging on itself and only contacting the plug and seat ring.

Multi-Port This term refers to any valve or manifold of valves with more than one inlet or outlet. For throttling control, the three-way body is used for blending (two inlets, one outlet) or as a divertor (one inlet, two outlets). A three-way valve is most commonly a special globelike body with special trim that allows flow both over and under the plug. Two rotary valves and a pipe tee can also be used. Special three-, four-, and five-way ball-valve designs are used for switching applications.

Special Application Valves

Digital Valves True digital valves consist of discrete solenoid-operated flow ports that are sized according to binary weighing. The valve can be designed with sharp-edged orifices or with streamlined nozzles that can be used for flow metering. Precise control of the throttling-control orifice is the strength of the digital valve. Digital valves are mechanically complicated and expensive, and they have considerably reduced maximum flow capacities over the globe and rotary valve styles.

Cryogenic Service Valves designed to minimize heat absorption for throttling liquids and gases below 80 K are called cryogenic service valves. These valves are designed with small valve bodies to minimize heat absorption and long bonnets between the valve and actuator to allow for extra layers of insulation around the valve. For extreme cases, vacuum jacketing can be constructed around the entire valve to minimize heat influx.

High Pressure Valves used for pressures nominally above 760

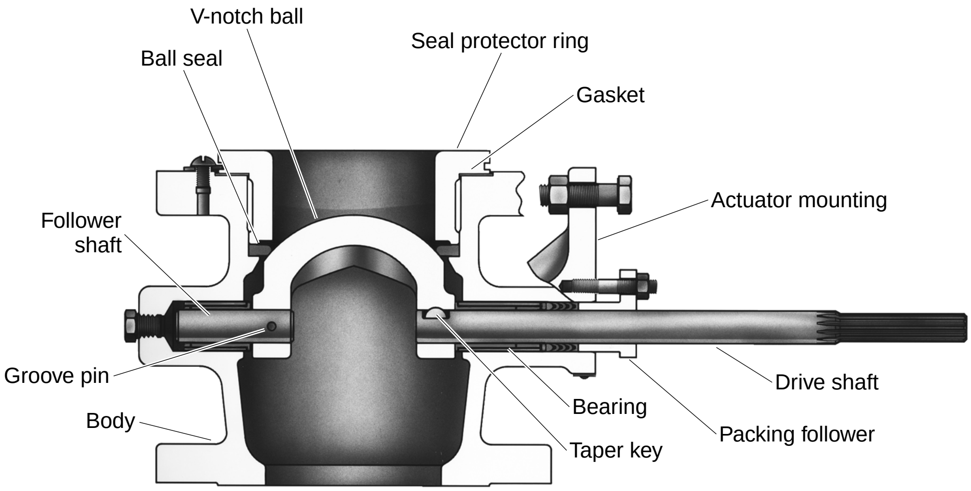


FIG. 8-68 Segmented ball valve. Partial view of actuator mounting shown 90° out of position. (*Courtesy Fisher-Rosemount.*)

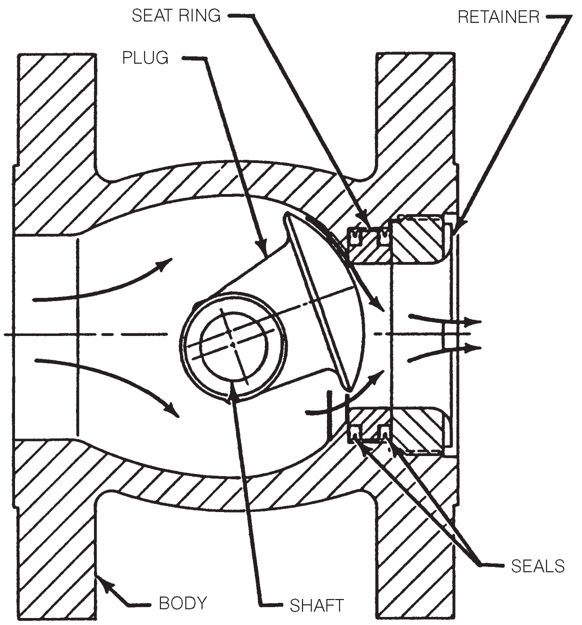


FIG. 8-69 Eccentric plug valve shown in erosion-resistant reverse flow direction. Shaded components can be made of hard metal or ceramic materials. (*Courtesy Fisher-Rosemount.*)

bar (11,000 psi, pressures above ANSI Class 4500) are often custom-designed for specific applications. Normally, these valves are of the plug type and use specially hardened plug and seat assemblies. Internal surfaces are polished, and internal corners and intersecting bores are smoothed to reduce high localized stresses in the valve body. Steam loops in the valve body are available to raise the body temperature to increase ductility and impact strength of the body material.

High-Viscous Process Used most extensively by the polymer industry, the valve for high-viscous fluids is designed with smooth finished internal passages to prevent stagnation and polymer degradation. These valves are available with integral body passages through which a heat-transfer fluid is pumped to keep the valve and process fluid heated.

Pinch The industrial equivalent of controlling flow by pinching a soda straw is the pinch valve. Valves of this type use fabric-reinforced elastomer sleeves that completely isolate the process fluid from the metal parts in the valve. The valve is actuated by applying air pressure directly to the outside of the sleeve, causing it to contract or pinch. Another method is to pinch the sleeve with a linear actuator with a specially attached foot. Pinch valves are used extensively for corrosive material service and erosive slurry service. This type of valve is used in applications with pressure drops up to 10 bar (145 psi).

Fire-Safe Valves that handle flammable fluids may have additional safety-related requirements for minimal external leakage, minimal internal (downstream) leakage, and operability during and after a fire. Being fire-safe does not mean being totally impervious to fire, but a sample valve must meet particular specifications such as American Petroleum Institute (API) 607, Factory Mutual Research Corp. (FM) 7440, or the British Standard 5146 under a simulated fire test. Due to very high flame temperature, metal seating (either primary or as a backup to a burned-out elastomer) is mandatory.

Solids Metering The control valves described earlier are primarily used for the control of fluid (liquid or gas) flow. Sometimes these valves, particularly the ball, butterfly, or sliding gate valves, are used to throttle dry or slurry solids. More often, special throttling mechanisms like venturi ejectors, conveyers, knife-type gate valves, or rotating vane valves are used. The particular solids-metering valve hardware depends on the volume, density, particle shape, and coarseness of the solids to be handled.

Actuators An actuator is a device that applies the force (torque) necessary to cause a valve's closure member to move. Actuators must overcome pressure and flow forces; friction from packing, bearings or guide surfaces, and seals; and provide the seating force. In rotary valves, maximum friction occurs in the closed position and the moment necessary to overcome it is referred to as breakout torque. The rotary valve shaft torque generated by steady-state flow and pressure forces is called dynamic torque. It may tend to open or close the valve depending on valve design and travel. Dynamic torque per unit pressure differential is largest in butterfly valves at roughly 70° open. In linear stem-motion valves, the flow forces should not exceed available actuator force, but this is usually accounted for by default when the seating force is provided.

Actuators often provide a failsafe function. In the event of an interruption in the power source, the actuator will place the valve in a predetermined safe position, usually either full open or full closed. Safety systems are often designed to trigger local failsafe action at specific valves to cause a needed action to occur, which may not be a complete process or plant shutdown.

Actuators are classified according to their power source. The nature of these sources leads naturally to design features that make their performance characteristics distinct.

Pneumatic Despite the availability of more sophisticated alternatives, the pneumatically driven actuator is still by far the most popular type. Historically the most common has been the spring and diaphragm design (Fig. 8-70). The compressed air input signal fills a chamber sealed by an elastomeric diaphragm. The pressure force on the diaphragm plate causes a spring to be compressed and the actuator stem to move. This spring provides the failsafe function and contributes to the dynamic stiffness of the actuator. If the accompanying valve is "push-down-to-close," the actuator depicted in Fig. 8-70 would be described as "air-to-close" or synonymously as fail-open. A

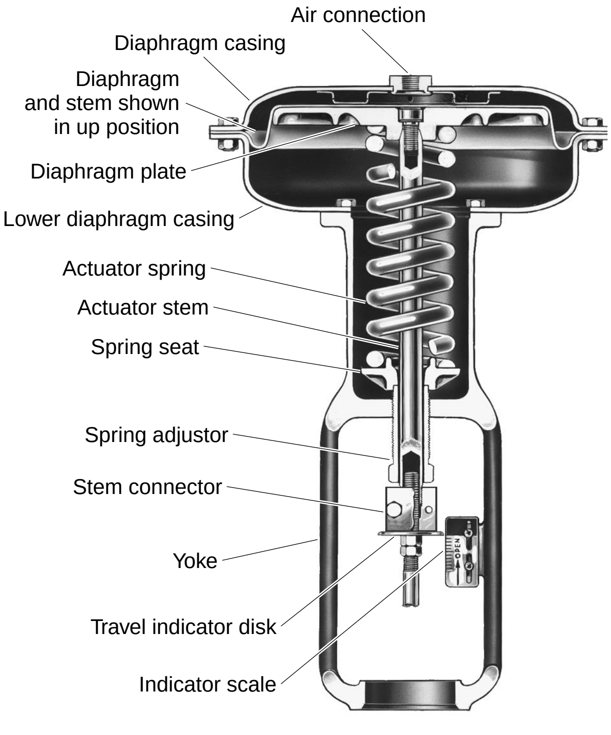


FIG. 8-70 Spring and diaphragm actuator with an "up" fail-safe mode. Spring adjuster allows slight alteration of bench set. (*Courtesy Fisher-Rosemount.*)

slightly different design yields "air-to-open" or fail-closed action. The spring is typically precompressed to provide a significant available force in the failed position (e.g., to provide seating load). The spring also provides a proportional relationship between the force generated by air pressure and stem position. The pressure range over which a spring and diaphragm actuator strokes in the absence of valve forces is known as the bench set. The chief advantages of spring and diaphragm actuators are their high reliability, low cost, adequate dynamic response, and failsafe action—all of which are inherent in their simple design.

Alternately, the pressurized chamber can be formed by a circular piston with a seal on its outer edge sliding within a cylindrical bore. Higher operating pressure (6 bar [~90 psig] is typical) and longer strokes are possible. Piston actuators can be spring-opposed but many times are in a dual-acting configuration (i.e., compressed air is applied to both sides of the piston with the net force determined from the pressure difference—see Fig. 8-71). Dynamic stiffness is usually higher with piston designs than with spring and diaphragm actuators; see "Positioner/Actuator Stiffness." Failsafe action, if necessary, is achieved without a spring through the use of additional solenoid valves, trip valves, or relays. See "Valve Control Devices."

Motion Conversion Actuator power units with translational output can be adapted to rotary valves that generally need 90° or less rotation. A lever is attached to the rotating shaft and a link with pivoting means on the end connects to the linear output of the power unit, an arrangement similar to an internal combustion engine crankshaft, connecting rod, and piston. When the actuator piston, or more commonly the diaphragm plate, is designed to tilt, one pivot can be eliminated (see Fig. 8-71). Scotch yoke and rack and pinion arrangements are also commonly used, especially with piston power units. Friction and changing mechanical advantage of these motion-conversion mechanisms means the available torque may vary greatly with travel. One notable exception is vane-style rotary actuators whose offset "piston" pivots, giving direct rotary output.

Hydraulic The design of typical hydraulic actuators is similar to double-acting piston pneumatic types. One key advantage is the high pressure (typically 35 to 70 bar [500 to 1000 psi]), which leads to high thrust in a smaller package. The incompressible nature of the hydraulic oil means these actuators have very high dynamic stiffness. The incompressibility and small chamber size connote fast stroking speed and good frequency response. The disadvantages include high initial cost, especially when considering the hydraulic supply. Maintenance is much more difficult than with pneumatics, especially on the hydraulic positioner.

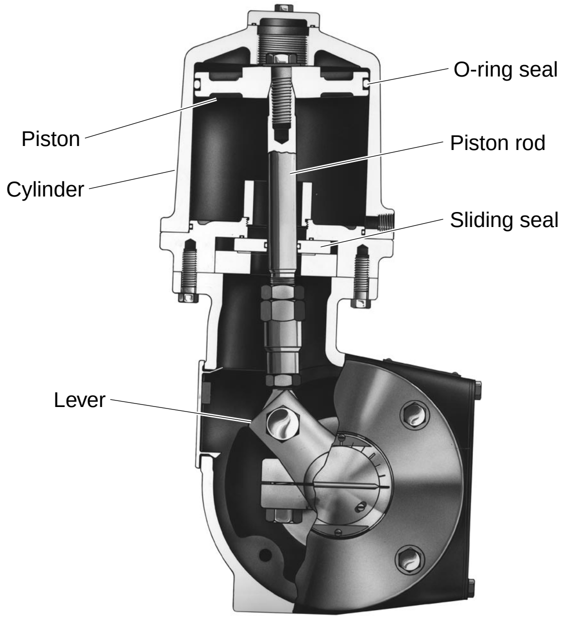


FIG. 8-71 Double-acting piston rotary actuator with lever and tilting piston for motion conversion. (*Courtesy Fisher-Rosemount.*)

Electrohydraulic actuators have similar performance characteristics and cost/maintenance ramifications. The main difference is that they contain their own electric-powered hydraulic pump. The pump may run continuously or be switched on when a change in position is required. Their main application is remote sites without an air supply when a failsafe spring return is needed.

Electric The most common electric actuators use a typical motor—three-phase AC induction, capacitor-start split-phase induction, or DC. Normally the motor output passes through a large gear reduction and, if linear motion output is required, a ball screw or thread. These devices can provide large thrust, especially given their size. Lost motion in the gearing system does create backlash, but if not operating across a thrust reversal, this type of actuator has very high stiffness. Usually the gearing system is self-locking, which means that forces on the closure member cannot move it by spinning a nonenergized motor. This behavior is called a lock-in-last-position failsafe mode. Some gear systems (e.g., low-reduction spur gears) can be backdriven. A solenoid-activated mechanical brake or locking current to motor field coils is added to provide lock-in-last-position fail mode. A battery backup system for a DC motor can guard against power failures. Otherwise, an electric actuator is not acceptable if fail-open/closed action is mandatory. Using electrical power requires environmental enclosures and explosion protection, especially in hydrocarbon-processing facilities; see the full discussion in "Valve Control Devices."

Unless sophisticated speed-control power electronics is used, position modulation is achieved via bang-bang control. Mechanical inertia causes overshoot, which is (1) minimized by braking and/or (2) hidden by adding dead band to the position control. Without these provisions, high starting currents would cause motors to overheat from constant "hunting" within the position loop. Travel is limited with power interruption switches or with force (torque) electromechanical cutouts when the closed position is against a mechanical stop (e.g., a globe valve). Electric actuators are often used for on/off service. Stepper motors can be used instead, and they, as their name implies, move in fixed incremental steps. Through gear reduction, the typical number of increments for 90° rotation range from 5000 to 10,000; hence positioning resolution at the actuator is excellent. Position overshoot is not an issue, and added dead band need only be a few steps away.

An electromagnetic solenoid can be used to directly actuate the plug on very small linear stem-motion valves. A solenoid is usually designed as a two-position device, so this valve control is on/off. Special solenoids with position feedback can provide proportional action for modulating control. Force requirements of medium-sized valves can be met with piloted plug designs, which use process pressure to assist the solenoid force. Piloted plugs are also used to minimize the size of common pneumatic actuators, especially when there is need for high seating load.

Manual A manually positioned valve is by definition not an automatic control valve, but it may be involved with process control. For rotary valves, the manual operator can be as simple as a lever, but a wheel driving a gear reduction is necessary in larger-size valves. Linear motion is normally created with a wheel turning a screw-type device. A manual override is usually available as an option for the powered actuators listed above. For spring-opposed designs, an adjustable travel stop will work as a one-way manual input. In more complex designs the handwheel can provide loop control override via an engagement means. Some gear-reduction systems of electric actuators allow the manual positioning to be independent from the automatic positioning without declutching.

Valve-Control Devices Devices mounted on the control valve that interface various forms of input signals, monitor and transmit valve position, or modify valve response are valve-control devices. In some applications, several auxiliary devices are used together on the same control valve. For example, mounted on the control valve, one may find a current-to-pressure transducer, a valve positioner, a volume booster relay, a solenoid valve, a trip valve, a limit switch, a process controller, and/or a stem-position transmitter. Figure 8-72 shows a valve positioner mounted on the yoke leg of a spring and diaphragm actuator.

As most throttling control valves are still operated by pneumatic actuators, the control-valve device descriptions that follow relate primarily to devices that are used with pneumatic actuators. The function of hydraulic and electrical counterparts are very similar. Specific details on a particular valve-control device are available from the vendor of the device.

Transducers The current-to-pressure transducer (I/P transducer) is a conversion interface that accepts a standard 4–20 mA input current from the process controller and converts it to a pneumatic output in a standard pneumatic pressure range (normally 0.2–1.0 bar [3–15 psig] or, less frequently, 0.4–2.0 bar [6–30 psig]). The output pressure generated by the transducer is connected directly to the pressure connection on a spring-opposed diaphragm actuator or to the input of a pneumatic valve positioner.

Figure 8-73*a* is the schematic of a basic I/P transducer. The transducer shown is characterized by (1) an input conversion that generates an angular displacement of the beam proportional to the input current, (2) a pneumatic amplifier stage that converts the resulting angular displacement to pneumatic pressure, and (3) a pressure area that serves as a means to return the beam back to very near its original position when the new output pressure is achieved. The result is a device that generates a pressure output that tracks the input current signal. The transducer shown in Fig. 8-73*a* is used to provide pressure to small load volumes (normally 4.0 in^3 or less), such as a positioner or booster input. With only one stage of pneumatic amplification, the flow capacity of this transducer is limited and not sufficient to provide responsive load pressure directly to a pneumatic actuator.

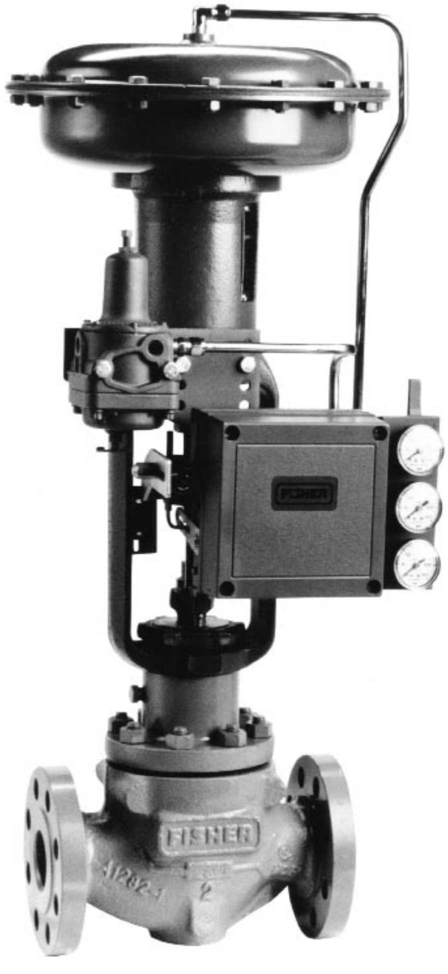


FIG. 8-72 Valve and actuator with valve positioner attached. (*Courtesy Fisher-Rosemount.*)

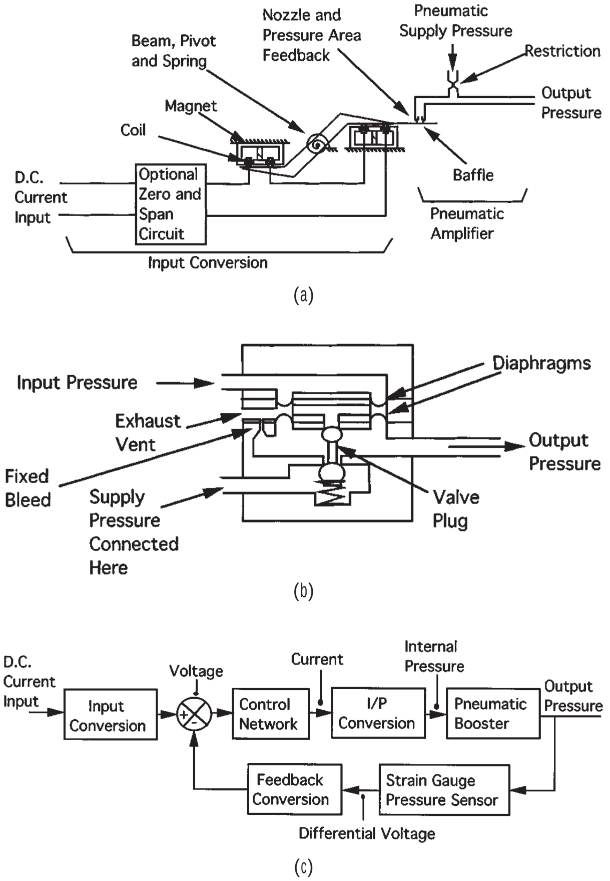


FIG. 8-73 Current to pressure transducer components parts: (*a*) direct current to pressure conversion; (*b*) pneumatic booster amplifier (relay); (*c*) block diagram of a modern I/P transducer.

The flow capacity of the transducer can be increased by adding a booster relay like the one shown in Fig. 8-73*b*. The flow capacity of the booster relay is nominally fifty to one hundred times that of the nozzle amplifier shown in Fig. 8-73*a* and makes the combined transducer/booster suitably responsive to operate pneumatic actuators. This type of transducer is stable into all sizes of load volumes and produces measured accuracy (see Instrument Society of America [ISA]-S51.1-1979, "Process Instrumentation Terminology" for the definition of measured accuracy) of 0.5 percent to 1.0 percent of span.

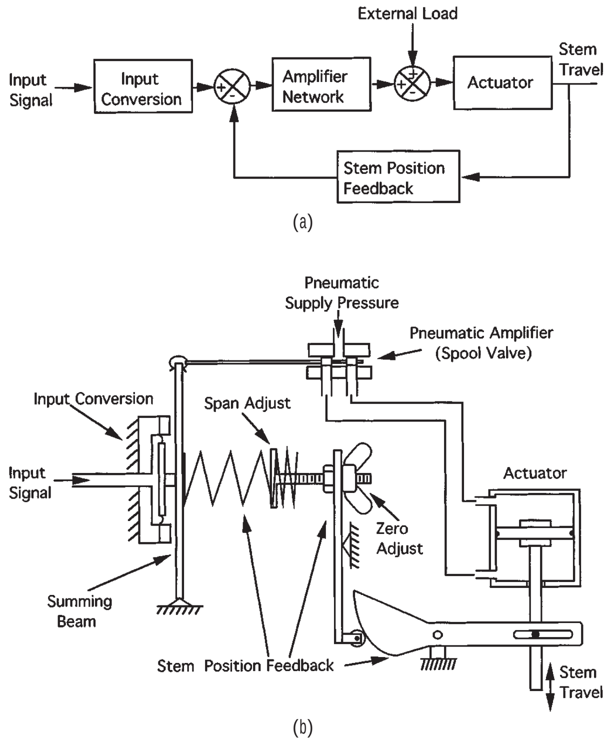
Better measured accuracy results from the transducer design shown in Fig. 8-73*c*. In this design, pressure feedback is taken at the output of the booster-relay stage and fed back to the main summer. This allows the transducer to correct for errors generated in the pneumatic booster as well as errors in the I/P-conversion stage. Also, particularly with the new analog electric and digital versions of this design, PID control is used in the transducer-control network to give extremely good static accuracy, fast dynamic response, and reasonable stability into a wide range of load volumes (small instrument bellows to large actuators). Also, environmental factors such as temperature change, vibration, and supply pressure fluctuation affect this type of transducer the least. Even a perfectly accurate I/P transducer cannot compensate for stem-position errors generated by friction, backlash, and varying force loads coming from the actuator and valve. To do this compensation, a different control-valve device, known as a valve positioner, is required.

Valve Positioners The valve positioner, when combined with an appropriate actuator, forms a complete closed-loop valve-position control system. This system makes the valve stem conform to the input signal coming from the process controller in spite of force loads that the actuator may encounter while moving the control valve. Usually, the valve positioner is contained in its own enclosure and is mounted on the control valve.

The key parts of the positioner/actuator system, shown in Fig. 8-74*a,* are (1) an input-conversion network, (2) a stem-position feedback network, (3) a summing junction, (4) an amplifier network, and (5) an actuator.

The input-conversion network shown is the interface between the input signal and the summer. This block converts the input current or pressure (from an I/P transducer or a pneumatic process controller) to a voltage, an electric current, a force, torque, displacement or other particular variable that can be directly used by the summer. The input conversion usually contains a means to adjust the slope and offset of the block to provide for a means of spanning and zeroing the positioner during calibration. In addition, means for changing the sense (known as "action") of the input/output characteristic are oftentimes addressed in this block. Also, exponential, logarithmic or other predetermined characterization can be put in this block to provide a characteristic that is useful in offsetting or reinforcing a nonlinear valve or process characteristic.

The stem-position feedback network converts stem travel to a useful form for the summer. This block includes the feedback linkage, which varies with actuator type. Depending on positioner design, the stem-position feedback network can provide span and zero and characterization functions similar to that described for the input-conversion block.



(a)

(b)

FIG. 8-74 Positioner/actuators: (*a*) generic block diagram; (*b*) an example of a pneumatic positioner/actuator.

The amplifier network provides signal conversion and suitable static and dynamic compensation for good positioner performance. Control from this block usually reduces down to a form of proportional or proportional plus derivative control. The output from this block in the case of a pneumatic positioner is a single connection to the spring and diaphragm actuator or two connections for push-pull operation of a springless piston actuator. The action of the amplifier network and the action of the stem-position feedback can be reversed together to provide for reversed positioner action.

By design, the gain of the amplifier network shown in Fig. 8-74*a* is made very large. Large gain in the amplifier network means that only a small proportional deviation will be required to position the actuator through its active range of travels. This means that the signals into the summer track very closely and that the gain of the input-conversion block and the stem-position feedback block determine the closed-loop relationship between the input signal and the stem travel.

Large amplifier gain also means that only a small amount of additional stem-travel deviation will result when large external force loads are applied to the actuator stem. For example, if the positioner's amplifier network has a gain of 50 and assuming that high packing-box friction loads require 25 percent of the actuator's range of thrust to move the actuator, then only 25 percent/50 or 0.5 percent deviation between input signal and output travel will result due to valve friction.

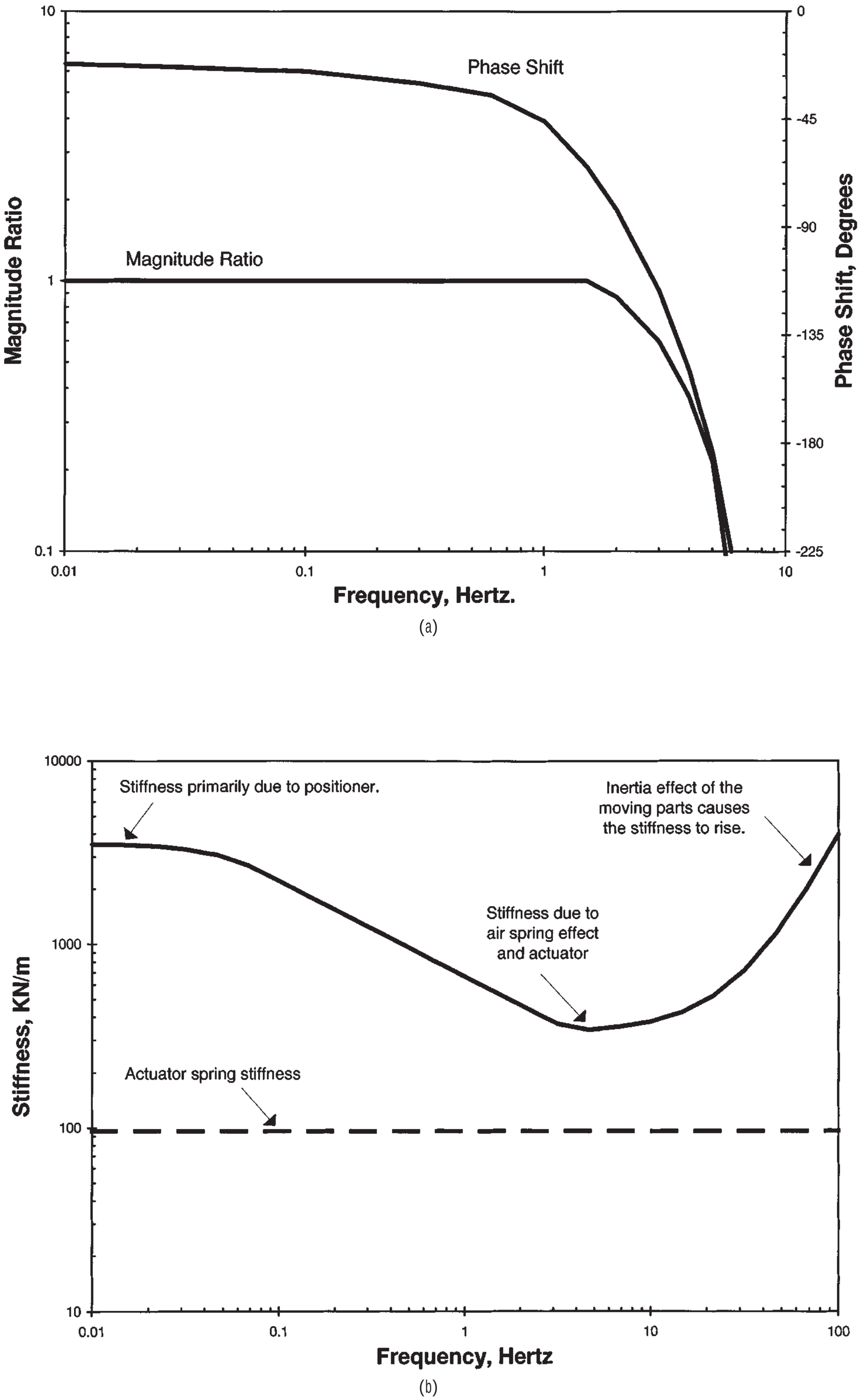
Figure 8-74*b* is an example of a pneumatic positioner/actuator. The input signal is a pneumatic pressure that (1) moves the summing beam, which (2) operates the spool valve amplifier, which (3) provides flow to and from the piston actuator, which (4) causes the actuator to move and continue moving until (5) the feedback force returns the beam to its original position and stops valve travel at a new position. Typical positioner operation is thereby achieved.

Static performance measurements related to positioner/actuator operation are: conformity, measured accuracy, hysteresis, dead band, repeatability, and locked stem-pressure gain. Definitions and standardized test procedures for determining these measurements can be found in ISA-S75.13-1989, "Method of Evaluating the Performance of Positioners with Analog Input Signals and Pneumatic Output".

Dynamics of Pneumatic Positioners Dynamically, the pneumatic positioner is characterized by the combined effects of gain and capacitance and nonlinear effects such as valve friction and flow-capacity saturation. Generally, there is a threshold level of input signal below which the positioner output will not respond at all. This band is on the order of 0.1 percent of input span for pneumatic positioners but can be larger if significant valve friction is present. Above this threshold level, but below the level that causes velocity saturation, the positioner is approximately linear and likened to a second-order low-pass filter (see Fig. 8-75*a*). Natural frequencies range from 0.3 to 3.0 Hz and damping ratios of 0.6 to 2.0 are common and dependent on positioner design and the physical size of the actuator volume. At higher drive levels, the flow capacity of the positioner is reached and attenuation of the resulting travel begins.

Positioner/Actuator Stiffness Minimizing the effect of dynamic loads on valve-stem travel is an important characteristic of the positioner/actuator. Stem position must be maintained in spite of changing reaction forces caused by valve throttling. These forces can be random in nature (buffeting force) or result from a negative sloped force/stem travel characteristic (negative gradient); either could result in valve-stem instability and loss of control. To reduce and eliminate the effect of these forces, the effective stiffness of the positioner/actuator must be made sufficiently high to maintain adequate stability of the valve stem.

The stiffness characteristic of the positioner/actuator varies with frequency. Figure 8-75*b* indicates the stiffness of the positioner/actuator is high at low frequencies and is directly related to the locked-stem pressure gain provided by the positioner. As frequency increases, a dip in the stiffness curve results from dynamic gain attenuation in the pneumatic amplifiers in the positioner. The value at the bottom of the dip is the sum of the mechanical stiffness of the spring in the actu-



(a)

(b)

FIG. 8-75 Frequency response curves for a pneumatic positioner/actuator: (*a*) input signal to stem travel for a 69-inch^2 spring and diaphragm actuator with a 1.5-inch total travel and 3–15 psig input pressure; (*b*) dynamic stiffness for the same positioner/actuator.

ator and the air spring effect produced by air enclosed in the actuator casing.

The air spring effect results from adiabatic expansion and compression of the air in the actuator casing. Numerically, the small perturbation value for air spring stiffness in Newtons/meter is given by Eq. (8-107).

$$\text{Air spring rate} = \frac{\gamma P_a A^2 a}{V} \tag{8-107}$$

where γ is ratio of specific heats (1.4 for air), P_a is the actuator pressure in Pascal absolute, A_a is the actuator pressure area in m^2, and V is the internal actuator volume in m^3.

Notice in the figure that the minimum stiffness value (mechanical spring stiffness + air spring stiffness) is several times larger than the stiffness produced by the spring in the actuator (shown as a dotted line) by itself. This indicates that the air spring stiffness is quite significant and worth considering in actuator design and actuator sizing. To the right of the dip, the inertia effects of the mass of the moving parts of the valve and actuator cause the overall system stiffness to rise with increasing frequency.

Positioner Application Positioners are widely used on pneumatic valve actuators. More often than not, they provide improved process-loop control because they reduce valve-related nonlinearity. Dynamically, positioners maintain their ability to improve control-valve performance for sinusoidal input frequencies up to about one half of the positioner bandwidth. At input frequencies greater than this, the attenuation in the positioner amplifier network gets large, and valve nonlinearity begins to affect final control-element performance more significantly. Because of this, the most successful use of the positioner occurs when the positioner-response bandwidth is greater than twice that of the most dominant time lag in the process loop.

Some typical examples of where the dynamics of the positioner are sufficiently fast to improve process control are the following:

1. *In a distributed control system (DCS) process loop with an electronic transmitter.* The DCS controller and the electronic transmitter have time constants that are dominant over the positioner response. Positioner operation is therefore beneficial in reducing valve-related nonlinearity.

2. *In a process loop with a pneumatic controller and a large process time constant.* Here the process time constant is dominant, and the positioner will improve the linearity of the final control element. Some common processes with large time constants that benefit from positioner application are liquid level, temperature, large volume gas pressure, and mixing.

3. Additional situations where valve positioners are used are as follows:

- On springless actuators where the actuator is not usable for throttling control without position feedback.
- When split ranging is required to control two or more valves sequentially. In the case of two valves, the smaller control valve is calibrated to open in the lower half of the input signal range and a larger valve is calibrated to open in the upper half of the input signal range. Calibrating the input command signal range in this way is known as split-range operation and increases the practical range of throttling process flows over that of a single valve.
- In open-loop control applications where best static accuracy is needed.

On occasion, positioner use can degrade process control. Such is the case when the process controller, the process, and the process transmitter have time constants that are similar or smaller than that of the positioner/actuator. This situation is characterized by low process-controller P gain (P gain < 0.5), and hunting or limit cycling of the process variable is observed. Improvements here can be made by doing one of the following:

- Install a dominant first-order low-pass filter in the loop ahead of the positioner and retune the process loop. This should allow increased proportional gain in the process loop and reduce hunting. Possible means for adding the filter include adding it to the firmware of the DCS controller, by adding an external RC network on the output of the process controller or by enabling the filter function in the input of the positioner if it is available. Also, some transducers, when connected directly to the actuator, form a dominant first-order lag that can be used to stabilize the process loop.
- Select a positioner with a faster response characteristic.

Booster Relays The booster relay is a single-stage power amplifier having a fixed gain relationship between the input and output pressures. The device is packaged as a complete standalone unit with pipe-thread connections for input, output, and supply pressure. The booster amplifier shown in Fig. 8-73*b* shows the basic construction of the booster relay. Enhanced versions are available that provide specific features such as: (1) variable gain to split the output range of a pneumatic controller to operate more than one valve or to provide additional actuator force; (2) low hysteresis for relaying measurement and control signals; (3) high flow capacity for increased actuator-stroking speed; and (4) arithmetic, logic, or other compensation functions for control system design.

A particular type of booster relay, called a dead-band booster, is shown in Fig. 8-76. This booster is designed to be used exclusively between the output of a valve positioner and the input to a pneumatic actuator. It is designed to provide extra flow capacity to stroke the actuator faster than with the positioner alone. The dead-band booster is designed intentionally with a large dead band (approximately 5 percent of the input span), elastomer seats for tight shutoff, and an adjustable bypass valve connected between the input and the output of the booster. The bypass valve is tuned to provide the best compromise between increased actuator stroking speed and positioner/actuator stability.

With the exception of the dead-band booster, the application of booster relays has diminished somewhat by the increased use of current-to-pressure transducers, electropneumatic positioners, and electronic control systems. Transducers and valve positioners serve much the same functionality as the booster relay in addition to interfacing with the electronic process controller.

Solenoid Valves The electric solenoid valve has two output states. When sufficient electric current is supplied to the coil, an internal armature moves against a spring to an extreme position. This motion causes an attached pneumatic or hydraulic valve to operate. When current is removed, the spring returns the armature and the attached solenoid valve to the deenergized position. An intermediate pilot stage is sometimes used when additional force is required to operate the main solenoid valve. Generally, solenoid valves are used to pressurize or vent the actuator casing for on/off control-valve application and safety shutdown applications.

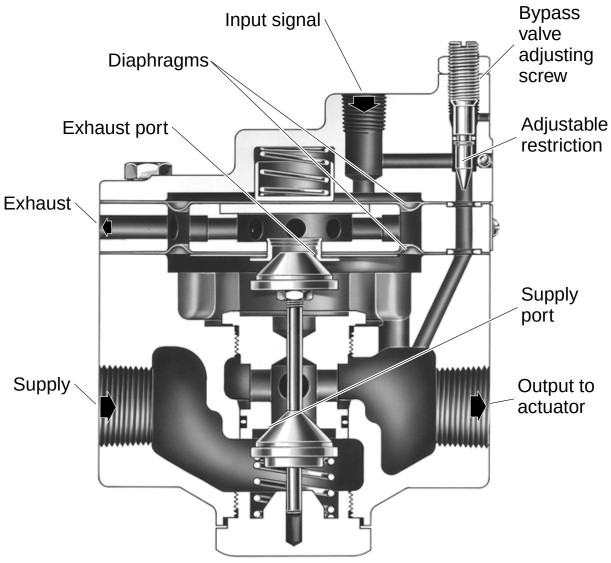


FIG. 8-76 Dead-band booster relay. (*Courtesy Fisher-Rosemount.*)

Trip Valves The trip valve is part of a system that is used where a specific valve action (i.e., fail up, fail down, or lock in last position) is required when pneumatic supply pressure to the control valve falls below a preset level. Trip systems are used primarily on springless piston actuators requiring fail-open or fail-closed action. An air storage or "volume" tank and a check valve are used with the trip valve to provide power to stroke the valve when supply pressure is lost. Trip valves are designed with hysteresis around the trip point to avoid instability when the trip pressure and the reset pressure settings are too close to the same value.

Limit Switches and Stem-Position Transmitters Travel-limit switches, position switches, and valve-position transmitters are devices that, when mounted on the valve, actuator, damper, louver, or other throttling element, detect the component's relative position. The switches are used to operate alarms, signal lights, relays, solenoid valves, or discrete inputs into the control system. The valve-position transmitter generates a 4–20-mA output that is proportional to the position of the valve.

Fire and Explosion Protection Electrical equipment can be a source of ignition in environments with combustible concentrations of gas, liquid, dust, fibers, or flyings. Most of the time it is possible to locate the electronic equipment away from these hazardous areas. However, where electric or electronic valve-mounted instruments must be used in areas where there is a hazard of fire or explosion, the equipment must be designed to meet requirements for safety. Articles 500 through 504 of the National Electrical Code address definitions of hazardous locations and the requirements for electrical devices in locations where fire or explosion hazard exists. NFPA (National Fire Protection Agency) 497M addresses the properties and group classification of gas, vapor, and dust for electrical devices in hazardous locations. With valve-mounted accessories, the approved protection concepts most often used for safety protection are explosion-proof, intrinsically safe, nonincendive, and dust-ignition-proof.

The explosion-proof enclosure is designed such that an explosion in the interior of the enclosure containing the electronic circuits will be contained. The enclosure will not allow sufficient flame to escape to the exterior to cause an ignition. Also, a surface temperature rating is given to the device. This rating must indicate a lower surface temperature than the ignition temperature of the gas in the hazardous area.

Explosion-proof enclosures are characterized by strong metal enclosures with special close-fitting access covers and breathers that contain an ignition to the inside of the enclosure. Field wiring in the hazardous environment is enclosed in a metal conduit of the mineral-insulated-cable type. All conduit and cable connections or cable terminations are threaded and explosion-proof. Conduit seals are put into the conduit or cable system at locations defined by the National Electric Code (Article 501) to prevent gas and vapor leakage and to prevent flames from passing from one part of the conduit system to the other.

The intrinsically safe (I.S.) control-valve device contains circuits that are incapable of releasing sufficient electrical or thermal energy to cause ignition of a specified hazardous mixture under normal or fault operating conditions of the circuit. I.S. circuits are designed with voltage- and current-limiting networks added where necessary to achieve approved levels of safety. I.S. field wiring need not be enclosed in metal conduit but must be kept separate from wiring for nonintrinsically safe circuits. Intrinsically safe field wiring must be energy limited, usually by a Zener diode barrier circuit located in the control room. The manufacture of the intrinsically safe control-valve devices must list the identification number of the control drawing on the nameplate attached to the approved device. The control drawing contains information showing approved combinations of accessories and other connected apparatus such as Zener diode energy barriers.

ANSI/ISA S12.12, "Nonincendive Electrical Equipment for Use in Class I and Class II, Division 2 and Class III, Division 1 and 2 Hazardous (Classified) Locations," addresses requirements for nonincendive electrical equipment and wiring. Nonincendive apparatus and/or field wiring refers to approved equipment or wiring that is incapable of imparting sufficient energy to ignite the specified hazardous atmosphere under normal circuit operating conditions. Nonincendive protection is considered for applications where hazardous concentrations of flammable gas and vapors or combustible dusts are only present under abnormal operating conditions or in those applications where easily ignited fibers or flyings are present in sufficient quantities to cause ignitable mixtures. For applications where hazardous concentrations of flammable gas and vapors or combustible dust are present continuously, intermittently, or periodically, more stringent protection (see Division 1, National Electrical Code, Article 500) offered by explosion-proof or intrinsically safe concepts is required.

The dust-ignition-proof protection concept excludes dust from entering the device enclosure and will not permit arcs, sparks, or heat generated by the device to cause ignition of external suspensions or accumulations of the dust. Enclosure requirements can be found in ANSI/UL 1203-1994, "Explosion-Proof and Dust-Ignition-Proof Electrical Equipment for Use in Hazardous Locations."

Certified testing and approval for control-valve devices used in hazardous locations is normally procured by the manufacturer of the device. The manufacturer often goes to a third party laboratory for testing and certification. Applicable approval standards are available from CSA, CENELEC, FM, SAA, and UL.

Environmental Enclosures Enclosures for valve accessories are sometimes required to provide protection from specific environmental conditions. The National Electrical Manufacturers Association (NEMA) provides descriptions and test methods for equipment used in specific environmental conditions in NEMA 250. Protection against rain, windblown dust, hose-directed water, and external ice formation are examples of environmental conditions that are covered by NEMA standards.

Also, the electronic control-valve device's level of immunity to, and emission of, electromagnetic interference (EMI) can be an issue in the chemical-valve environment. EMI requirements for the control-valve devices are presently mandatory in the European Community but voluntary in the United States, Japan, and the rest of the world. International Electrotechnical Commission (IEC) 801, Parts 1 through 4, "Electromagnetic Compatibility for Industrial Process Measurement and Control Equipment," defines tests and requirements for control-device immunity. Immunity and emission standards are addressed in CENELEC (European Committee for Electrotechnical Standardization) EN 50 081-1:1992, EN 50 081-2:1993, EN 50 082-1:1992, and prEN 50 082-2:1994.

Digital Field Communications An increasing number of valve-mounted devices are available that support digital communications in addition to, or in place of, the traditional 4–20 mA current signal. These control-valve devices have increased functionality, resulting in reduced setup time, improved control, combined functionality of traditionally separate devices, and control-valve diagnostic capability. Digital communications also allow the control system to become completely distributed where, for example, the process PID controller could reside in the valve positioner or in the process transmitter.

The high-performance, all-digital, multidrop communication protocol for use in the process-control industry is known as fieldbus. Presently there are several regional and industry-based fieldbus standards including the French standard, FIP (NFC 4660x approved by UTE), the German standard, Profibus (DIN 19245 approved by DKE), and proprietary standards by DCS vendors. As of 1997, none of these fieldbus standards have been adopted by international standards organizations.

The International Electrotechnical Commission (IEC) Standards Committee 65C Working Group 6 (IEC SC65C WG6) and the ISA Standards and Practices Committee 50 (ISA SP50) are presently working on a fieldbus standard, but at the time of this writing, the standard is unfinished. One interim solution supported by some valve-device products is the hybrid communication method, where both analog and digital communication capabilities are present in the same device. This scheme has the advantage of allowing the communicating valve-control device to be retrofit into a traditional 4–20 mA current loop and still support digital communications between the final control element and the control room. Here the current signal is used to communicate the primary signal value, and the digital communication channel carries secondary variable information, configuration information, calibration information, and alert and diagnostic information.

An example of a hybrid protocol that is open (not proprietary) and

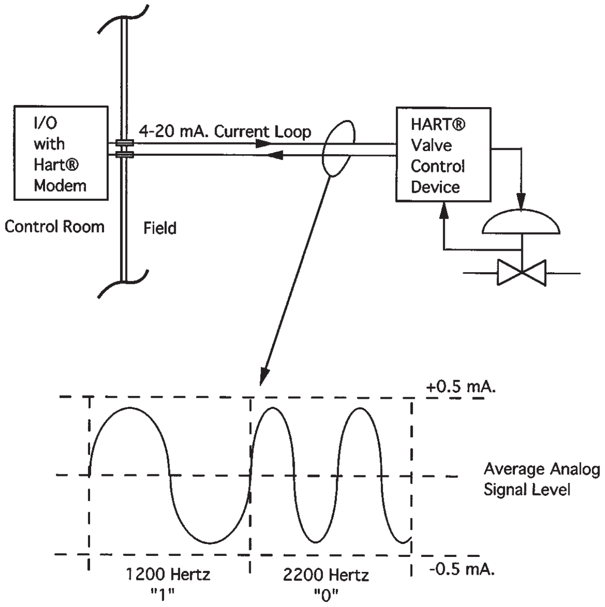


FIG. 8-77 Hybrid point-to-point communications between the control room and the control valve device.

in use by several manufacturers of control-valve devices is known as HART®* (Highway Addressable Remote Transducer) protocol (see Fig. 8-77). With this protocol, the digital communications occur over the same two wires that provide the 4–20 mA process control signal without disrupting the process signal. The protocol uses the frequency-shift keying (FSK) technique where two individual frequencies, one representing the mark and the other representing the space, are superimposed on the 4–20 mA current signal. As the average value of the signals used is zero, there is no DC offset value added to the 4–20 mA signal.

Valve Application Technology Functional requirements and the properties of the controlled fluid determine which valve and actuator types are best for a specific application. If demands are modest and no unique valve features are required, the valve-design style selection may be determined solely by cost. If so, general-purpose globe or angle valves provide exceptional value, especially in sizes less than 3-inch NPS and hence are very popular. Beyond type selection, there are many other valve specifications that must be determined properly in order to ultimately yield-improved process control.

Materials and Pressure Ratings Valves must be constructed from materials that are sufficiently immune to corrosive or erosive action by the process fluid. Common body materials are cast iron, steel, stainless steel, high-nickel alloys, and copper alloys such as bronze. Trim materials usually need a greater immunity due to the higher fluid velocity in the throttling region. High hardness is desirable in erosive and cavitating applications. Heat-treated and precipitation-hardened stainless steels are common. High hardness is also good for guiding, bearing, and seating surfaces; cobalt-chromium alloys are utilized in cast or wrought form and frequently as welded overlays called hard facing. In less stringent situations, chrome plating, heat-treated nickel coatings, and ion nitriding are used. Tungsten carbide and ceramic trim are warranted in extremely erosive services. See Sec. 28, "Materials of Construction," for specific material properties.

Since the valve body is a pressurized vessel, it is usually designed to comply with a standardized system of pressure ratings. Two common systems are described in the standards ANSI B16.34 and DIN 2401. Internal pressure limits under these standards are divided into broad classes, with specific limits being a function of material and temperature. Manufacturers also assign their own pressure ratings based on internal design rules. A common insignia is "250 WOG," which means a pressure rating of 250 psig (~17 bar) in water, oil, or gas at ambient temperature. "Storage and Process Vessels" in Sec. 10 provides introductory information on compliance of pressure-vessel design to industry codes (e.g., ASME Boiler and Pressure Vessel Code—Section VIII, ASME B31.3 Chemical Plant and Petroleum Refinery Piping).

Valve bodies are also standardized to mate with common piping connections: flanged, butt-weld end, socket-weld end, and screwed end. Dimensional information for some of these joints and class pressure-temperature ratings are included in Sec. 10, "Process Plant Piping." Control valves have their own standardized face-to-face dimensions that are governed by ISA Standards S75.03, 04, 12, 14, 15, 16, 20, and 22. Butterfly valves are also governed by API 609 and Manufacturers Standardization Society (MSS) SP-67 and 68.

Sizing Throttling control valves must be selected to pass the required flow rate given expected pressure conditions. Sizing is not merely matching the end connection size with surrounding piping; it is a key step in ensuring that the process can be properly controlled. Sizing methods range from simple models based on elementary fluid mechanics to very complex models when unusual thermodynamics or nonideal behaviors occur. Basic sizing practices have been standardized upon (e.g., ISA S75.01) and are implemented as PC-based programs by manufacturers. The following is a discussion of very basic sizing equations and the associated physics.

Regardless of the particular process variable being controlled (e.g., temperature, level, pH), the output of a control valve is flow rate. The throttling valve performs its function of manipulating flow rate by virtue of being an adjustable resistance to flow. Flow rate and pressure conditions are normally known when a process is designed and the valve resistance range must be matched accordingly. In the tradition of orifice and nozzle discharge coefficients, this resistance is embodied in the valve flow coefficient C_v. The mass flow rate (w) in kg/h is given for a liquid by

$$w = 27.3 C_v \sqrt{\rho(p_1 - p_2)} \tag{8-108}$$

where p_1 and p_2 are upstream and downstream static pressure in bar, respectively. The density of the fluid ρ is expressed in kg/m^3. This equation is valid for nonvaporizing, turbulent-flow conditions for a valve with no attached fittings. The relationship can be derived from the principles of conservation of mass and energy. A more complete presentation of sizing relationships is given in ISA S75.01, including provisions for pipe reducers, vaporizing liquids, and Reynolds number effects.

While the above equation gives the relationship between pressure and flow from a macroscopic point of view, it does not explain what is going on inside the valve. Valves create a resistance to flow by restricting the cross sectional area of the flow passage and also by forcing the fluid to change direction as it passes through the body and trim. The conservation of mass principle dictates that, for steady flow, density × average velocity × cross sectional area equals a constant. The average velocity of the fluid stream at the minimum restriction in the valve is therefore much higher than at the inlet. Note that due to the abrupt nature of the flow contraction that forms the minimum passage, the main fluid stream may separate from the passage walls and form a jet that has an even smaller cross section, the so-called vena contracta. The ratio of minimum stream area to the corresponding passage area is called the contraction coefficient. As the fluid expands from the minimum cross sectional area to the full passage area in the downstream piping, large amounts of turbulence are generated. Direction changes can also induce significant amounts of turbulence.

Some of the potential energy that was stored in the fluid by pressurizing it (e.g., the work done by a pump) is first converted into the kinetic energy of the fast-moving fluid at the vena contracta. Some of that kinetic energy turns into the kinetic energy of turbulence. As the turbulent eddies break down into smaller and smaller structures, viscous effects ultimately convert all of the turbulent energy into heat. Therefore, a valve converts fluid energy from one form to another.

For many valve constructions, it is reasonable to approximate the

[*] HART is a trademark owned by Rosemount, Inc.

fluid transition from the valve inlet to the minimum cross section of the flow stream as an isentropic or lossless process. This minimum pressure, p_{vc}, can be estimated from the Bernoulli relationship. See Sec. 6 ("Fluid and Particle Mechanics") for more background information. Downstream of the vena contracta, the flow is definitely not lossless due to all the turbulence that is generated. As the flow passage area increases and the fluid slows down, some of the kinetic energy of the fluid is converted back to potential energy as the pressure recovers. The remaining energy that is permanently lost via turbulence accounts for the permanent pressure or head loss of the valve. The relative amount of pressure that is recouped determines whether the valve is considered to be high or low recovery. See Fig. 8-78 for an illustration of how the mean pressure changes as fluid moves through a valve. The flow-passage geometry at and downstream of the vena contracta primarily determines the amount of recovery. The amount of recovery is quantified by the ***liquid*** pressure recovery factor F_L where

$$F_L = \sqrt{\frac{p_1 - p_2}{p_1 - p_{\text{vc}}}} \tag{8-109}$$

A key limitation of sizing Eq. (8-109) is the limitation to incompressible fluids. For gases and vapors, density is dependent on pressure. For convenience, compressible fluids are often assumed to follow the ideal-gas-law model. Deviations from ideal behavior are corrected for, to first order, with nonunity values of compressibility factor Z. (See Sec. 2, "Physical and Chemical Data," for definitions and data for common fluids.) For compressible fluids

$$w = 94.8 C_v P_1 Y \sqrt{\frac{x M_w}{T_1 Z}} \tag{8-110}$$

where P_1 is in bar absolute, T_1 is inlet temperature in K, M_w is the molecular weight, and x is the dimensionless pressure-drop ratio $(p_1 - p_2)/p_1$. The expansion factor Y accounts for changes in the fluid density as the fluid passes through the valve and for variation in the contraction coefficient with pressure drop. For convenience the experimental data is approximated by a simple relationship:

$$Y = 1 - \frac{1.4x}{3x_T\gamma} \qquad \text{for} \qquad x \le \frac{x_T\gamma}{1.4} \tag{8-111}$$

where γ is the ratio of specific heats and x_T is the pressure drop ratio factor. Even though a fluid may be compressible, if the value of x is small, the flow will behave as though it is incompressible. In the limit as x goes to zero, Eq. (8-110) reduces to the incompressible form Eq. (8-108) with ρ expressed via the ideal-gas equation of state.

Compressible fluids exhibit a phenomenon known as choking. Given a nozzle geometry with fixed inlet conditions, the mass flow rate will increase as P_2 is decreased up to a maximum amount at the critical pressure drop. The velocity at the vena contracta has reached sonic and a standing shock has formed. This shock causes a step change in pressure as flow passes through it, and further reduction in P_2 does not increase mass flow. x_T is a parameter of the flow model that relates to the critical pressure-drop ratio but also accounts for valve geometry effects. The value of x_T varies with flow-path geometry; a rough estimate for conventional valves is one-half. In the choked case,

$$x > \frac{x_T\gamma}{1.4} \qquad \text{and} \qquad Y = 0.67 \tag{8-112}$$

Noise Control Sound is a fluctuation of air pressure that can be detected by the human ear. Sound travels through any fluid (e.g., the air) as a compression/expansion wave. This wave travels radially outward in all directions from the sound source. The pressure wave induces an oscillating motion in the transmitting medium that is superimposed on any other net motion it may have. These waves are reflected, refracted, scattered, and absorbed as they encounter solid objects. Sound is transmitted through solids in a complex array of types of elastic waves. Sound is characterized by its amplitude, frequency, phase, and direction of propagation.

Sound strength is therefore location-dependent and is often quantified as a sound pressure level (L_p) in dB based on the root-mean-square (rms) sound-pressure (p_S) value, where

$$L_p = 10 \log_{10}\left(\frac{p_S}{p_{\text{reference}}}\right)^2 \tag{8-113}$$

For airborne sound, the reference pressure is 2×10^{-5} Pa (29×10^{-10} psi), which is nominally the human threshold of hearing at 1000 Hz. The corresponding sound pressure level is 0 dB. Conversation is about 50 dB, and a jackhammer operator is subject to 100 dB. Extreme levels such as a jet engine at takeoff might produce 140 dB at a distance of 3 m, which is a pressure amplitude of 200 Pa (29×10^{-3} psi). These examples demonstrate both the sensitivity and wide dynamic range of the human ear.

Traveling sound waves carry energy. Sound intensity I is a measure of the power passing through a unit area in a specified direction and is related to p_S. Measuring sound intensity in a process plant gives clues as to the location of the source. As one moves away from the source, the fact that the energy is spread over a larger area requires that sound pressure level decrease. For example, doubling one's distance from a point source reduces the L_p by 6 dB. Viscous action from the induced fluid motion absorbs additional acoustic energy. However, in free air, this viscous damping is negligible over short distances (on the order of a meter).

Noise is a group of sounds with many nonharmonic frequency components of varying amplitudes and random phase. The turbulence generated by a throttling valve creates noise. As a valve converts potential energy to heat, some of it becomes acoustic energy as an intermediate step. Valves handling large amounts of compressible fluid through a large pressure change create the most noise because more total power is being transformed. Liquid flows are noisy only

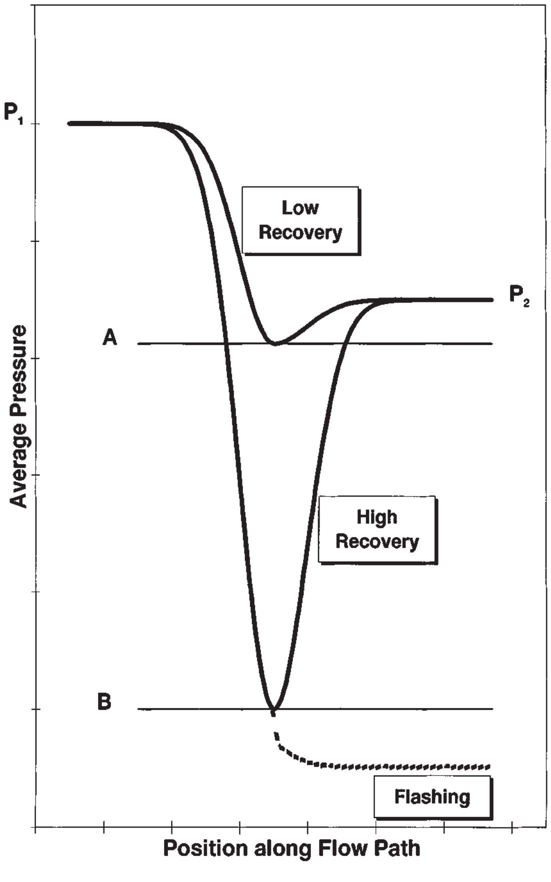


FIG. 8-78 Generic depictions of average pressure at subsequent cross sections throughout a control valve. F_Ls selected for illustration are 0.9 and 0.63 for low and high recovery, respectively. Internal pressure in the high-recovery valve is shown as a dashed line for flashing conditions ($p_2 < p_v$) with p_v = B.

under special circumstances as will be seen in the next subsection. Due to the random nature of turbulence and the broad distribution of length and velocity scales of turbulent eddies, valve-generated sound is usually random, broad-spectrum noise. Total sound pressure level from two such statistically uncorrelated sources is (in dB):

$$L_p = 10 \log_{10}\left[\frac{(p_{S1})^2 + (p_{S2})^2}{(p_{\text{reference}})^2}\right] \tag{8-114}$$

For example, two sources of equal strength combine to create an L_p that is 3 dB higher.

While noise is annoying to listen to, the real reasons for being concerned about noise are its impact on people and equipment. Hearing loss can occur due to long-term exposure to moderately high or even short exposure to very high noise levels. The U.S. Occupational Safety and Health Act (OSHA) has specific guidelines for permissible levels and exposure times. The human ear has a frequency-dependent sensitivity to sound. When the effect on humans is the criteria, L_p measurements are weighted to account for the ear's response. This so-called *A*-weighted scale is defined in ANSI S1.4 and is commonly reported as L_{pA}. Figure 8-79 illustrates the difference between actual and perceived airborne sound pressure level. At sufficiently high levels, noise and the associated vibration can damage equipment.

There are two approaches to fluid-generated noise control—source or path treatment. Path treatment means absorbing or blocking the transmission of noise after it has been created. The pipe itself is a barrier. The sound pressure level inside a standard schedule pipe is roughly 40–60 dB higher than on the outside. Thicker walled pipe reduces levels somewhat more, and adding acoustical insulation on the outside of the pipe reduces ambient levels up to 10 dB per inch of thickness. Since noise propagates relatively unimpeded inside the pipe, barrier approaches require the entire downstream piping system to be treated in order to be totally effective. In-line silencers place absorbent material inside the flow stream, thus reducing the level of the internally propagating noise. Noise reductions up to 25 dB can be achieved economically with silencers.

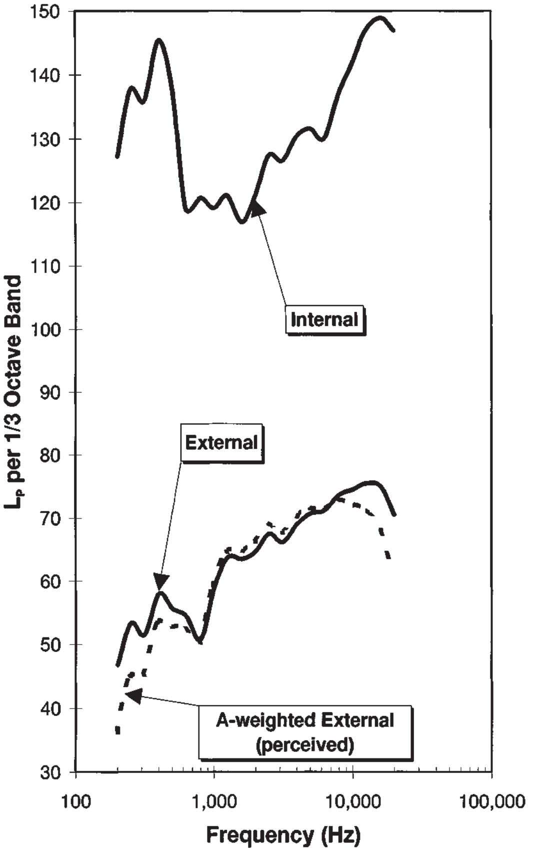


FIG. 8-79 Valve-generated sound pressure level spectrums.

The other approach to valve noise problems is the use of quiet trim. Two basic strategies are used to reduce the initial production of noise—dividing the flow stream into multiple paths and using several flow resistances in series. L_p is proportional to mass flow and is dependent on vena contracta velocity. If each path is an independent source, it is easy to show from Eq. (8-114) that p_s^2 is inversely proportional to the number of passages; additionally, smaller passage size shifts the predominate spectral content to higher frequencies, where structural resonance may be less of a problem. Series resistances or multiple stages can reduce maximum velocity and/or produce back pressure to keep jets issuing from multiple passages acting independently. While some of the basic principles are understood, predicting noise for a particle-flow passage requires some empirical data as a basis. Valve manufacturers have developed noise-prediction methods for the valves they build. ISA S75.17 is a public-domain methodology for standard (nonlow noise) valve types, although treatment of some multistage, multipath types is underway. Low-noise hardware consists of special cages in linear stem valves, perforated domes or plates and multichannel inserts in rotary valves, and separate devices that use multiple fixed restrictions.

Cavitation and Flashing From the discussion on pressure recovery it was seen that the pressure at the vena contracta can be much lower than the downstream pressure. If the pressure on a liquid falls below its vapor pressure (p_v), the liquid will vaporize. Due to the effect of surface tension, this vapor phase will first appear as bubbles. These bubbles are carried downstream with the flow, where they collapse if the pressure recovers to a value above p_v. This pressure-driven process of vapor-bubble formation and collapse is known as cavitation.

Cavitation has three negative side effects in valves—noise and vibration, material removal, and reduced flow. The bubble-collapse process is a violent asymmetrical implosion that forms a high-speed microjet and induces pressure waves in the fluid. This hydrodynamic noise and the mechanical vibration that it can produce are far stronger than other noise-generation sources in liquid flows. If implosions occur adjacent to a solid component, minute pieces of material can be removed, which, over time, will leave a rough, cinderlike surface.

The presence of vapor in the vena contracta region puts an upper limit on the amount of liquid that will pass through a valve. A mixture of vapor and liquid has a lower density than the liquid alone. While Eq. (8-108) is not applicable to two-phase flows because pressure changes are redistributed due to varying density and the two phases do not necessarily have the same average velocity, it does suggest that lower density reduces total mass flow rate. Figure 8-80 illustrates a typical flow-rate-to-pressure-drop relationship. As with compressible gas flow at a given p_1, flow increases as p_2 is decreased until the flow chokes (i.e., no additional fluid will pass). The transition between incompressible and choked flow is gradual because, within the convoluted flow passages of valves, the pressure is actually an uneven distribution at each cross section, and, consequently, vapor-formation zones increase gradually. In fact, isolated zones of bubble formation or incipient cavitation often occur at pressure drops well below that at which a reduction in flow is noticeable. The similarity between liquid and gas choking is not serendipitous; it is surmised that the two-phase fluid is traveling at the mixture's sonic velocity in the throat when choked. Complex fluids with components having varying vapor pressures and/or entrained noncondensable gases (e.g., crude oil) will exhibit soft vaporization/implosion transitions.

There are several methods to reduce cavitation or at least its negative side effects. Material damage is slowed by using harder materials and by directing the cavitating stream away from passage walls (e.g., with an angle body flowing down). Sometimes the system can be designed to place the valve in a higher p_2 location or add downstream resistance, which creates back pressure. A low recovery valve has a higher minimum pressure for a given p_2 and so is a means to eliminate the cavitation itself, not just its side effects. In Fig. 8-78, if p_v < "B" neither valve will cavitate substantially. For p_v > "B" but < "A," the

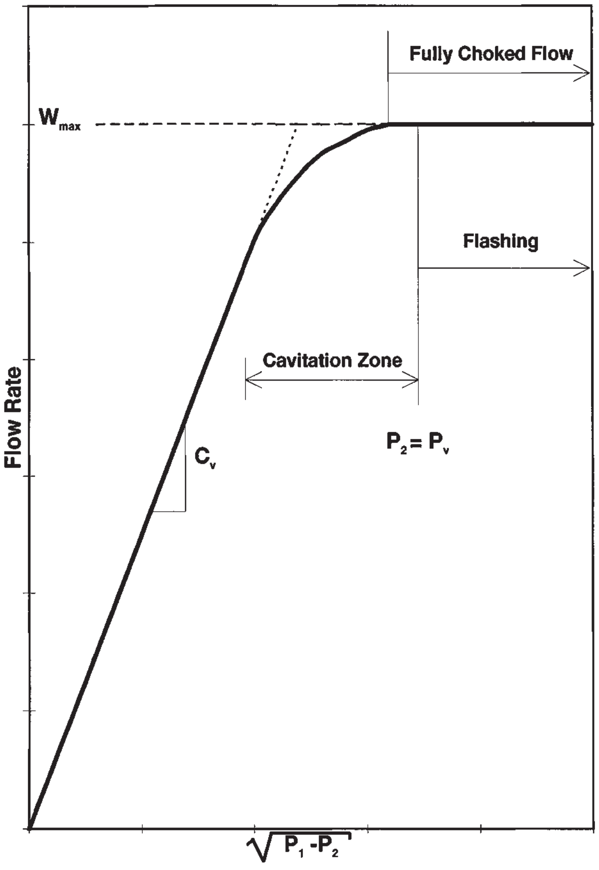


FIG. 8-80 Liquid flow rate versus pressure drop (assuming constant P_1 and P_v).

high recovery valve will cavitate substantially, but the low recovery valve will not. Special anticavitation trims are available for globe/angle valves and more recently for some rotary valves. These trims use multiple contraction/expansion stages or other distributed resistances to boost F_L to values sometimes near unity.

If p_2 is below p_v, the two-phase mixture will continue to vaporize in the body outlet and/or downstream pipe until all liquid phase is gone, a condition known as flashing. The resulting huge increase in specific volume leads to high velocities, and any remaining liquid droplets acquire much of the higher vapor-phase velocity. Impingement of these droplets can produce material damage, but it differs from cavitation damage because it exhibits a smooth surface. Hard materials and directing the two-phase jets away from solid surfaces are means to avoid this damage.

Seals, Bearings, and Packing Systems In addition to their control function, valves often need to provide shutoff. ANSI B16.104, FCI 70-2 (1991), and IEC 534-4 all recognize six standard classifications and define their as-shipped qualification tests. Class I is an amount agreed to by user and supplier with no test needed. Classes II, III, and IV are based on an air test with maximum leakage of 0.5 percent, 0.1 percent, and 0.01 percent of rated capacity, respectively. Class V restricts leakage to 5×10^{-6} ml of water per second per mm of port diameter per bar differential. Class VI allows 0.15 to 6.75 ml per minute of air to escape depending on port size; this class implies the need for interference-fit elastomeric seals. With the exception of Class V, all classes are based on standardized pressure conditions that may not represent actual conditions. Therefore, it is difficult to estimate leakage in service. Leakage normally increases over time as seals and seating surfaces become nicked or worn. Leak passages across the seat-contact line, known as wire drawing, may form and become worse over time—even in hard metal seats under sufficiently high pressure differentials.

Polymers used for seat and plug seals and internal static seals include: PTFE (polytetrafluoroethylene) and other fluorocarbons, polyethylene, nylon, polyether-ether-ketone, and acetal. Fluorocarbons are often carbon or glass-filled to improve mechanical properties and heat resistance. Temperature and chemical compatibility with the process fluid are the key selection criteria. Polymer-lined bearings and guides are used to decrease friction, which lessens dead band and reduces actuator force requirements. See Sec. 28, “Materials of Construction,” for properties.

Packing forms the pressure-tight seal, where the stem protrudes through the pressure boundary. Packing is typically made from PTFE or, for high temperature, a bonded graphite. If the process fluid is toxic, more sophisticated systems such as dual packing, live-loaded, or a flexible metal bellows may be warranted. Packing friction can significantly degrade control performance. Pipe, bonnet, and internal-trim joint gaskets are typically a flat sheet composite. Gaskets intended to absorb dimensional mismatch are typically made from filled spiral-wound flat stainless steel wire with PTFE or graphite filler. The use of asbestos in packing and gaskets has largely been eliminated.

Flow Characteristics The relationship between valve flow and valve travel is called the valve-flow characteristic. The purpose of flow characterization is to make loop dynamics independent of load, so that a single controller tuning remains optimal for all loads. Valve gain is one factor affecting loop dynamics. In general, gain is the ratio of change in output to change in input. The input of a valve is travel (y) and the output is flow (w). Since pressure conditions at the valve can depend on flow (hence travel), valve gain is

$$\frac{dw}{dy} = \frac{\partial w}{\partial C_v}\frac{dC_v}{dy} + \frac{\partial w}{\partial p_1}\frac{dp_1}{dy} + \frac{\partial w}{\partial p_2}\frac{dp_2}{dy} \qquad (8\text{-}115)$$

An inherent valve flow characteristic is defined as the relationship between flow rate and travel, under constant pressure conditions. Since the last two terms in Eq. (8-115) are zero in this case, the inherent characteristic is necessarily also the relationship between flow coefficient and travel.

Figure 8-81 shows three common inherent characteristics. A linear characteristic has a constant slope, meaning the inherent valve gain is a constant. The most popular characteristic is equal-percentage, which gets its name from the fact that equal changes in travel produce equal-percentage changes in the existing flow coefficient. In other words, the slope of the curve is proportional to C_v or equivalently that inherent valve gain is proportional to flow. The equal-percentage characteristic can be expressed mathematically by

$$C_v(y) = (\text{rated } C_v)\exp\left[\left(\frac{y}{\text{rated } y} - 1\right)\ln R\right] \qquad (8\text{-}116)$$

This expression represents a set of curves parameterized by R. Note that C_v (y = 0) equals (rated C_v)/R rather than zero; real equal-percentage characteristics deviate from theory at some small travel to meet shutoff requirements. An equal-percentage characteristic provides perfect compensation for a process that has gain inversely proportional to flow (e.g., liquid pressure). Quick opening does not have a standardized mathematical definition. Its shape arises naturally from high-capacity plug designs used in on/off service globe valves.

Frequently, pressure conditions at the valve will change with flow rate. This so-called process influence [the last two terms on the right hand side of Eq. (8-115)] combine with inherent gain to express the installed valve gain. The flow-versus-travel relationship for a specific set of conditions is called the installed flow characteristic. Typically, valve Δp decreases with load, since pressure losses in the piping system increase with flow. Figure 8-82 illustrates how allocation of total system head to the valve influences the installed flow characteristics. For a linear or quick-opening characteristic, this transition toward a concave down shape would be more extreme. This effect of typical process pressure variation, which causes equal-percentage character-

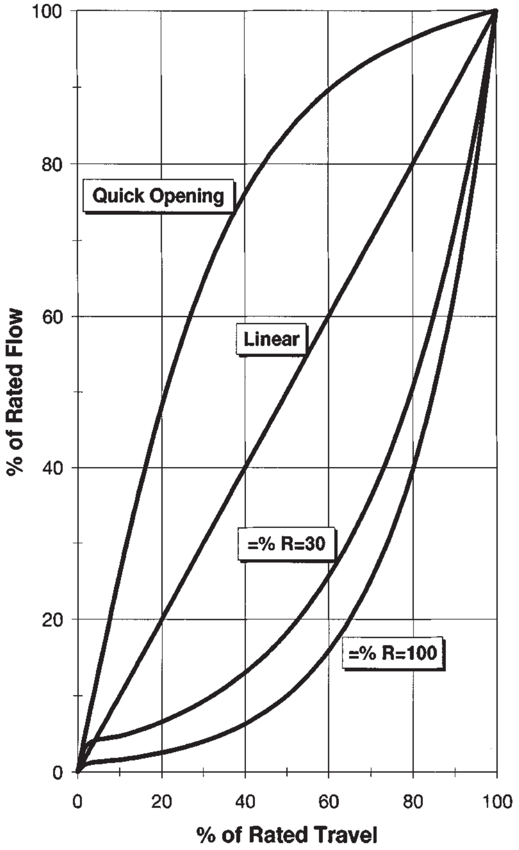


FIG. 8-81 Typical inherent flow characteristics.

istics to have fairly constant installed gain, is one reason the equal-percentage characteristic is the most popular.

Due to clearance flow, flow force gradients, seal friction, and the like, flow cannot be throttled to an arbitrarily small value. Installed rangeability is the ratio of maximum to minimum controllable flow. The actuator and positioner, as well as the valve, influence the installed rangeability. Inherent rangeability is defined as the ratio of the largest to the smallest C_v within which the characteristic meets specified criteria (see ISA S75.11). The *R* value in the equal-percentage definition is a theoretical rangeability only. While high installed rangeability is desirable, it is also important not to oversize a valve; otherwise, turndown (ratio of maximum normal to minimum controllable flow) will be limited.

Sliding stem valves are characterized by altering the contour of the plug when the port and plug determine the minimum (controlling) flow area. Passage area versus travel is also easily manipulated in characterized cage designs. Inherent rangeability varies widely, but typical values are 30 for contoured plugs and 20–50 for characterized cages. While these types of valves can be characterized, the degree to which manufacturers conform to the mathematical ideal is revealed by plotting measured C_v versus travel. Note that ideal equal percentage will plot as a straight line on a semilog graph. Custom characteristics that compensate for a specific process are possible.

Rotary stem-valve designs are normally offered only in their naturally occurring characteristic, since it is difficult to appreciably alter this. If additional characterization is required, the positioner or controller may be characterized. However, these approaches are less direct, since it is possible for device nonlinearity and dynamics to distort the compensation.

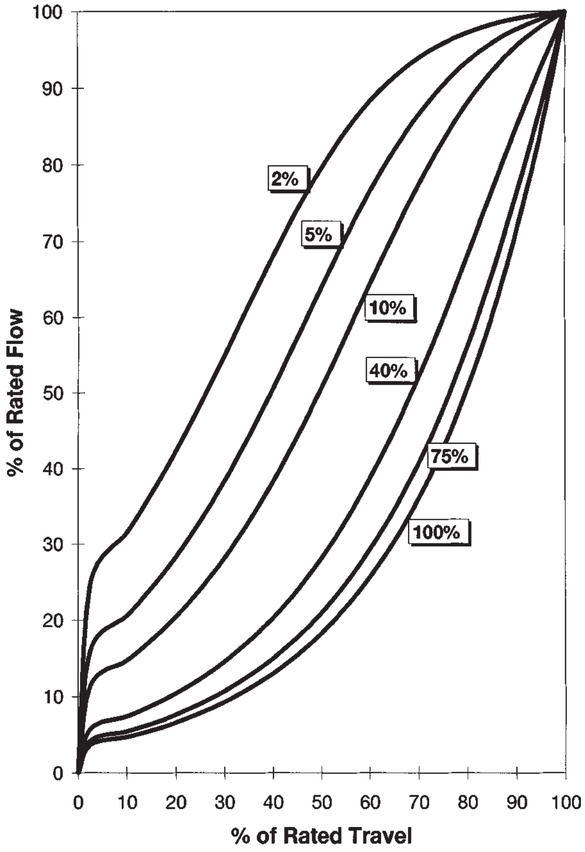


FIG. 8-82 Installed flow characteristic as a function of percent of total system head allocated to the control valve (assuming constant head pump, no elevation head loss, and an *R* equal 30 equal-percentage inherent characteristic).

OTHER PROCESS VALVES

In addition to the throttling control valve, other types of process valves are used to manipulate the process.

Valves for On/Off Applications Valves are often required for service that is primarily nonthrottling in nature. Valves in this category, depending on the service requirements, may be of the same design as the types used for throttling control or, as in the case of gate valves, different in design. Valves in this category usually have tight shutoff when they are closed and low pressure drops when they are wide open. The on/off valve can be operated manually, such as by handwheel or lever; or automatically, with pneumatic or electric actuators.

Batch Batch process operation is an application requiring on/off valve service. Here the valve is opened and closed to provide reactant, catalyst, or product to and from the batch reactor. Like the throttling control valve, the valve used in this service must be designed to open and close thousands of times. For this reason, valves used in this application are often the same valves used in continuous throttling applications. Ball valves are especially useful in batch operations. The ball valve has a straight-through flow passage that reduces pressure drop in the wide-open state and provides tight shutoff capability when closed. In addition, the segmented ball valve provides for shearing action between the ball and the ball seat that promotes closure in slurry service.

Isolation A means for pressure-isolating control valves, pumps, and other piping hardware for installation and maintenance is another common application for an on/off valve. In this application, the valve is required to have tight shutoff so that leakage is stopped when the piping system is under repair. As the need to cycle the valve in this application is far less than that of a throttling control valve, the wear characteristics of the valve are less important. Also, because there are many required in a plant, the isolation valve needs to be reliable, simple in design and simple in operation. The gate valve, shown in Figure 8-83, is the most widely used valve in this application.

The gate valve is composed of a gate-like disc that moves perpendicular to the flow stream. The disc is moved up and down by a threaded screw that is rotated to effect disc movement. Because the disc is large and at right angles to the process pressure, large seat loading for tight shutoff is possible. Wear produced by high seat loading during the movement of the disk prohibits the use of the gate valve for throttling applications.

Pressure Relief Valves Definitions for pressure relief valves, relief valves, pilot-operated pressure relief valves and safety valves, are found in the ASME Boiler and Pressure Vessel Code, Section VIII, Division 1, "Rules for Construction of Pressure Vessels," Paragraphs UG-125 and UG-126. The pressure-relief valve is an automatic pressure relieving device designed to open when normal conditions are exceeded and to close again when normal conditions are restored. Within this class there are relief valves, pilot operated pressure relief valves, and safety valves.

Relief valves (see Fig. 8-84) have spring-loaded disks that close a main orifice against a pressure source. As pressure rises, the disk begins to rise off the orifice and a small amount of fluid passes through the valve. Continued rise in pressure above the opening pressure causes the disk to open the orifice in a proportional fashion. The main orifice reduces and closes when the pressure returns to the set pressure. Additional sensitivity to over-pressure conditions can be improved by adding an auxiliary pressure relief valve (pilot) to the basic pressure relief valve. This combination is known as a pilot-operated pressure relief valve.

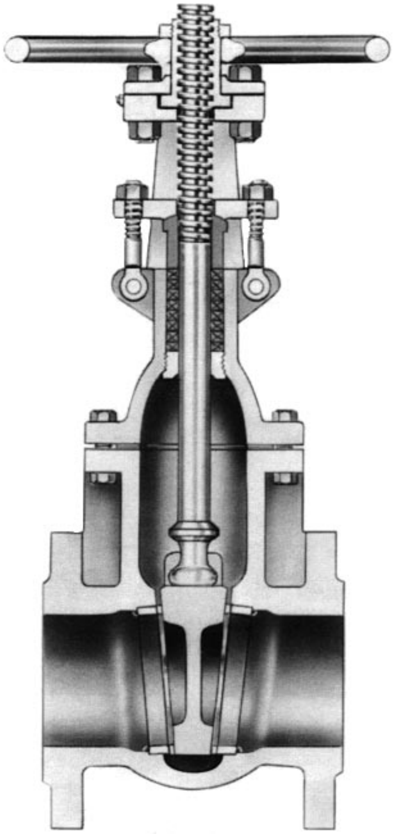


FIG. 8-83 Gate valve. (*Courtesy Crane Valves.*)

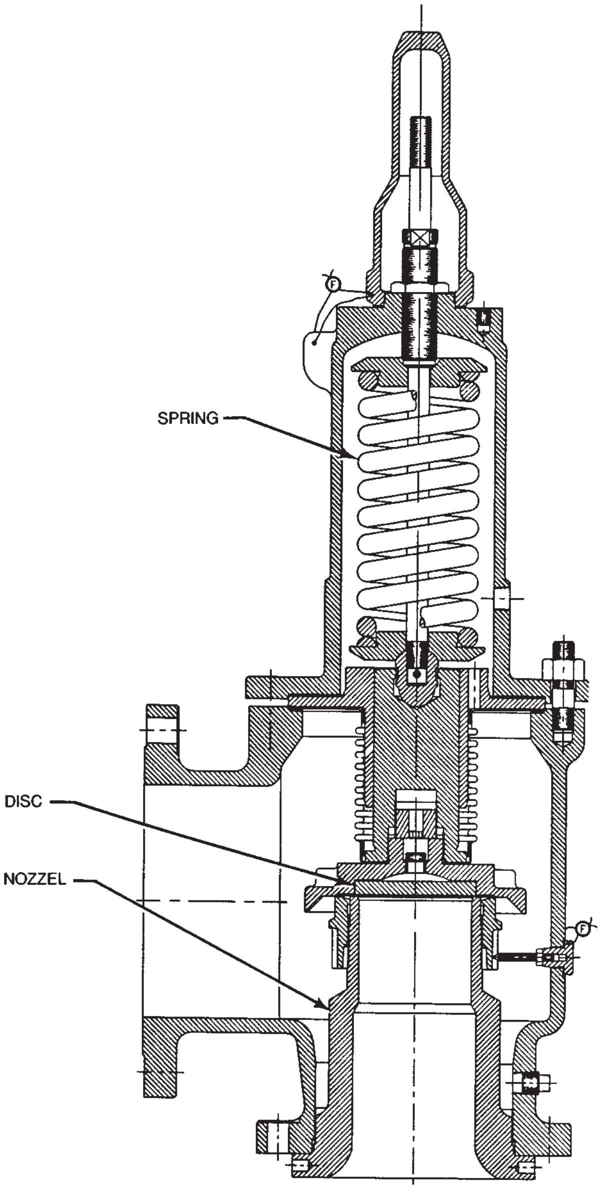


FIG. 8-84 Relief valve. (*Courtesy Teledyne Fluid Systems, Farris Engineering.*)

The safety valve is similar to the relief valve except it is designed to open fully, or pop, with only a small amount of pressure over the rated limit. Conventional safety valves are sensitive to downstream pressure and may have unsatisfactory operating characteristics in variable back pressure applications. The balanced safety relief valve is available and minimizes the effect of downstream pressure on performance.

Check Valves The purpose of a check valve is to allow relatively unimpeded flow in the desired direction but to prevent flow in the reverse direction. Two common designs are swing-type and lift-type check valves—the names of which denote the motion of the closure member. In the forward direction, flow forces overcome the weight of the member or a spring to open the flow passage. With reverse pressure conditions, flow forces drive the closure member into the valve seat, thus providing shutoff.

ADJUSTABLE SPEED PUMPS

An alternative to throttling a process with a process-control valve and a fixed speed pump is by adjusting the speed of the process pump and not using a throttling control valve at all. Pump speed can be varied by using variable-speed prime movers such as turbines, motors with magnetic or hydraulic couplings, and electric motors. Each of these methods of modulating pump speed has its own strengths and weaknesses but all offer energy savings and dynamic performance advantages over throttling with a control valve.

The centrifugal pump directly driven by a variable-speed electric motor is the most commonly used hardware combination for adjustable speed pumping. The motor is operated by an electronic-motor speed controller whose function is to generate the voltage or current waveform required by the motor to make the speed of the motor track the input command input signal from the process controller.

The most popular form of motor speed control for adjustable-speed pumping is the voltage-controlled pulse-width-modulated (PWM) frequency synthesizer and AC squirrel-cage induction motor combination. The flexibility of application of the PWM motor drive and its 90 percent+ electrical efficiency along with the proven ruggedness of the traditional AC induction motor makes this combination popular.

From an energy-consumption standpoint, the power required to maintain steady process flow with an adjustable-speed pump system (three-phase PWM drive and a squirrel-cage induction motor driving a centrifugal pump on water) is less than that required with a conventional control valve and a fixed speed pump. Figure 8-85 shows this to be the case for a system where 100 percent of the pressure loss is due to flow velocity losses. At 75 percent flow, Fig. 8-85 shows the constant speed-pump/control-valve use at a 10.1-kW rate where throttling with the adjustable speed pump and no control valve used at a 4.1-kW rate. This trend of reduced energy consumption is true for the entire range of flows, although amounts vary.

From a dynamic-response standpoint, the adjustable speed pump has a dynamic characteristic that is more suitable in process-control applications than those characteristics of control valves. The small amplitude response of an adjustable speed pump does not contain the dead band or the dead time commonly found in the small amplitude response of the control valve. Nonlinearities associated with frictions in the valve and discontinuities in the pneumatic portion of the control-valve instrumentation are not present with electronic variable-speed drive technology. As a result, process control with the adjustable speed pump does not exhibit limit cycles, problems related to low controller gain and generally degraded process loop performance caused by control valve nonlinearities.

Unlike the control valve, the centrifugal pump has poor or nonexistent shutoff capability. A flow check valve or an automated on/off valve may be required to achieve shutoff requirements. This requirement may be met by automating an existing isolation valve in retrofit applications.

REGULATORS

A regulator is a compact device that maintains the process variable at a specific value in spite of disturbances in load flow. It combines the functions of the measurement sensor, controller, and final control element into one self-contained device. Regulators are available to control pressure, differential pressure, temperature, flow, liquid level, and other basic process variables. They are used to control the differential across a filter press, heat exchanger, or orifice plate. Regulators are used for monitoring pressure variables for redundancy, flow check, and liquid surge relief.

Regulators may be used in gas blanketing systems to maintain a protective environment above any liquid stored in a tank or vessel as the liquid is pumped out. When the temperature of the vessel is suddenly cooled, the regulator maintains the tank pressure and protects the walls of the tank from possible collapse. Regulators are known for their fast dynamic response. The absence of time delay that often comes with more sophisticated control systems makes the regulator useful in applications requiring fast corrective action.

Regulators are designed to operate on the process pressures in the pipeline without any other sources of energy. Upstream and downstream pressures are used to supply and exhaust the regulator. Exhausting is back to the downstream piping so that no contamination or leakage to the external environment occurs. This makes regulators useful in remote locations where power is not available or where external venting is not allowed.

The regulator is limited to operating on processes with clean, nonslurry process fluids. The small orifice and valve assemblies contained in the regulator can plug and malfunction if the process fluid that operates the regulator is not sufficiently clean.

Regulators are normally not suited to systems that require constant set point adjustment. Although regulators are available with capability to respond to remote set point adjustment, this feature adds complexity to the regulator and may be better addressed by a control-valve-based system. In the simplest of regulators, tuning of the regulator for best control is accomplished by changing a spring, an orifice, or a nozzle.

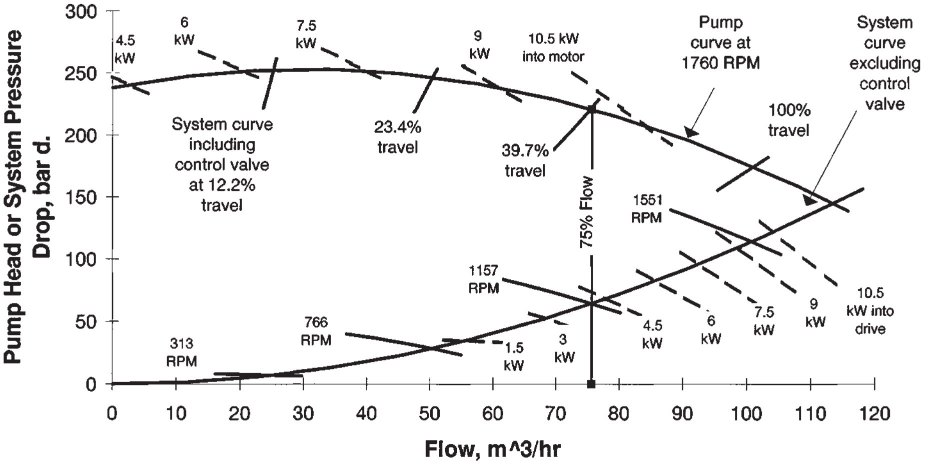
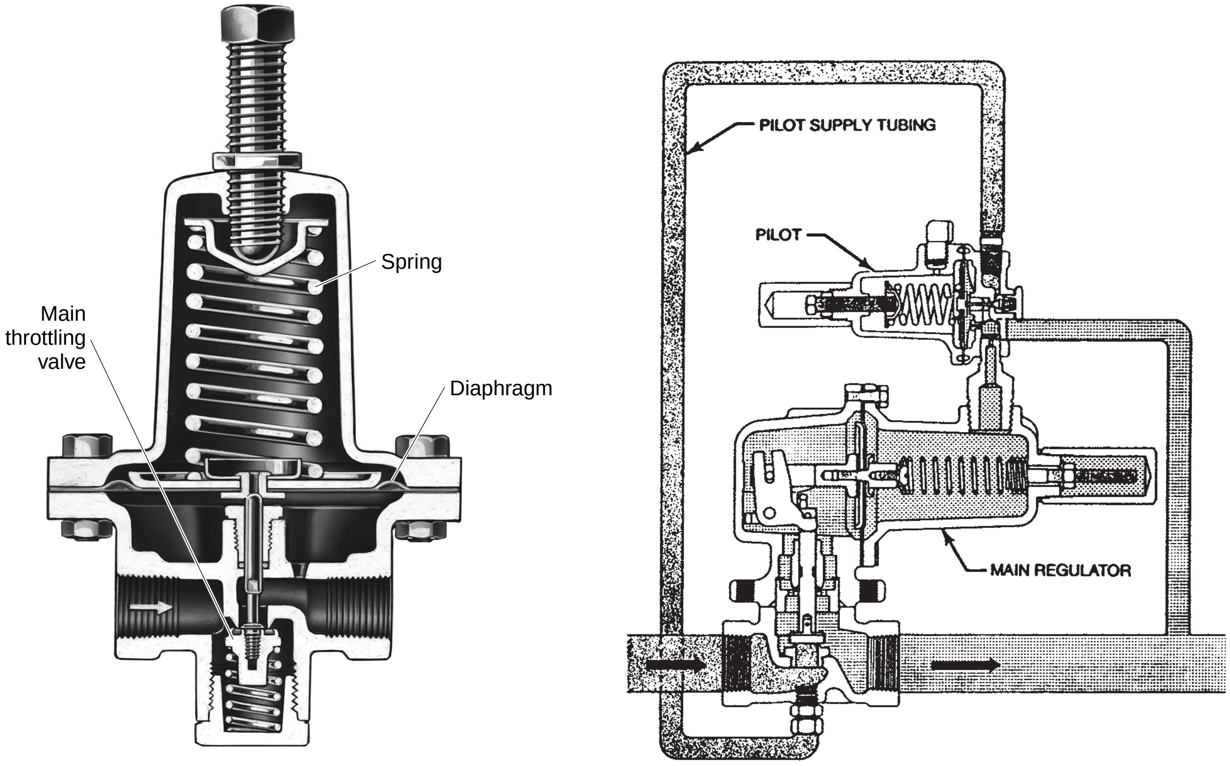


FIG. 8-85 Pressure, flow, and power for a throttling process using (1) a control valve and a constant speed pump and (2) an adjustable speed pump.



(a)

(b)

FIG. 8-86 Regulators: (*a*) self-operated; (*b*) pilot-operated. (*Courtesy Fisher-Rosemount.*)

Self-Operated Regulators Self-operated regulators are the simplest form of regulator. This regulator (see Fig. 8-86*a*) is composed of a main throttling valve, a diaphragm or piston to sense pressure, and a spring. The self-contained regulator is completely operated by the process fluid, and no outside control lines or pilot stage is used. In general, self-operated regulators are simple in construction, easy to operate and maintain, and are usually stable devices. Except for some of the pitot tube types, self-operated regulators have very good dynamic response characteristics. This is because any change in the controlled variable registers directly and immediately upon the main diaphragm to produce a quick response to the disturbance.

The disadvantage of the self-operated regulator is that it is not generally capable of maintaining a set point as load flow is increased. Because of the proportional nature of the spring and diaphragm-throttling effect, offset from set point occurs in the controlled variable as flow increases. Figure 8-87 shows a typical regulation curve for the self-contained regulator.

Reduced set point offset with increasing load flow can be achieved

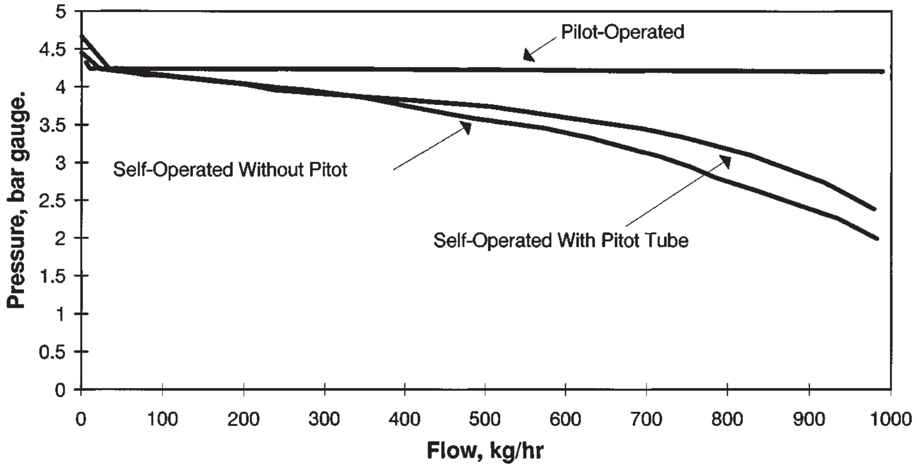


FIG. 8-87 Pressure regulation curves for three regulator types.

by adding a pitot tube to the self-operated regulator. The tube is positioned somewhere near the vena contracta of the main regulator valve. As flow though the valve increases, the measured feedback pressure from the pitot tube drops below the control pressure. This causes the main valve to open or boost more than it would if the static value of control pressure were acting on the diaphragm. The resultant effect keeps the control pressure closer to set point and thus prevents a large drop in process pressure during high-load-flow conditions. Figure 8-87 shows the improvement that the pitot-tube regulator provides over the regulator without the tube. A side effect of adding a pitot-tube method is that the response of the regulator can be slowed due to the restriction provided by the pitot tube.

Pilot-Operated Regulators Another category of regulators uses a pilot stage to provide the load pressure on the main diaphragm. This pilot is a regulator itself that has the ability to multiply a small change in downstream pressure into a large change in pressure applied to the regulator diaphragm. Due to this high-gain feature, pilot-operated regulators can achieve a dramatic improvement in steady-state accuracy over that achieved with a self-operated regulator. Figure 8-87 shows for regulation at high flows the pilot-operated regulator is best of the three regulators shown.

The main limitation of the pilot-operated regulator is stability. When the gain in the pilot amplifier is raised too much, the loop can become unstable and oscillate or hunt. The two-path pilot regulator (see *b*) is also available. This regulator combines the effects of self-operated and the pilot-operated styles and mathematically produces the equivalent of proportional plus reset control of the process pressure.

Over-Pressure Protection Figure 8-87 shows a characteristic rise in control pressure that occurs at low or zero flow. This lockup tail is due to the effects of imperfect plug and seat alignment and the elastomeric effects of the main throttle valve. If for some reason the main throttle valve fails to completely shut off, or if the valve shuts off but the control pressure continues to rise for other reasons, the lockup tail could get very large, and the control pressure could rise to extremely high valves. Damage to the regulator or the downstream pressure volume could occur.

To avoid this situation, some regulators are designed with a built-in over-pressure relief mechanism. Over-pressure relief circuits usually are composed of a spring-opposed diaphragm and valve assembly that vents the downstream piping when the control pressure rises above the set point pressure.

PROCESS CONTROL AND PLANT SAFETY

General References: *Guidelines for Safe Automation of Chemical Processes,* AIChE Center for Chemical Process Safety, New York, 1993.

Accidents in chemical plants make headline news, especially when there is loss of life or the general public is affected in even the slightest way. This increases the public's concern and may lead to government action. The terms *hazard* and *risk* are defined as follows:

- *Hazard.* A potential source of harm to people, property, or the environment
- *Risk.* Possibility of injury, loss, or an environmental accident created by a hazard

Safety is the freedom from hazards and thus the absence of any associated risks. Unfortunately, absolute safety cannot be realized.

The design and implementation of safety systems must be undertaken with a view of two issues:

- *Regulatory.* The safety system must be consistent with all applicable codes and standards as well as "generally accepted good engineering practices."
- *Technical.* Just meeting all applicable regulations and "following the crowd" does not relieve a company of its responsibilities. The safety system must work.

The regulatory environment will continue to change. As of this writing, the key regulatory instrument is OSHA 29 CFR 1910.119 that pertains to process safety management within plants in which certain chemicals are present.

In addition to government regulation, industry groups and professional societies are producing documents ranging from standards to guidelines. Instrument Society of America Standard S84.01, "Application of Safety Instrumented Systems for the Process Industries," is in draft form at the date of this writing. The *Guidelines for Safe Automation of Chemical Processes* from the American Institute of Chemical Engineers' Center for Chemical Process Safety (1993) provides a comprehensive coverage of the various aspects of safety, and, although short on specifics, it is very useful to operating companies developing their own specific safety practices (that is, it does not tell you what to do, but it helps you decide what is proper for your plant).

The ultimate responsibility for safety rests with the operating company; OSHA 1910.119 is clear on this. Each company is expected to develop (and enforce) its own practices in the design, installation, testing, and maintenance of safety systems. Fortunately, some companies make these documents public. Monsanto's *Safety System Design Practices* was published in its entirety in the proceedings of the International Symposium and Workshop on Safe Chemical Process Automation, Houston, Texas, September 27–29, 1994 (available from the American Institute of Chemical Engineers' Center for Chemical Process Safety).

ROLE OF AUTOMATION IN PLANT SAFETY

As microprocessor-based controls displaced hardwired electronic and pneumatic controls, the impact on plant safety has definitely been positive. When automated procedures replace manual procedures for routine operations, the probability of human errors leading to hazardous situations is lowered. The enhanced capability for presenting information to the process operators in a timely manner and in the most meaningful form increases the operator's awareness of the current conditions in the process. Process operators are expected to exercise due diligence in the supervision of the process, and timely recognition of an abnormal situation reduces the likelihood that the situation will progress to the hazardous state. Figure 8-88 depicts the layers of safety protection in a typical chemical plant.

Although microprocessor-based process controls enhance plant safety, their primary objective is efficient process operation. Manual

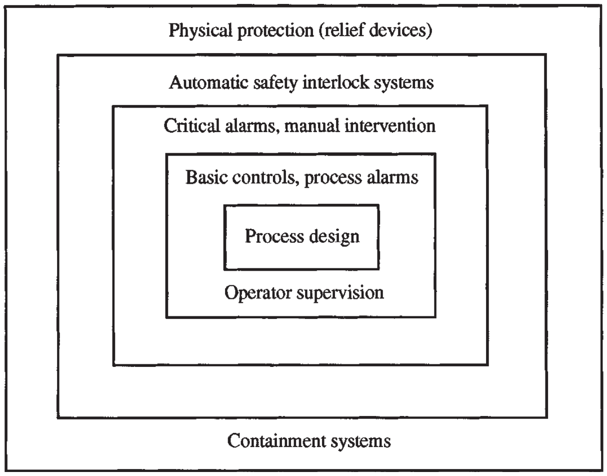

Physical protection (relief devices)
Automatic safety interlock systems
Critical alarms, manual intervention
Basic controls, process alarms
Process design
Operator supervision
Containment systems

FIG. 8-88 Layers of safety protection in chemical plants.

operations are automated to reduce variability, to minimize the time required, to increase productivity, and so on. Remaining competitive in the world market demands that the plant be operated in the best manner possible, and microprocessor-based process controls provide numerous functions that make this possible. Safety is never compromised in the effort to increase competitiveness, but enhanced safety is a by-product of the process-control function and is not a primary objective.

By attempting to maintain process conditions at or near their design values, the process controls also attempt to prevent abnormal conditions from developing within the process. Although process controls can be viewed as a protective layer, this is really a by-product and not the primary function. Where the objective of a function is specifically to reduce risk, the implementation is normally not within the process controls. Instead, the implementation is within a separate system specifically provided to reduce risk. This system is generally referred to as the safety interlock system.

As safety begins with the process design, an inherently safe process is the objective of modern plant designs. When this cannot be achieved, process hazards of varying severity will exist. Where these hazards put plant workers and/or the general public at risk, some form of protective system is required. Process safety management addresses the various issues, ranging from assessment of the process hazard to assuring the integrity of the protective equipment installed to cope with the hazard. When the protective system is an automatic action, it is incorporated into the safety interlock system, not within the process controls.

INTEGRITY OF PROCESS CONTROL SYSTEMS

Ensuring the integrity of process controls involves both hardware issues, software issues, and human issues. Of these, the hardware issues are usually the easiest to assess and the software issues the most difficult.

The hardware issues are addressed by providing various degrees of redundancy, by providing multiple sources of power and/or an uninterruptible power supply, and the like. The manufacturers of process controls provide a variety of configuration options. Where the process is inherently safe and infrequent shutdowns can be tolerated, nonredundant configurations are acceptable. For more demanding situations, an appropriate requirement might be that no single component failure can render the process-control system inoperable. For the very critical situations, triple-redundant controls with voting logic might be appropriate. The difficulty is assessing what is required for a given process.

Another difficulty is assessing the potential for human errors. If redundancy is accompanied with increased complexity, the resulting increased potential for human errors must be taken into consideration. Redundant systems require maintenance procedures that can correct problems in one part of the system while the remainder of the system is in full operation. When conducting maintenance in such situations, the consequences of human errors can be rather unpleasant.

The use of programmable systems for process control present some possibilities for failures that do not exist in hard-wired electromechanical implementations. Probably the one of most concern is latent defects or "bugs" in the software, either the software provided by the supplier or the software developed by the user. The source of this problem is very simple. There is no methodology available that can be applied to obtain absolute assurance that a given set of software is completely free of defects. Increased confidence in a set of software is achieved via extensive testing, but no amount of testing results in absolute assurance that there are no defects. This is especially true of real-time systems, where the software can easily be exposed to a sequence of events that was not anticipated. Just because the software performs correctly for each event individually does not mean that it will perform correctly when two (or more) events occur at nearly the same time. This is further complicated by the fact that the defect may not be in the programming; it may be in how the software was designed to respond to the events.

The testing of any collection of software is made more difficult as the complexity of the software increases. Software for process control has progressively become more complex, mainly because the requirements have progressively become more demanding. To remain competitive in the world market, processes must be operated at higher production rates, within narrower operating ranges, closer to equipment limits, and so on. Demanding applications require sophisticated control strategies, which translate into more complex software. Even with the best efforts of both supplier and user, complex software systems are unlikely to be completely free of defects.

CONSIDERATIONS IN IMPLEMENTATION OF SAFETY INTERLOCK SYSTEMS

Where hazardous conditions can develop within a process, a protective system of some type must be provided. Sometimes these are in the form of process hardware such as pressure relief devices. However, sometimes logic must be provided for the specific purpose of taking the process to a state where the hazardous condition cannot exist. The term *safety interlock system* is normally used to designate such logic.

The purpose of the logic within the safety interlock system is very different from the logic within the process controls. Fortunately, the logic within the safety interlock system is normally much simpler than the logic within the process controls. This simplicity means that a hardwired implementation of the safety interlock system is usually an option. Should a programmable implementation be chosen, this simplicity means that latent defects in the software are less likely to be present. Most safety systems only have to do simple things, but they must do them very, very well.

The difference in the nature of process controls and safety interlock systems leads to the conclusion that these two should be physically separated (see Fig. 8-89). That is, safety interlocks should not be piggy-backed onto a process-control system. Instead, the safety interlocks should be provided by equipment, either hard-wired or programmable, that is dedicated to the safety functions. As the process controls become more complex, faults are more likely. Separation means that faults within the process controls have no consequences in the safety interlock system.

Modifications to the process controls are more frequent than modifications to the safety interlock system. Therefore, physically separating the safety interlock system from the process controls provides the following benefits:

1. The possibility of a change to the process controls leading to an unintentional change to the safety interlock system is eliminated.
2. The possibility of a human error in the maintenance of the process controls having consequences for the safety interlock system is eliminated.
3. Management of change is simplified.
4. Administrative procedures for software-version control are more manageable.

Separation also applies to the measurement devices and actuators.

Although the traditional point of reference for safety interlock systems is a hard-wired implementation, a programmed implementation is an alternative. The potential for latent defects in software implementation is a definite concern. Another concern is that solid-state components are not guaranteed to fail to the safe state. The former is addressed by extensive testing; the latter is addressed by manufacturer-supplied and/or user-supplied diagnostics that are routinely executed by the processor within the safety interlock system. Although issues must be addressed in programmable implementations, the hard-wired implementations are not perfect either.

Where a programmed implementation is deemed to be acceptable, the choice is usually a programmable logic controller (PLC) that is dedicated to the safety function. PLCs are programmed using the traditional relay ladder diagrams used for hard-wired implementations. The facilities for developing, testing, and troubleshooting PLCs are excellent. However, for PLCs used in safety interlock systems, administrative procedures must be developed and implemented to address the following issues:

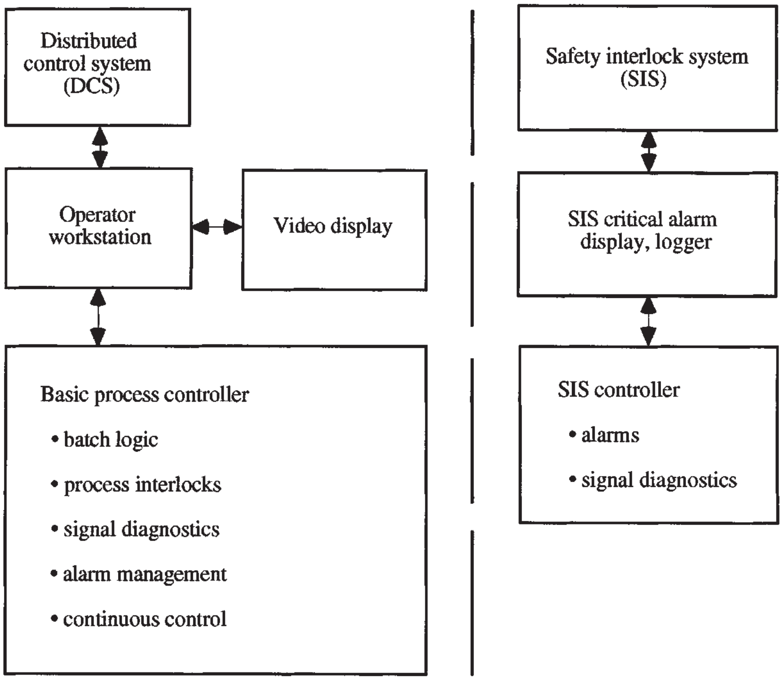


FIG. 8-89 Total control system with parallel tasks.

1. Version controls for the PLC program must be implemented and rigidly enforced. Revisions to the program must be reviewed in detail and thoroughly tested before implementing in the PLC. The various versions must be clearly identified so that there can be no doubt as to what logic is provided by each version of the program.

2. The version of the program that is currently being executed by the PLC must be known with absolute certainty. It must simply not be possible for a revised version of the program undergoing testing to be downloaded to the PLC.

Constant vigilance is required to prevent lapses in such administrative procedures.

INTERLOCKS

An interlock is a protective response initiated on the detection of a process hazard. The interlock system consists of the measurement devices, logic solvers, and final control elements that recognize the hazard and initiate an appropriate response. Most interlocks consist of one or more logic conditions that detect out-of-limit process conditions and respond by driving the final control elements to the safe states. For example, one must specify that a valve fails open or fails closed.

The potential that the logic within the interlock could contain a defect or bug is a strong incentive to keep it simple. Within process plants, most interlocks are implemented with discrete logic, which means either hard-wired electromechanical devices or programmable logic controllers.

Interlocks within process plants can be broadly classified as follows:

1. *Safety interlocks.* These are designed to protect the public, the plant personnel, and possibly the plant equipment from process hazards.

2. *Process interlocks.* These are designed to prevent process conditions that would unduly stress equipment (perhaps leading to minor damage), lead to off-specification product, and so on.

Basically, the process interlocks address hazards whose consequences essentially lead to a monetary loss, possibly even a short plant shutdown. The more serious hazards are addressed by the safety interlocks.

Implementation of process interlocks within process control systems is perfectly acceptable. Furthermore, it is also permissible (and probably advisable) that responsible operations personnel be authorized to bypass or ignore a process. Safety interlocks must be implemented within the separate safety interlock system. Bypassing or ignoring safety interlocks by operations personnel is simply not permitted. When this is necessary for actions such as verifying that the interlock continues to be functional, such situations must be infrequent and incorporated into the design of the interlock.

Safety interlocks are assigned to categories that reflect the severity of the consequences should the interlock fail to perform as intended. The specific categories used within a company is completely at the discretion of the company. However, most companies use categories that distinguish among the following:

1. *Hazards that pose a risk to the public.* Complete redundancy is normally required.

2. *Hazards that could lead to injury of company personnel.* Partial redundancy is often required (for example, redundant measurements but not redundant logic).

3. *Hazards that could result in major equipment damage and consequently lengthy plant downtime.* No redundancy is normally required for these, although redundancy is always an option.

Situations that result in minor equipment damage that can be quickly repaired do not generally require a safety interlock; however, a process interlock might be appropriate.

A process hazards analysis is intended to identify the safety interlocks required for a process and to provide the following for each:

1. The hazard that is to be addressed by the safety interlock.
2. The classification of the safety interlock.
3. The logic for the safety interlock, including inputs from measurement devices and outputs to actuators.

The process hazards analysis is conducted by an experienced, multidisciplinary team that examines the process design, the plant equipment, operating procedures, and so on, using techniques such as hazard and operability studies (HAZOP), failure mode and effect analysis (FEMA), and others. The process hazards analysis recommends appropriate measures to reduce the risk, including (but not limited to) the safety interlocks to be implemented in the safety interlock system.

Diversity is recognized as a useful approach to reduce the number of defects. The team that conducts the process hazards analysis does not implement the safety interlocks but provides the specifications for the safety interlocks to another organization for implementation. This organization reviews the specifications for each safety interlock, seeking clarifications as necessary from the process hazards analysis team and bringing any perceived deficiencies to the attention of the process hazards analysis team.

Diversity can be used to further advantage in redundant configurations. Where redundant measurement devices are required, different technology can be used for each. Where redundant logic is required, one can be programmed and one hard-wired.

Reliability of the interlock systems has two aspects:

1. It must react should the hazard arise.
2. It must not react when there is no hazard.

Emergency shutdowns often pose risks in themselves, and therefore they should be undertaken only when truly appropriate. The need to avoid extraneous shutdowns is not just to avoid disruption in production operations.

Although safety interlocks can inappropriately initiate shutdowns, the process interlocks are usually the major source of problems. It is possible to configure so many process interlocks that it is not possible to operate the plant.

TESTING

As part of the detailed design of each safety interlock, written test procedures must be developed for the following purposes:

1. Assure that the initial implementation complies with the requirements defined by the process hazards analysis team.
2. Assure that the interlock (hardware, software, and I/O) continues to function as designed. The design must also determine the time interval on which this must be done. Often these tests must be done with the plant in full operation.

The former is the responsibility of the implementation team and is required for the initial implementation and following any modification to the interlock. The latter is the responsibility of plant maintenance, with plant management responsible for seeing that it is done on the specified interval of time.

Execution of each test must be documented, showing when it was done, by whom, and the results. Failures must be analyzed for possible changes in the design or implementation of the interlock.

These tests must encompass the complete interlock system, from the measurement devices through the final control elements. Merely simulating inputs and checking the outputs is not sufficient. The tests must duplicate the process conditions and operating environments as closely as possible. The measurement devices and final control elements are exposed to process and ambient conditions and thus are usually the most likely to fail. Valves that remain in the same position for extended periods of time may stick in that position and not operate when needed. The easiest component to test is the logic; however, this is the least likely to fail.